A Bio-Inspired Integration Model of Basal Ganglia and Cerebellum for Motion Learning of a Musculoskeletal Robot*

ZHANG Jinhan · CHEN Jiahao · ZHONG Shanlin · QIAO Hong

DOI: 10.1007/s11424-024-3414-7
Received: 5 October 2023 / Revised: 30 December 2023

Abstract It is a significant research direction for highly complex musculoskeletal robots that how to develop the ability of motion learning and generalization. The cooperations of multiple brain regions are crucial to improving motion performance. Inspired by the neural mechanisms of structures, functions, and interconnections of basal ganglia and cerebellum, a biologically inspired integration model for motor learning of musculoskeletal robots is proposed. Based on the neural characteristics of the basal ganglia, the basal ganglia actor network, which mainly simulates the dorsal striatum, outputs motion commands, and the basal ganglia critic network, which simulates the ventral striatum, estimates action-state values. Their network parameters are updated using the soft actor-critic method. Based on the sensorimotor prediction mechanism of the cerebellum, the cerebellum network evaluates the state feature extraction quality of the basal ganglia actor network and then updates the weights of its feature layer. This learning method is proven to converge to the optimal policy. Furthermore, drawing on the mechanism of dopaminergic dynamic regulation in the basal ganglia, the adaptive adjustment of target entropy and the dopaminergic experience replay are proposed to further improve the integration model, which contributes to the exploration-exploitation trade-off of motor learning. The bio-inspired integration model is validated on a musculoskeletal system. Experimental results indicate that this model can effectively control the musculoskeletal robot to accomplish the motion task from random starting locations to random target positions with high precision and robustness.

Keywords Basal ganglia and cerebellum, bio-inspired integration model, motion learning, musculoskeletal robot, reinforcement learning.

ZHANG Jinhan · CHEN Jiahao · ZHONG Shanlin · QIAO Hong (Corresponding author)
State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190*, China; School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing* 100049*, China.*
Email: zhangjinhan2018@ia.ac.cn; jiahao.chen@ia.ac.cn; shanlin.zhong@ia.ac.cn; hong.qiao@ia.ac.cn.

*This research was supported by Major Project of Science and Technology Innovation 2030 Brain Science and Brain-Inspired Intelligence under Grant No. 2021ZD0200408, the National Natural Science Foundation of China under Grant Nos. 62203439 and 62203443, and Major program of the National Natural Science Foundation of China under Grant Nos. T2293720, T2293723, and T2293724.
◇ *This paper was recommended for publication by Guest Editor LIN Zongli.*

1 Introduction

Robots that can achieve dexterous, flexible and high-precision operations have extensive application requirements in modern industrial systems[1–4]. Accomplishing general and diverse tasks is both a challenge and an opportunity for robotics research. In addition to traditional joint-link rigid robots, musculoskeletal robots that simulate biological characteristics have been focused on improving robot operating performance in recent years[5–7]. Musculoskeletal robots are tendon-driven systems modeled after the arrangement of bones and muscles of organisms. The robots are driven by muscle units, and common driving methods include motor drive and pneumatic drive[8]. There exist many typical musculoskeletal robot platforms, such as ECCEROBOT[9], Kengoro[10], Pneuborn[11], ZAR5[12], lower-limb robot with multifilament muscles[8], robot hand with SCP actuators[13], etc.

Musculoskeletal robots have some irreplaceable advantages. Firstly, with a variety of different joint rotations and muscle coordination, the musculoskeletal robots are able to complete operating tasks flexibly and smoothly. Secondly, due to the inherent feedback from skeletal models to muscle models, the musculoskeletal systems have anti-interference capability[14]. Therefore, musculoskeletal robots are given high expectations to achieve highly sophisticated motions by imitating human motion characteristics. But it should not be ignored that such robots are also complex nonlinear and multi-redundant systems[15]. The muscle force produced by a muscle unit is a highly complex function of muscle length, muscle contraction velocity, and muscle activation[16]. One muscle may have multiple attachment points on several bones, which increases the difficulty of analytically solving joint torques. The output force of each muscle will affect the rotation of multiple joints, and conversely the rotation of each joint is affected by multiple muscles, so it is impossible to control each muscle individually to complete the overall motion tasks. There are many sets of muscle excitation patterns for a musculoskeletal robot to complete one motion task, so it is urgent to effectively explore the action space to obtain optimal or suboptimal control signals.

One of the key issues in musculoskeletal robot research is to improve the ability of motion learning and generalization to complete high-precision and high-reliability tasks. Many representative methods have been proposed to improve the performance of musculoskeletal robots. The model-based methods explicitly analyze the kinematics and dynamics of musculoskeletal systems, and solve muscle forces or muscle control signals according to desired joint torques. Thelen, et al.[17] proposed the computed muscle control (CMC) method, which uses static optimization along with feedforward and feedback controls to drive a musculoskeletal model to achieve the desired motion trajectory. It is a classic algorithm for solving muscle excitation signals. Jäntsch, et al.[18] demonstrated the adaptability of CMC on a hardware platform with eleven tendon-driven compliant muscles. Furthermore, for constrained musculoskeletal systems, Stanev and Moustakas[19] proposed the Task Space Dynamic Inverse Kinematics method to solve the inverse kinematics, and the Task Space Computed Muscle Control method to solve the muscle excitations for forward dynamics simulation, which provides a new way to deal with musculoskeletal problems from the perspective of task space. Additionally, Jäntsch, et al.[20]

established models of muscle force and actuator velocity, and then used the Dynamic Surface Control method and an adaptive neural network to compensate for the friction terms of joints and muscles, which effectively improved the trajectory tracking capability of a musculoskeletal robot. However, as the number of muscles increases and their arrangement becomes more diverse, it is more difficult to establish accurate musculoskeletal models. Thus model-based methods are not suitable for complex musculoskeletal robots. In contrast, model-free methods use data-driven approaches to guide robot learning without building realistic models. Advanced control theories are used in the research of musculoskeletal robot control. Combining control theory with the concept of motion primitives to improve skill learning capabilities has been widely studied[21–23]. Li, et al.[21] proposed a skill learning-based hierarchical control strategy, which learns motor skills from demonstrators by fusing dynamic motion primitives and Gaussian mixture models, and then ensures that the exoskeleton robot can complete complex interactive tasks by using an adaptive neural controller. Adaptive impedance control[24, 25] has achieved remarkable results in the compliant interaction between upper or lower limb exoskeleton robots and humans. Li, et al.[24] designed an optimal reference impedance model and proposed an adaptive neural network combined with a high-gain observer, which approximates robot dynamics and deadzone effects, drive the robot to track desired trajectories. In addition, deep reinforcement learning methods, such as deep deterministic policy gradient (DDPG)[26], trust region policy optimization (TRPO)[27], and proximal policy optimization (PPO)[28], were used to control the lower limb musculoskeletal model to walk in multiple directions with varying speeds[29, 30]. But model-free methods usually require a large amount of exploration and iteration to obtain the optimal control policy, resulting in low learning efficiency and poor generalization performance.

In recent years, the interdisciplinary research of information science and neuroscience has provided new ideas for motion learning and control of musculoskeletal systems, and the effectiveness of neural heuristic methods to improve motion performance has also been confirmed[31–38]. Based on the time-varying muscle synergy mechanism of the central nervous system, Chen and Qiao[33] constructed phasic and tonic muscle synergy models to characterize the features of muscle excitations, and used a radial basis function network to obtain muscle excitations signals by combining modulated muscle synergies according to different movement goals. Inspired by the gain primitive model of cortical network, Zhong and Wu[37] proposed a recurrent neural network modulated by gain primitives to achieve the control dimensionality reduction of musculoskeletal robots from the number of actuators to the number of primitives, where a parameter adaptation algorithm regulated by monoamine mechanism is applied to improve the learning efficiency of gain primitives, and the prior motion experiences and learned primitives are introduced to enhance motion generalization ability. Drawing on the speed-accuracy trade-off theory in neuroscience and behavioral sciences, Zhou, et al.[38] proposed a hierarchical movement learning framework, in which a basal ganglia network is modeled to realize adaptive motion planning on the basis of Fitts' Law, and an improved policy gradient algorithm is used to generate muscle excitations via muscle co-contraction policy. However, these methods still have some shortcomings. Firstly, some neural heuristic methods lack feedback control, and the utilization of environmental states needs to be strengthened. Secondly, the neural mechanism mainly originates from one brain region, and the cooperation of multiple brain regions needs to be considered to accomplish complex movements. Thirdly, the motion task of musculoskeletal robots is to reach from a fixed position to random positions, which needs to be generalized to the task of reaching from random starting positions to random target positions.

The solutions to the above issues are as follows. In an effort to accomplish the random reaching task of the end-effector, feedback of environment state needs to be added to the controller. In reinforcement learning, the widely used actor-critic architecture essentially belongs to feedback control. Correspondingly, in computational neuroscience, the basal ganglia are often modeled using the actor-critic method[39–41]. Thereby the basal ganglia serve as the primary controller. Then the complementary roles of different brain regions in motor learning and control are considered. The basal ganglia process the reward prediction error, calculated as the difference between the expected and received reward of an action, while the cerebellum is responsible for the sensory prediction error, calculated as the difference between the predicted and effective sensory feedback[42]. That is, neural processing in the basal ganglia is more relevant to the task, while that in the cerebellum is more relevant to environmental sensory.

There have been several studies combining the basal ganglia and cerebellum to control robotic systems. Dasgupta, et al.[43] constructed a reservoir actor-critic network of basal ganglia and an input correlation learning model of cerebellum, and then used a reward modulated heterosynaptic plasticity rule of thalamus to dynamically add the outputs of these two learning systems. Wang, et al.[44–46] used an actor-critic model and a motivated developmental network to establish models of basal ganglia and cerebellum respectively, in which radial basis function network and Q-learning are employed to be the actor and the critic respectively, and then used a thalamic model to combine the outputs from two models in an adaptive ratio. Ruan, et al.[47] modeled the cerebellum and basal ganglia using an actor and a critic model respectively, and established a thalamus model using a tropism mechanism, which serves as a relay bridge to cooperate with the information interaction between the cerebellum and basal ganglia in the motor cortex and participates in processing reward signals. To further study the combination scheme of these two brain areas, the integration model should be considered to be improved from the following aspects. Firstly, draw on the anatomical structures of brain regions to design the basal ganglia network and cerebellum network, so that these models are biologically interpretable and credible. Secondly, simulate the subcortical interconnections of two brain regions to design their communication, rather than just letting them work in parallel and then combining their outputs through the thalamus. Thirdly, introduce the neural mechanisms of brain areas in movement regulation to facilitate the exploration-exploitation trade-off during the learning process.

The motivation and purpose for this paper is how to propose a novel biological plausible integration model of these two brain areas and its learning rules, which can control the musculoskeletal robot to accomplish the random reaching task with high movement precision and fast learning efficiency. The main contributions of this article are as follows:

1) The bio-inspired integration model of basal ganglia and cerebellum is proposed, and the

convergence of the method is proved. The biologically plausible integration model is designed inspired by the anatomy of basal ganglia and cerebellum and their subcortical interconnections. Based on the neural mechanism of the basal ganglia, the basal ganglia network is divided into an actor network and a critic network, which are learned using the soft actor-critic method. The cerebellum network evaluates the quality of feature extraction in the basal ganglia actor network and updates the network weights of its feature layer. Then the cerebellum network weights are updated according to the difference of the optimization function of actor network parameters. By alternately using policy evaluation and policy improvement, it is demonstrated that the method can converge to the optimal policy.

2) The adaptive adjustment method for target entropy is proposed. Drawing on the neural relationship the entropy of action signals and the dopamine proportion in the basal ganglia, the formula for target entropy with respect to dopamine proportion is derived. This proportion as a hyperparameter changes exponentially with the relationship between target entropy and policy entropy, regulating the exploration-exploitation trade-off in motor learning.

3) The dopaminergic experience replay method is proposed. Following the principle of emphasizing the utilization of distinguished experience, the recent experience set and the optimal experience set are designed from the replay buffer respectively, and then a mini-batch of transitions for network updates are sampled from these two sets. By borrowing the dopamine proportion again, the ratio of different types of transitions is adjusted. The recent and optimal experiences are enhanced while older experiences are taken into account, which also promotes the exploration-exploitation trade-off.

The remainder of this paper is organized as follows. In Section 2, the structures and functions of the basal ganglia and cerebellum, as well as their interconnections, are introduced sequentially, which lays the foundation for proposing biological heuristic methods. In Section 3, the construction and learning rules of the bio-inspired integration model are described in detail, and then performance improvement methods are proposed, including adaptive adjustment of target entropy and dopaminergic experience replay. The experimental verification is carried out in Section 4. Differences from similar previous studies are discussed in Section 5. Finally, the full text is summarized in Section 6.

2 Preliminaries

2.1 Basal Ganglia

The basal ganglia, located at the base of the forebrain and the top of the midbrain, are a group of subcortical nuclei. The neural pathways of the basal ganglia are as follows[48, 49]. The striatum, as the largest structure in the basal ganglia, is the input nucleus of the basal ganglia. The substantia nigra pars reticulata (SNr) and the internal globus pallidus (GPi) usually work synergistically, and can be regarded as a complex, which are the outputs of the basal ganglia. The striatum receives signals from the cerebral cortex. The dorsal striatum (DS, consisting of the caudate nucleus and the putamen) then transmits signals to other parts of the basal ganglia through both direct and indirect pathways. In the direct pathway, the DS sends inhibitory signals to the SNr-GPi complex. In the indirect pathway, the DS firstly projects inhibitory signals to the external globus pallidus (GPe), which in turn sends inhibitory signals to the subthalamic nucleus (STN). The STN then relays excitatory signals to the SNr-GPi. Finally, the signals are output by the SNr-GPi to the thalamus, and then transmitted to the cerebral cortex and other brain regions. In addition, the ventral striatum (VS, consisting of the nucleus accumbens and the olfactory tubercle) receives dopaminergic input from the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA). The neural connections of the basal ganglia are shown in Figure 1.

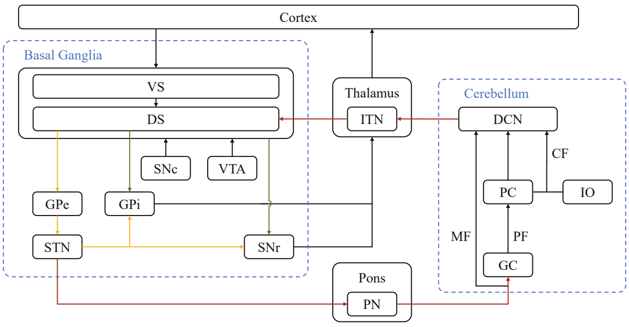


Figure 1 Neural circuits of basal ganglia, cerebellum, and their interconnections. Green and yellow lines indicate direct and indirect pathways of dorsal striatum respectively. Red lines indicate interconnections of basal ganglia and cerebellum. The investigation and summary of neural circuits and mechanisms are used to design network structures and learning rules

The basal ganglia are essential brain structures for motor control. Their main functions are as follows[50, 51]. Firstly, the basal ganglia play an important role in action selection. The basal ganglia make decisions about which of several possible actions to perform at a given time. They are able to select desired actions while suppressing unwanted potentially competing actions. Secondly, the basal ganglia have the function of motor learning. They are frequently modeled as reinforcement learning models in the motor learning research. In trial-and-error training, they achieve skill learning and motor adaptation based on reward prediction errors. In addition, they also have an important effect on motor tasks such as eye movements, voluntary movements, and executive functions, and have advanced cognitive functions such as working memory and emotion regulation.

2.2 Cerebellum

The cerebellum is located in the posterior cranial fossa. Its structure is as follows[52, 53]. The input signals to the cerebellum come from mossy fibers (MF) and climbing fibers (CF), and the sole output of the cerebellum is deep cerebellar nuclei (DCN). The MF encode sensory information such as environment and motion, which transmits signals to the DCN through two pathways. On the one hand, the MF project signals directly to the DCN. On the other hand, they also transmit signals to granule cells (GC), and then to Purkinje cells (PC) through parallel fibers (PF). The PC, as the only output of the cerebellar cortex, eventually send inhibitory signals to the DCN. The CF, originated from inferior olivary (IO) nucleus, mainly encodes error signals. They project these signals to the PC and DCN. At last, the signals output by the cerebellum are transmitted through the thalamus to the motor cortex or other brain regions, thereby affecting the firing of their neurons. The neural circuit of the cerebellum is shown in Figure 2.

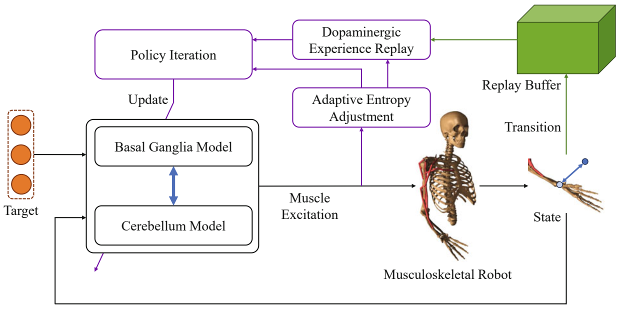


Figure 2 Framework of bio-inspired integration method for motion learning of musculoskeletal robot. The integration model of basal ganglia and cerebellum serves as the controller, generating muscle control signals to drive the movement of the musculoskeletal robot. The adaptive adjustment of target entropy trade-offs the exploration-exploitation of network learning. Dopaminergic experience replay samples the recent and optimal experiences for gradient update. Policy iteration combines the above modules and is proved to converge to the optimal policy

The cerebellum also plays an important role in motor control. With the assistance of the cerebellum, the precision and adaptability of movements can be significantly improved. Its main functions are described below[50, 54]. Firstly, the cerebellum is capable of motor correction. The cerebellum does not generate motor control commands, but instead regulates or compensates motor signals from other brain regions to effectively coordinate movement. Secondly, the cerebellum also has the function of motor learning. During trial-and-error training, the cerebellum continuously learns from sensory prediction errors, thereby improving the ability of motion evaluation and correction. Thirdly, the cerebellum also plays an important part in predicting movement results, coordinating voluntary movements, and maintaining posture balance. In addition, it also involves cognitive functions such as language, emotion, and attention.

2.3 Interconnections of Basal Ganglia and Cerebellum

The basal ganglia and cerebellum jointly regulate the processes of motion learning and control. The traditional view is that the basal ganglia and cerebellum transmit their respective outputs to the same cerebral cortical area for information interaction through different corticothalamic pathways. So they used to be considered as two separate subcortical systems with only cortical-level communication. However, recent studies have found that there also exist subcortical interactions between them.

Anatomically, a series of studies using retrograde transneuronal transport of rabies virus revealed two disynaptic pathways between the basal ganglia and cerebellum[55–57]. On the one hand, the STN of the basal ganglia firstly transmits signals to the pontine nuclei (PN), and then the PN transmits signals to the cerebellar cortex. On the other hand, the dentate nucleus (DN) of the cerebellum projects to the intralaminar thalamic nuclei (ITN), which in turn projects to the putamen and caudate of the basal ganglia. Actually, the DN is the largest nucleus in the DCN, and the putamen and caudate make up the DS. The above interaction mode is shown in Figure 1.

Moreover, the complementary roles of the basal ganglia and cerebellum in motion regulation can also demonstrate the above connections, which bring reference and inspiration to network modeling and learning rules. Firstly, the GC in the cerebellar cortex, in addition to being traditionally thought to encode the sensorimotor context, also encode information about reward expectation and are involved in reward-based learning[58]. This finding corroborates with the pathway from the STN to the cerebellar cortex. Secondly, the disynaptic pathway projecting from the DCN to the DS may convey cerebellar predictions to the basal ganglia with short latency. This type of communication not only engages the striatum in sensorimotor adaptation[56], but also facilitates more rapid coordination between these two brain regions to cope with complex environments[59, 60]. Meanwhile, this pathway can guide changes in striatal plasticity[59]. The above research reveals the significance of this pathway. Thirdly, dopaminergic signals from the basal ganglia, which are associated with predicting rewards, not only act on the VS, but also project to the nucleo-olivary pathway. In this way, the IO encode teaching signals from the basal ganglia to drive motor learning in the cerebellum[42, 61].

The basal ganglia and cerebellum coordinate with each other to collectively improve the quality and adaptability of movements[42]. Inspired by their structures, functions and interconnections, a neural heuristic control model will be constructed.

3 Methods

3.1 System Framework

Borrowing ideas from neural mechanisms such as the mutual communication between the basal ganglia and the cerebellum, dopaminergic modulation of the target policy entropy and the experience replay, the bio-inspired integration model of the basal ganglia and the cerebellum is proposed, which effectively improves the learning efficiency and motion adaptability of the musculoskeletal robot.

The system framework of this study is shown in Figure 2. The process of system operation is as follows. In each episode, a target position is randomly given. At each timestep t, the coordinates of the target are combined with other current environmental information, such as the joint angles and joint angular velocities of the musculoskeletal arm, the energy of muscle signals, etc., to compose the observation state $\boldsymbol{s}_t$. The bio-inspired model acts as a controller that generates the muscle command $\boldsymbol{a}_t$ based on the current state $\boldsymbol{s}_t$. Then this command $\boldsymbol{a}_t$ is projected to the cerebral cortex through the thalamus to drive the end-effector of the musculoskeletal model to the target position. One timestep later, the reward signal $r\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right)$ and the next state $\boldsymbol{s}_{t+1}$ are obtained. Together with the current state $\boldsymbol{s}_t$ and muscle command $\boldsymbol{a}_t$, they form a transition $(\boldsymbol{s}_t, \boldsymbol{a}_t, r(\boldsymbol{s}_t, \boldsymbol{a}_t), \boldsymbol{s}_{t+1})$ and are stored in the replay buffer $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\boldsymbol{s}_t, \boldsymbol{a}_t, r(\boldsymbol{s}_t, \boldsymbol{a}_t), \boldsymbol{s}_{t+1})\}$. By using a batch of samples extract from the buffer, the stochastic gradients are calculated for the motion learning of the bio-inspired model. Furthermore, the adaptive regulation methods based on dopamine proportion are proposed. Based on the change of this proportion, the value of the target policy entropy and the samples of experience replay are dynamically adjusted to realize the exploration-exploitation trade-off in the motor learning process. Through the above approaches, the musculoskeletal model can be accurately controlled to reach any target point.

3.2 Modeling and Learning of Basal Ganglia Network and Cerebellum Network

A bio-inspired integration model of the basal ganglia and cerebellum is proposed to generate the muscle control signal $\boldsymbol{a}$ based on the state $\boldsymbol{s}$. In the following, the modeling methods of the basal ganglia network and the cerebellum network, the interconnection mode of the two networks, and the learning rules of the integration model are introduced in sequence, and then the convergence of network learning is proved.

The basal ganglia network is established based on the connection of nuclei in the basal ganglia. The modeling and learning of the basal ganglia network refer to the soft actor-critic (SAC) method[62, 63]. The neural argument for using SAC is as follows. Firstly, the basal ganglia and the actor-critic architecture share similarities[39–41]. DS, GPe, and STN focus on motion learning, while SNr and GPi are related to action selection and command output, so they have similar functions to the actor. VS is responsible for value estimation, while SNc and VTA are in charge of computing temporal difference (TD) errors, so they have similar functions to the critic. Secondly, the output of basal ganglia is the probability density function of action[64]. Thus, the algorithm needs to output stochastic policy instead of deterministic policy. Thirdly, entropy is often introduced as a measure of the probabilistic irregularity of basal ganglia neuron activity over timescales[64, 65]. Entropy reflects changes in dopamine proportion, which has important implications for the exploration-utilization tradeoff during motor learning. Therefore, based on the above biological mechanisms, the use of SAC is reasonable. $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$ and $Q_\theta(\boldsymbol{s},\boldsymbol{a})$ are used to denote the actor and critic of the basal ganglia network respectively, where ϕ and θ are network parameters. $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$ and $Q_\theta(\boldsymbol{s},\boldsymbol{a})$ are shown in Figure 3.

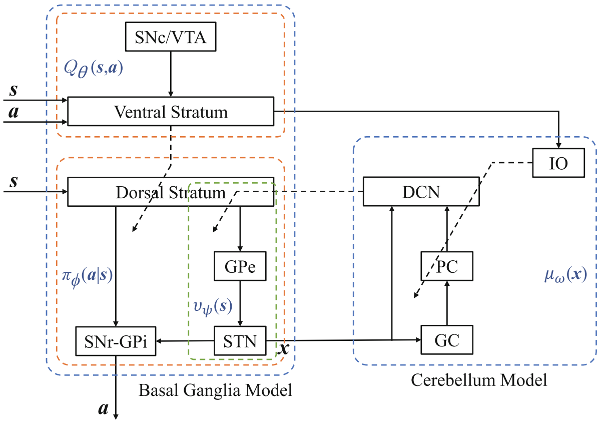


Figure 3 Network architecture of bio-inspired integration model of basal ganglia, cerebellum, and interconnections

The cerebellum network $\mu_\omega(\boldsymbol{x})$ is also designed in the same way as the connection of nuclei in the cerebellum, where ω represents the network parameters. In the cerebellum network, sensory signals are transmitted forward along MF-DCN pathway and MF-GC-PC-DCN pathway, and error signals are encoded in the IO to modify the network weights ω. $\mu_\omega(\boldsymbol{x})$ is shown in Figure 3.

Analogous to the information processing flow, the indirect pathway of $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$ can be divided into two stages, namely environmental feature extraction DS-GPe-STN and action selection STN-SNr/GPi[66]. The quality of feature extraction is related to the effect of the next action selection. Consequently, the cerebellum network $\mu_\omega(\boldsymbol{x})$ is introduced to monitor the sensorimotor context in order to improve the feature extraction capability of the basal ganglia network. According to the interconnection modes of the two brain areas, the communication methods of these networks are constructed as shown in Figure 3. The results of feature extraction are passed from the STN to the GC, and the outputs of the cerebellar evaluation is transmitted from the DCN to the DS. Let $\nu_\psi(\boldsymbol{s})$ represent the DS-GPe-STN network, where ψ is its parameters.

After setting up the network structures, the learning rules are designed as follows. The basal ganglia network is designed to learn a tractable policy $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$ and a soft Q-value function $Q_\theta(\boldsymbol{s},\boldsymbol{a})$. Part of the calculation rules for Q and π are derived from the SAC method[62, 63]. The training objective of π is to maximize both expected return and entropy:

$$J(\phi) = \max_{\pi_\phi} \mathbb{E}_{(\boldsymbol{s}_t,\boldsymbol{a}_t)\sim p_{\pi_\phi}} \left[\sum_{t=0}^{T} \gamma^t \left(r\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) + \alpha \mathcal{H}\left(\pi_\phi\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right)\right)\right], \tag{1}$$

where p_π is the marginal distribution following the policy π, $\gamma \in [0,1]$ is discount factor, $\alpha > 0$ is the temperature parameter, and $\mathcal{H}(\pi) = \mathbb{E}_\pi[-\ln \pi]$ is the policy entropy. According to the Bellman equation, the soft Q-value function $Q\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)$ is defined as follows:

$$Q\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) = r\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) + \gamma \mathbb{E}_{\boldsymbol{s}_{t+1}\sim p_\pi}\left[V\left(\boldsymbol{s}_{t+1}\right)\right]. \tag{2}$$

The relationship between the soft Q-value function $Q\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)$ and the soft state value function $V\left(\boldsymbol{s}_t\right)$ is as follows:

$$V\left(\boldsymbol{s}_t\right) = \mathbb{E}_{\boldsymbol{a}_t\sim\pi}\left[Q\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) - \alpha \ln \pi\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right]. \tag{3}$$

For each update timestep t, after the robot interacts with the environment, the basal ganglia critic network dominated by VS, namely $Q_\theta\left(\boldsymbol{s},\boldsymbol{a}\right)$, is firstly updated. The training objective of $Q_\theta\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)$ is to minimize the soft Bellman residual:

$$\begin{aligned} J_Q(\theta) &= \mathbb{E}_{(\boldsymbol{s}_t,\boldsymbol{a}_t)\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_\theta\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) - \widehat{Q}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)\right)^2\right] \\ &= \mathbb{E}_{(\boldsymbol{s}_t,\boldsymbol{a}_t)\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_\theta\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) - \left(r\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) + \gamma \mathbb{E}_{\boldsymbol{s}_{t+1}\sim p_\pi}\left[V_{\overline{\theta}}\left(\boldsymbol{s}_{t+1}\right)\right]\right)\right)^2\right], \end{aligned} \tag{4}$$

where $\widehat{Q}(\boldsymbol{s}_t,\boldsymbol{a}_t)$ is the target value of $Q_\theta(\boldsymbol{s}_t,\boldsymbol{a}_t)$, and $V_{\overline{\theta}}(\boldsymbol{s}_t) = \mathbb{E}_{\boldsymbol{a}_t\sim\pi_\phi}[Q_{\overline{\theta}}(\boldsymbol{s}_t,\boldsymbol{a}_t) - \alpha \ln \pi_\phi(\boldsymbol{a}_t|\boldsymbol{s}_t)]$ is implicitly parameterized by (3). $Q_{\overline{\theta}}(\boldsymbol{s}_t,\boldsymbol{a}_t)$ is the soft Q-value target function, and $\overline{\theta} \leftarrow \tau\theta + (1-\tau)\overline{\theta}$ is the exponentially moving average, which is beneficial to the stability of the training process. τ is the smoothing coefficient. The gradient of $J_Q(\theta)$ is estimated with an unbiased estimator:

$$\widehat{\nabla}_\theta J_Q(\theta)=\nabla_\theta Q_\theta\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)\left(Q_\theta\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)-\left(r\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)+\gamma\left(Q_{\overline{\theta}}\left(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1}\right)-\alpha\ln\pi_\phi\left(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1}\right)\right)\right)\right). \tag{5}$$

Thus, the parameters of $Q_\theta\left(\boldsymbol{s},\boldsymbol{a}\right)$ are updated as follows:

$$\theta\leftarrow\theta-\lambda_Q\widehat{\nabla}_\theta J_Q(\theta), \tag{6}$$

where λ_Q and the following λ_π, λ_ν, λ_μ, λ_α are the respective learning rates.

Then, the basal ganglia actor network dominated by DS, namely $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$, is updated. The training objective of $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$ is to minimize the Kullback-Leibler divergence:

$$\begin{aligned}\widetilde{J}_\pi(\phi)&=D_{\mathrm{KL}}\left(\pi_\phi\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\middle\|\frac{\exp\left(\frac{1}{\alpha}Q_\theta\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)\right)}{Z\left(\boldsymbol{s}_t\right)}\right)\\&=D_{\mathrm{KL}}\left(\pi_\phi\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\middle\|\exp\left(\frac{1}{\alpha}Q_\theta\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)-\ln Z\left(\boldsymbol{s}_t\right)\right)\right)\\&=\mathbb{E}_{\boldsymbol{s}_t\sim\mathcal{D},\boldsymbol{a}_t\sim\pi_\phi}\left[\ln\left(\frac{\pi_\phi\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)}{\exp\left(\frac{1}{\alpha}Q_\theta\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)-\ln Z\left(\boldsymbol{s}_t\right)\right)}\right)\right]\\&=\mathbb{E}_{\boldsymbol{s}_t\sim\mathcal{D},\boldsymbol{a}_t\sim\pi_\phi}\left[\ln\pi_\phi\left(\boldsymbol{a}|\boldsymbol{s}_t\right)-\frac{1}{\alpha}Q_\theta\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)+\ln Z\left(\boldsymbol{s}_t\right)\right],\end{aligned} \tag{7}$$

where $Z\left(\boldsymbol{s}\right)=\sum_{\boldsymbol{a}}\exp\left(\frac{1}{\alpha}Q\left(\boldsymbol{s},\boldsymbol{a}\right)\right)$ is the partition function to normalize the distribution. To simplify the calculation, multiply $\widetilde{J}_\pi(\phi)$ by α, and remove the constant $\ln Z\left(\boldsymbol{s}_t\right)$ that does not affect the gradient:

$$J_\pi(\phi)=\mathbb{E}_{\boldsymbol{s}_t\sim\mathcal{D},\boldsymbol{a}_t\sim\pi_\phi}\left[\alpha\ln\pi_\phi\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)-Q_\theta\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)\right]. \tag{8}$$

The approach to minimize $J_\pi(\phi)$ makes use of the reparameterization trick, which uses a neural network transformation f_ϕ to evaluate $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$:

$$\boldsymbol{a}=f_\phi\left(\boldsymbol{s},\varepsilon\right), \tag{9}$$

where ε is independent noise, sampled from some fixed distribution. In practice, actions are obtained by a squashed Gaussian policy:

$$\boldsymbol{a}=\tanh\left(\chi_\phi\left(\boldsymbol{s}\right)+\sigma_\phi\left(\boldsymbol{s}\right)\odot\varepsilon\right), \tag{10}$$

where χ_ϕ is the mean output by a Gaussian network, σ_ϕ is standard deviation derived from $\ln\sigma_\phi$ output by the Gaussian network, and $\varepsilon\sim\mathcal{N}(0,\mathbf{1})$. The reparameterization trick rewrites the expectation over actions as the expectation over noise, so that the expectation in $J_\pi(\phi)$ no longer depends on the policy parameters, which effectively reduces the variance of the gradient estimator. $J_\pi(\phi)$ is rewritten as follows:

$$J_\pi(\phi)=\mathbb{E}_{\boldsymbol{s}_t\sim\mathcal{D},\varepsilon_t\sim\mathcal{N}}\left[\alpha\ln\pi_\phi\left(f_\phi\left(\boldsymbol{s}_t,\varepsilon_t\right)|\boldsymbol{s}_t\right)-Q_\theta\left(\boldsymbol{s}_t,f_\phi\left(\boldsymbol{s}_t,\varepsilon_t\right)\right)\right], \tag{11}$$

where π_ϕ is implicitly defined in terms of f_ϕ. Then $J_\pi(\phi)$ is optimized with unbiased gradient estimation:

$$\widehat{\nabla}_\phi J_\pi(\phi) = \nabla_\phi \alpha \ln\left(\pi_\phi\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right) + \left(\nabla_{\boldsymbol{a}_t} \alpha \ln\left(\pi_\phi\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right) - \nabla_{\boldsymbol{a}_t} Q_\theta\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right)\right) \nabla_\phi f_\phi\left(\boldsymbol{s}_t, \varepsilon_t\right). \quad (12)$$

Thereby, the parameters of $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$ are updated to ϕ_π for the first time:

$$\phi_\pi \leftarrow \phi - \lambda_\pi \widehat{\nabla}_\phi J_\pi(\phi). \quad (13)$$

Next, the basal ganglia network interacts with the cerebellum network, and the network parameters are further updated. In the STN-GC pathway, state feature $\nu_\psi(\boldsymbol{s}_t)$ encoded in the STN is projected to the cerebellum network $\mu_\omega(\boldsymbol{x})$. The cerebellum network outputs $\mu_\omega(\nu_\psi(\boldsymbol{s}_t))$ after evaluating the quality of feature extraction. In the DCN-DS pathway, $\mu_\omega(\nu_\psi(\boldsymbol{s}_t))$ is projected to the DS as a loss to modify the network weights ψ in the feature extraction stage $\nu_\psi(\boldsymbol{s})$. Before updating ψ, save the complement ϕ_π of ψ in ϕ_π:

$$\phi'_\pi \leftarrow \phi_\pi \setminus \psi, \quad (14)$$

ψ is updated as follows:

$$\psi \leftarrow \psi - \lambda_\nu \widehat{\nabla}_\psi \mu_\omega(\nu_\psi(\boldsymbol{s}_t)). \quad (15)$$

Sequentially, the parameters of $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$ are updated to ϕ for the second time:

$$\phi \leftarrow \psi \cup \phi'_\pi. \quad (16)$$

The effect of two updates of $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$ is measured by using (11):

$$\Delta J_\pi = h\left(J_\pi(\phi)|_{\boldsymbol{a}_t=f_\phi(\boldsymbol{s}_t,\varepsilon_t)} - J_\pi(\phi_\pi)|_{\boldsymbol{a}_t=f_{\phi_\pi}(\boldsymbol{s}_t,\varepsilon_t)}\right), \quad (17)$$

where h is any squashing function. h limits ΔJ to a finite range to prevent its gradient from being too large. Here $h = \tanh(\cdot)$ is used. The calculation process of ϕ involves the cerebellum network parameters ω, so $J_\pi(\phi)$ is helpful for optimizing ω. Although ϕ_π bears no relation to ω, $J_\pi(\phi_\pi)$ can serve as a baseline to stabilize the learning process. Measurement error ΔJ_π is computed in the basal ganglia and then projected to the cerebellum. The IO encodes this error signal to guide the learning of $\mu_\omega(\boldsymbol{x})$:

$$\omega \leftarrow \omega - \lambda_\mu \widehat{\nabla}_\omega \Delta J. \quad (18)$$

In practical implementation, to avoid the problem of Q-value overestimation bias, two different Q-value networks $Q_{\theta_i}\left(\boldsymbol{s}, \boldsymbol{a}\right)$ with parameters θ_i are used to approximate $Q_\theta\left(\boldsymbol{s}, \boldsymbol{a}\right)$, and they are trained independently[67]. Their minimum is used for the gradient updates of (5) and (12):

$$Q_\theta\left(\boldsymbol{s}, \boldsymbol{a}\right) = \min_{i\in\{1,2\}} Q_{\theta_i}\left(\boldsymbol{s}, \boldsymbol{a}\right). \quad (19)$$

Correspondingly, the minimum of two different Q-value target networks $Q_{\overline{\theta}_i}\left(\boldsymbol{s}, \boldsymbol{a}\right)$ is also used to calculate $Q_{\overline{\theta}}\left(\boldsymbol{s}, \boldsymbol{a}\right)$ of (5):

$$Q_{\overline{\theta}}\left(\boldsymbol{s}, \boldsymbol{a}\right) = \min_{i\in\{1,2\}} Q_{\overline{\theta}_i}\left(\boldsymbol{s}, \boldsymbol{a}\right). \quad (20)$$

In the stage of motor evaluation, consistent with the performance of brain regions[42], the basal ganglia critic network $Q_\theta\left(\boldsymbol{s}, \boldsymbol{a}\right)$ and the cerebellum network $\mu_\omega(\boldsymbol{x})$ no longer participate in updating network parameters, and the basal ganglia actor network $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$ outputs control signals $\boldsymbol{a}_t$.

Last but not least, as the foundation of the bio-inspired integration model, it is proved that by repeatedly alternating policy evaluation and policy improvement, policy iteration ensures that any policy π will converge to the optimal policy π^*. See the Appendix for proofs.

Lemma 3.1 (Policy evaluation) *Let the soft Bellman backup operator in the basal ganglia model be $\mathcal{T}^\pi$, the soft state-action value function $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ with $\|\mathcal{A}\|_\infty < \infty$, and define $Q^{k+1} = \mathcal{T}^\pi Q^k$. Then given any policy π, the sequence Q^k will converge to Q^π as $k \rightarrow \infty$.*

Lemma 3.2 (Policy improvement) *Let $\pi_{\text{old}} \in \Pi$ and let π_{new} be the new policy optimized by the integration model. Then $Q^{\pi_{\text{new}}}\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right) \geq Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right)$ for all $\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right) \in \mathcal{S} \times \mathcal{A}$ with $\|\mathcal{A}\|_\infty < \infty$.*

Theorem 3.3 (Policy iteration) *Through repeatedly applying the policy evaluation and policy improvement in the integration model, any policy $\pi \in \Pi$ will converge to an optimal policy π^* such that $Q^{\pi^*}\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right) \geq Q^\pi\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right)$ for all $\pi \in \Pi$ and $\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right) \in \mathcal{S} \times \mathcal{A}$, assuming $\|\mathcal{A}\|_\infty < \infty$ and reward is bounded.*

3.3 Adaptive Adjustment of Target Entropy

For this bio-inspired integration model, it is crucial to realize the exploration-exploitation trade-off of the learning process. By further analyzing (1), the two variables, namely α and $\mathcal{H}(\pi)$, have an important impact on the learning process, and determine different behavior patterns of the robot. On the one hand, smaller α or $\mathcal{H}(\pi)$ corresponds to more exploitation. If these two parameters are too small, the basal ganglia actor network $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$ will be more inclined to learn greedy deterministic policy and thus the network learning will fall into a local optimal. In extreme cases, as $\alpha \rightarrow 0$, the maximum entropy term $\mathcal{H}(\pi)$ becomes useless, and the problem of (1) gradually degenerates into the standard maximum expected return objective. On the other hand, larger α or $\mathcal{H}(\pi)$ corresponds to more exploration. These large parameters make the policy behavior too random throughout the learning process, which is either not ideal in most situations.

In order to achieve an exploration-exploitation balance, the method to automatically adjust the regularization parameter α is introduced[63]. Formally, the objective of (1) is transformed into a conditional optimization problem, where the expected return is maximized and the policy satisfies the minimum entropy constraint:

$$\begin{aligned}
&\max_{\pi_0,\cdots,\pi_T} \mathbb{E}_{p_\pi}\left[\sum_{t=0}^{T}\gamma^t r\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right)\right] \\
&\text{s.t.} \quad \mathcal{H}(\pi_t) \geq \mathcal{H}^*, \quad \forall t,
\end{aligned} \tag{21}$$

where $\mathcal{H}(\pi_t) = \gamma^t \mathbb{E}_{(\boldsymbol{s}_t, \boldsymbol{a}_t)\sim p_\pi}\left[-\ln\left(\pi_t\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right)\right]$, and $\mathcal{H}^*$ is the target entropy, i.e., the desired minimum policy entropy. Here let $\mathcal{H}^* = -|\mathcal{A}|$, namely the negative dimension of the action space $\mathcal{A}$. The above problem is solved as follows:

$$J(\alpha) = \mathbb{E}_{\boldsymbol{a}_t \sim \pi_\phi} \left[-\alpha \ln \pi_\phi \left(\boldsymbol{a}_t | \boldsymbol{s}_t \right) - \alpha \mathcal{H}^* \right]. \tag{22}$$

Hence, α is updated as follows:

$$\alpha \leftarrow \alpha - \lambda_\alpha \widehat{\nabla}_\alpha J(\alpha). \tag{23}$$

The SAC method with automating adjustment of α shows great performance in most challenging tasks[63, 68].

This method still has some deficiencies that need to be further improved. This method focuses on the adjustment of α, but ignores the regulation of $\mathcal{H}^*$. The Lagrange multiplier method sets the target entropy $\mathcal{H}^*$ to a constant when solving the dual problem of constrained optimization. This leads to the fact that the policy entropy $\mathcal{H}(\pi_t)$ obtained is not the maximum value, but only approaches the constant target entropy $\mathcal{H}^*$. This is inconsistent with the optimization objective of maximizing policy entropy in (1). For the sake of solving the above issue, a bio-inspired adaptive adjustment rule of target entropy $\mathcal{H}^*$ modulated by dopamine ratio ρ is proposed.

Neuroscience research has found that the activity levels of dopamine in the basal ganglia can affect the probability distribution of output signal, which plays a significant role in regulating the exploration-utilization trade-off in the learning process. The relationship between the entropy of the basal ganglia output signal and the dopamine proportion can be described by an affine function[64], which inspires the way to establish the formula for the two.

The method of solving the upper and lower bounds of the target entropy $\mathcal{H}^*$ is as follows. For the purpose of network optimization, each control signal $\boldsymbol{a}$ in the action space $\mathcal{A}$ is normalized to the $[-1, 1]$ interval. When the robot interacts with the environment, these actions $\boldsymbol{a}$ should be rescaled to the $[0, 1]$ interval to control the musculoskeletal model normally. Such a box-constrained space is denoted as $\mathcal{A} = \mathrm{Box}\left([-1, 1], |\mathcal{A}|\right)$. As recommended in [63], the lower bound is $\mathcal{H}^*_{\min} = -|\mathcal{A}|$. The upper bound $\mathcal{H}^*_{\max}$ occurs when all actions are uniformly sampled within the action space $\mathcal{A}$, and thus the corresponding stochastic policy function $\pi_{\max}\left(\boldsymbol{a}|\boldsymbol{s}\right)$ is as follows:

$$\pi_{\max}\left(\boldsymbol{a}|\boldsymbol{s}\right) = 2^{-|\mathcal{A}|}. \tag{24}$$

Then the maximum value of the target entropy is calculated:

$$\begin{aligned} \mathcal{H}^*_{\max} &= \mathbb{E}_{(\boldsymbol{s},\boldsymbol{a}) \sim p_\pi} \left[-\ln \left(\pi_{\max}\left(\boldsymbol{a}|\boldsymbol{s}\right) \right) \right] \\ &= -\int_{\mathcal{A}} \pi_{\max}\left(\boldsymbol{a}|\boldsymbol{s}\right) \ln \left(\pi_{\max}\left(\boldsymbol{a}|\boldsymbol{s}\right) \right) \mathrm{d}\boldsymbol{a} \\ &= -\int_{\mathcal{A}} 2^{-|\mathcal{A}|} \ln 2^{-|\mathcal{A}|} \mathrm{d}\boldsymbol{a} \\ &= -2^{-|\mathcal{A}|} \ln 2^{-|\mathcal{A}|} \cdot (1 - (-1))^{|\mathcal{A}|} \\ &= |\mathcal{A}| \ln 2. \end{aligned} \tag{25}$$

Therefore, the function of the target entropy $\mathcal{H}^*$ with respect to the dopamine proportion ρ, as shown in Figure 4, is obtained as follows:

$$\mathcal{H}^*(\rho) = -\rho|\mathcal{A}|\ln 2\mathrm{e} + |\mathcal{A}|\ln 2, \quad \rho \in [0,1]. \tag{26}$$

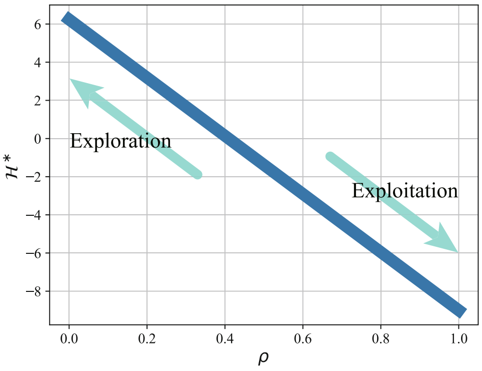


Figure 4 Target entropy function $\mathcal{H}^*$ with respect to dopamine proportion ρ. According to neuroscience research, their relationship conforms to an affine function. After determining the maximum and minimum values of the target entropy, the analytical formula of the target entropy function can be obtained. If the entropy is large, the learning tends to explore; On the contrary, the learning tends to exploit. Here let $|\mathcal{A}| = 9$

The algorithm of adaptive adjustment of target entropy is described as follows. For each update timestep t, it is judged whether $|\mathcal{H}^*(\rho) - \mathcal{H}(\pi_t)| \leq \xi$ is satisfied, where $\xi > 0$ is the threshold. If this condition is met, the dopamine proportion ρ is updated while ensuring that it does not exceed the maximum value:

$$\rho \leftarrow \min\{\rho \cdot \iota, 1\}, \tag{27}$$

where ι is the exponentially increasing factor. To avoid the singularity of adjustment, let the initial dopamine proportion ρ_{init} be 0.01. In addition, with the aim of making $\mathcal{H}^*(\rho)$ change smoothly rather than always being stuck at a certain value, the above update process is forced to be executed once after κ timesteps since the last update.

Under the action of this method, combined with (23), the policy entropy $\mathcal{H}(\pi)$ approaches the target entropy $\mathcal{H}^*$ continuously, and then correspondingly $\mathcal{H}^*$ decreases gradually, which in turn makes $\mathcal{H}(\pi)$ decrease along with it. In the early stage of training, $\mathcal{H}(\pi)$ is relatively large, making policy exploration more random. On the contrary, in the later stage of training, $\mathcal{H}(\pi)$ is relatively small, and the stability policy is gradually obtained. Thus, the exploration-exploitation trade-off is achieved, which also improves the performance of motor learning.

3.4 Dopaminergic Experience Replay

The bio-inspired integration model, as a type of off-policy reinforcement learning, requires the use of effective experience replay. Experience replay is a replay memory technique, where the experience $(\boldsymbol{s}_t, \boldsymbol{a}_t, r(\boldsymbol{s}_t, \boldsymbol{a}_t), \boldsymbol{s}_{t+1})$ of the agent is stored in a replay buffer $\mathcal{D}$ at each timestep t, and then a mini-batch of experiences are sampled from this buffer $\mathcal{D}$ with certain strategies for updating the learning rules. It is a valid mechanism to improve sample efficiency and algorithm stability[69]. Instead of uniform sampling that treats all past transitions equally, better sample and utilization of valuable transitions in the buffer is conducive to the performance improvement of reinforcement learning[70, 71].

Dopamine has long been implicated in working memory. Dopaminergic levels can intervene and improve working memory, thereby affecting the functional activity in basal ganglia[72, 73]. Drawing on the above neural mechanism, the Dopaminergic experience replay (DAER) method is proposed, where the experience replay is modulated by the dopamine proportion ρ. The core idea of DAER is that firstly the recent experience set $\mathcal{B}_r$ and the optimal experience set $\mathcal{B}_o$ are constructed from $\mathcal{D}$, and secondly mini-batch transitions are sampled from the two sets for gradient update according to the proportion allocated by ρ. Here, the function symbol $g_s\left(\mathcal{S}, n_{\mathcal{S}}\right)$ is defined as sampling $n_{\mathcal{S}}$ transitions from the set $\mathcal{S}$ with a strategy S.

With the increase of learning timesteps, the motion performance of the robot is gradually improved. So the more recent experiences are of higher value. Due to the poor learning effect at the beginning, utilization of older experiences should be reduced. Thus, the most recent experiences are sampled to construct $\mathcal{B}_r$:

$$\mathcal{B}_r = g_r\left(\mathcal{D}, \min\left\{\left\lfloor n_{\mathcal{D}} \cdot \rho\right\rfloor, n_r\right\}\right), \tag{28}$$

where $n_{\mathcal{D}}$ is the capacity of $\mathcal{D}$, $n_r < n_{\mathcal{D}}$ is the capacity of $\mathcal{B}_r$, and $\lfloor x\rfloor = \max\{z \in \mathbb{Z} | z \leq x\}$ is the floor function. As the learning process progresses, ρ gradually approaches 1, so $\mathcal{B}_r$ gradually stores all transitions. But by adding the restriction n_r, the oldest samples are always excluded.

In order to further take advantage of experiences with excellent performance in recent moments, the optimal experience set $\mathcal{B}_o$ is constructed:

$$\mathcal{B}_o = g_o\left(g_r\left(\mathcal{D}, 2m\right), m\right), \tag{29}$$

where m is the number of mini-batch samples. The construction of $\mathcal{B}_o$ is divided into two steps. Firstly, the most recent $2m$ transitions are sample from $\mathcal{D}$, that is, $\mathcal{B}_r' = g_r\left(\mathcal{D}, 2m\right)$. Secondly, the optimal m transitions are selected by sorting these $2m$ transitions, that is, $\mathcal{B}_o = g_o\left(\mathcal{B}_r', m\right)$. The basis for sorting is the absolute TD error of a transition, denoted as $|\delta|$. Since two Q-value networks are used, $|\delta|$ is obtained by calculating the average absolute TD error of two networks:

$$\begin{aligned} |\delta| &= \frac{1}{2}\sum_{i=1}^{2}\left|Q_{\theta_i}\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right) - \widehat{Q}\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right)\right| \\ &= \frac{1}{2}\sum_{i=1}^{2}\left|Q_{\theta_i}\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right) - \left(r\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right) + \gamma\left(Q_{\overline{\theta}}\left(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}\right) - \alpha \ln \pi_{\phi}\left(\boldsymbol{a}_{t+1} | \boldsymbol{s}_{t+1}\right)\right)\right)\right|. \end{aligned} \tag{30}$$

At last, a total of m transitions are uniformly sampled from $\mathcal{B}_r$ and $\mathcal{B}_o$ in proportion to ρ:

$$\mathcal{B}_m = g_u\left(\mathcal{B}_r, \left\lfloor m \cdot \left(1 - \frac{\rho}{\mathit{\Gamma}}\right)\right\rfloor\right) \cup g_u\left(\mathcal{B}_o, \left\lceil m \cdot \frac{\rho}{\mathit{\Gamma}}\right\rceil\right), \tag{31}$$

where $\mathit{\Gamma} > 1$ is a constant for regulating the ratio, and $\lceil x\rceil = \max\{z \in \mathbb{Z} | z \geq x\}$ is the ceiling function.

The advantages of DAER are analyzed as follows. Firstly, the recent and optimal experiences are sampled with a higher probability than older ones. Assuming uniform sampling is used alone, since all experiences have the same probability of being sampled, older experiences will be sampled more times than newer ones, which is not desirable. As the agent continues to explore the action space and improve its action policy, newer data, which contain more valuable information than old ones, deserves enhanced use. Secondly, older experience still has a certain probability of being extracted. As ρ increases, $\mathcal{B}_r$ gradually stores stores most of transitions of $\mathcal{D}$, so older data also have a chance to be sampled. This ensures that the integration model does not only estimate the value functions from recent transitions and avoids overfitting. Thirdly, this method facilitates the exploration-exploitation trade-off for the integration model. Considering (31), when ρ is smaller, $\mathcal{B}_r$ provides more experiences, allowing the agent to explore more different states and actions; when ρ is larger, the sample size provided by $\mathcal{B}_o$ gradually increases, strengthening the agent's exploitation of superior experiences to achieve desired motion. Therefore, by using DAER in network learning rules, the learning efficiency and generalization performance of the integration model will be further improved.

4 Experiments

4.1 Experiment Setup

The bio-inspired integration model is used for motion learning and control of a musculoskeletal robot. The biomechanical robotic system is implemented in MuJoCo[74], which has functions of robot kinematics and dynamics simulation and provides tools for modeling biological muscles. The robot system is built by simulating the arrangement of bones and muscles in the human upper limb, as shown in Figure 5(a). Nine muscles are used here, each of which is an actuator. Movement is generated by these muscles pulling rigid bones through tendon attachment points, where one degree of freedom is set in the shoulder joint and one at the elbow joint. The actual force of each muscle is

$$F_M = F_0 \cdot \left(act \cdot F_L(L) \cdot F_V(V) + F_P(L) \right), \tag{32}$$

where the constant F_0 is the peak active force at zero velocity, F_L is the active force function of scaled muscle length L, F_V is the active force function of scaled muscle velocity V, F_P is the passive force function that is always present regardless of activation, and act is the muscle activation signal. F_L, F_V, and F_P are all complex nonlinear functions[16]. Therefore, the musculoskeletal robot is a strongly nonlinear system. It is very challenging that how to control such a robot to achieve high-precision and high-robust motion tasks.

The experimental task is that the bio-inspired model generates a set of sequential muscle excitation signals, which can drive the end-effector of the upper limb musculoskeletal robot from random starting locations to random target positions with high precision, as shown in Figure 5(b). The range of motion is limited to a circle $\mathcal{O}$ in the vertical plane. The center of the circle $\mathcal{O}$ is $P_0 = (0.177 \text{ m}, 0.461 \text{ m})$, which is the middle of the wrist when the shoulder angle is $-10°$ and the elbow angle is $90°$. The radius of the circle $\mathcal{O}$ is $r = 0.14$ m, which is the farthest movement distance of the end-effector.

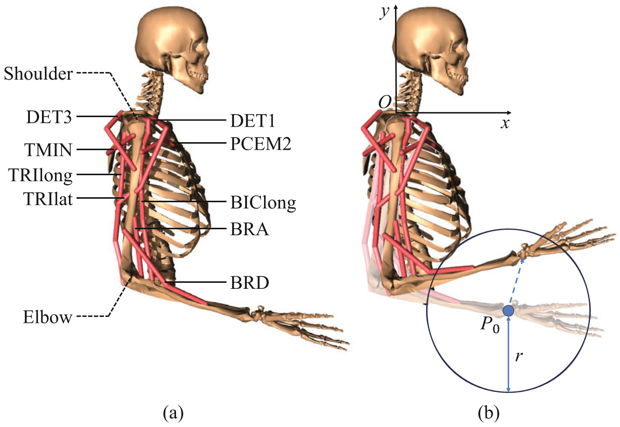


Figure 5 Experimental platform for musculoskeletal robot. (a) The musculoskeletal robot contains nine muscle actuators and two degrees of freedom. (b) The experimental task is to control the end-effector of the musculoskeletal robot from random starting locations to random target positions within a circle

4.2 Random Reaching Task

Inspired by the anatomical structures and interconnection manner of the basal ganglia and the cerebellum, the basal ganglia actor network $\pi_\phi(\boldsymbol{a}|\boldsymbol{s})$, the basal ganglia critic network $Q_\theta(\boldsymbol{s},\boldsymbol{a})$, and the cerebellum network $\mu_\omega(\boldsymbol{x})$ are constructed as Figure 3, and their learning rules are proposed. Each type of nucleus corresponds to a layer of network. The parameters of these networks are shown in Table 1. For each episode of movement, the initial coordinates of the end-effector and the coordinates of the motion target are randomly given inside the circle $\mathcal{O}$. The maximum duration of each episode is 500 ms, where each environment timestep is 1 ms. For each timestep t, the environment state $\boldsymbol{s}_t$, as the input of π_ϕ, contains the joint angles $\boldsymbol{q}_t$ and joint angular velocities $\dot{\boldsymbol{q}}_t$ of the robot, the coordinates of the target point $\boldsymbol{x}^*$ and the end-effector $\boldsymbol{x}_t$, the vector between the target point and the end-effector $\boldsymbol{x}^*-\boldsymbol{x}_t$, and the energy of muscle control signals $\|\boldsymbol{a}_t\|_2^2$. According to the state $\boldsymbol{s}_t$, π_ϕ outputs the action command $\boldsymbol{a}_t$ to control the robot to interact with the environment, thereby obtaining a transition $(\boldsymbol{s}_t,\boldsymbol{a}_t,r(\boldsymbol{s}_t,\boldsymbol{a}_t),\boldsymbol{s}_{t+1})$. $\boldsymbol{a}_t$ is a vector containing 9 elements, corresponding to the excitation signals of 9 muscles. The reward signal $r(\boldsymbol{s}_t,\boldsymbol{a}_t)$ is designed as follows:

$$r_t=-\eta_1\left\|\boldsymbol{x}^*-\boldsymbol{x}_t\right\|_2-\eta_2\left\|\boldsymbol{v}_t\right\|_2-\eta_3\left\|\boldsymbol{a}_t\right\|_2^2-\eta_4\zeta, \tag{33}$$

where $[\eta_1,\eta_2,\eta_3,\eta_4]=[10,1,0.01,1]$. $\boldsymbol{v}_t$ is the velocity of the end-effector. ζ denotes the additional reward when the end-effector reaches the target point, which is beneficial for improving the performance of the motion reaching task:

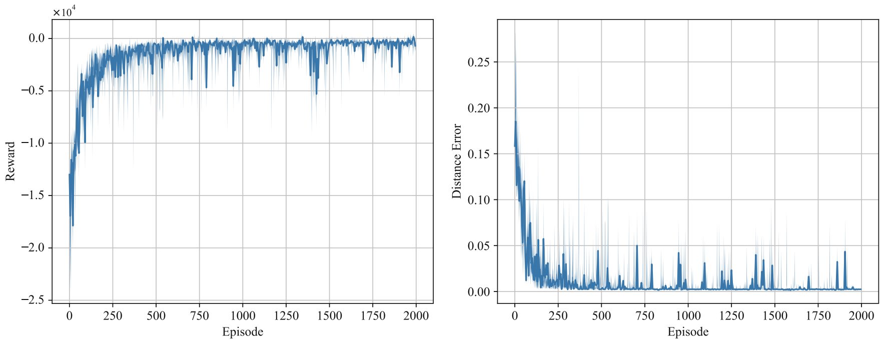
$$\zeta=\begin{cases}-1000, & \text{if } \|\boldsymbol{x}_t-\boldsymbol{x}^*\|\le\delta,\\ 0, & \text{if } \|\boldsymbol{x}_t-\boldsymbol{x}^*\|>\delta,\end{cases} \tag{34}$$

where $\delta = 3\,\mathrm{mm}$ is the distance threshold. This transition is then pushed to the repaly buffer $\mathcal{D}$, mini-batch of which are sampled through the dopaminergic experience replay method for unbiased gradient updating of network weights. Meanwhile, the automating entropy adjustment mechanism also plays a role to modulate network learning.

Table 1 Parameters of bio-inspired integration model

Parameter	Symbol	Value
Number of network hidden units	N_h	256
Adam learning rate	λ_Q, λ_π, λ_ν, λ_μ, λ_α	3×10^{-4}
Discount factor	γ	0.99
Target smoothing coefficient	τ	0.005
Dopamine proportion increasing factor	ι	1.001
Entropy threshold	ξ	0.01
Timestep threshold	κ	2000
Replay buffer size	$n_{\mathcal{D}}$	10^6
Recent buffer size	n_r	10^5
Sampling ratio	$\mathit{\Gamma}$	3
Minibatch size	m	256
Target update interval	n_u	1
Gradient steps	n_g	1

After training, the reward curve is drawn as shown in Figure 6(a). For special attention whether the robot reaches the target point, the distance error at the end of each episode is shown in Figure 6(b). Therefore, the integration model can control the musculoskeletal robot to accomplish the motion goal from random starting points to random target points. To evaluate the adaptive adjustment effect of policy entropy, the change in entropy and dopamine ratio are shown in Figure 7(a) and Figure 7(b) respectively. The policy entropy $\mathcal{H}(\pi)$ follows the target entropy $\mathcal{H}^*(\rho)$ from the maximum to the minimum gradually, thus ensuring the exploration-exploitation trade-off of model learning.



(a) (b)

Figure 6 Learning curves of bio-inspired integration model. (a) and (b) are the curves of reward and distance error during the learning process respectively

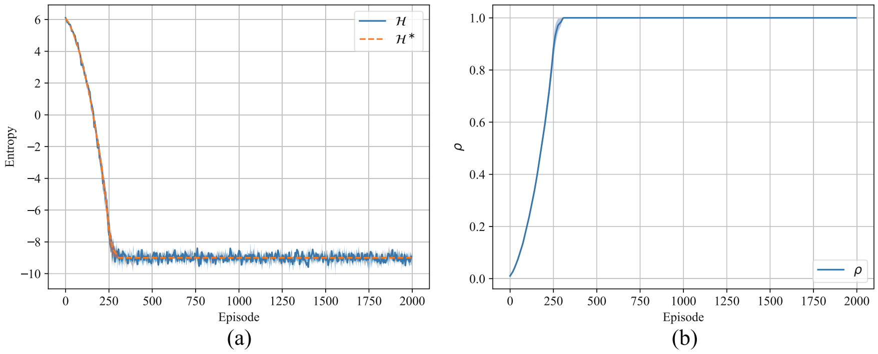


Figure 7 Adaptive adjustment process of target entropy. (a) Orange and blue lines show the changes in target entropy and policy entropy respectively. (b) The change in dopamine proportion is shown. Corresponding to the curves of reward and distance error in Figure 6, in the early stage of network learning, the dopamine proportion and entropy change dynamically; When learning converges, the dopamine proportion and entropy do not change substantially

Next, the effect of movement learning is evaluated. 20 pairs of starting positions and target positions are randomly set inside the circle $\mathcal{O}$, and the trained network π_ϕ is used to generate the control command $\boldsymbol{a}$. The motion trajectories of the robot end-effector are shown in Figure 8, and the average movement distance error is $2.147 \pm 0.176\,\text{mm}$. Hence, the bio-inspired model can control the musculoskeletal robot to accomplish movements with high precision and high generalization.

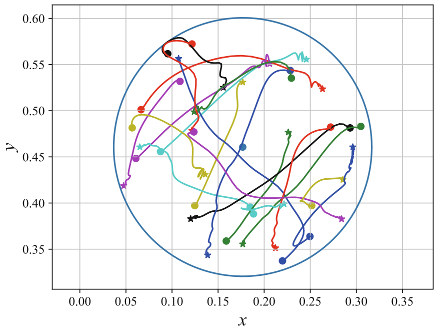


Figure 8 Motion trajectories of end-effector from random starting positions to random target positions during the evaluation process. Circles represent starting positions, stars represent target positions, and the same color indicates one movement

In order to better analyze the learning ability of the proposed method, comparison experiments are carried out between this bio-inspired method and other typical reinforcement learning methods, such as DDPG[26], PPO[28], and SAC[62, 63]. They have been proven to achieve good results in controlling complex robotic systems. The comparison experiment is still the task of controlling the end-effector of this musculoskeletal robot to reach arbitrary positions within the circle $\mathcal{O}$. The design of comparison methods, such as action space, environment space, reward function, etc., are consistent with the bio-inspired model mentioned above. The model parameters of DDPG, PPO, and SAC are shown in Tables 2, 3, and 4, respectively. The experimental results on rewards and distance errors are shown in Figure 9. The bio-inspired model evaluates not only state-action information by using the basal ganglia network, but also state feature extraction by using the cerebellum network. Besides it also realizes the exploration-exploitation trade-off of motor learning in terms of both policy entropy and experience replay with the help of dopaminergic modulation. Consequently, compared with DDPG, PPO, and SAC, the bio-inspired integration model can achieve higher precision motion tasks with fewer iterations and more stable convergence.

Table 2 Parameters of deep deterministic policy gradient

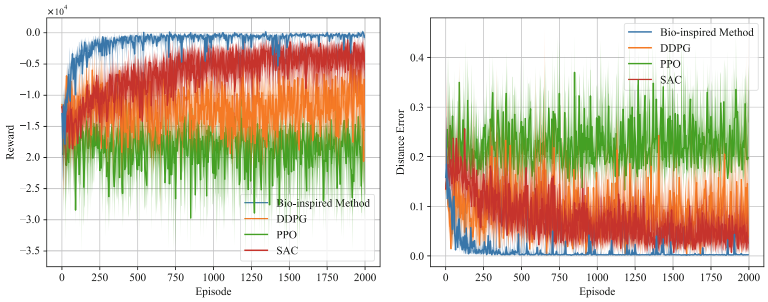
Parameter	Value
Number of network hidden layers	2
Number of network hidden units	256
Adam learning rate	3×10^{-4}
Discount factor	0.99
Target smoothing coefficient	0.005
Exploration noise	$\mathcal{N}(0, 0.1^2)$
Replay buffer size	10^6
Minibatch size	256

Table 3 Parameters of proximal policy optimization

Parameter	Value
Number of network hidden layers	2
Number of network hidden units	256
Adam learning rate	3×10^{-4}
Discount factor	0.99
GAE parameter	0.95
Number of epochs	10
Clip ratio	0.2
Minibatch size	256

Table 4 Parameters of soft actor critic

Parameter	Value
Number of network hidden layers	2
Number of network hidden units	256
Adam learning rate	3×10^{-4}
Discount factor	0.99
Target smoothing coefficient	0.005
Target entropy	-9
Replay buffer size	10^6
Minibatch size	256
Target update interval	1
Gradient steps	1

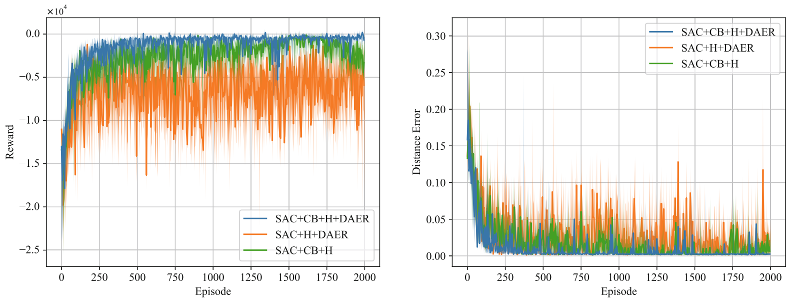


(a) (b)

Figure 9 Comparison of different learning methods for musculoskeletal robot. (a) and (b) are the curves of reward and distance error during the learning process respectively. Compared with other reinforcement learning methods, the bio-inspired integration model has faster convergence speed and higher stability, and its motion error is smaller after learning converges

4.3 Ablation Study

In the bio-inspired integration model, the basic component is the basal ganglia network (abbreviated as BG), others contain the cerebellum network (abbreviated as CB), adaptive adjustment of target entropy (abbreviated as H), dopaminergic experience replay (abbreviated as DAER). Thus it needs to be examined that the individual contribution of each component to performance. In the ablation study of the integration model, CB and DAER are removed separately. But H cannot be removed alone because the change of ρ in DAER depends on H. If H is removed, DAER cannot be used. The experiment tasks are described in Subsection 4.2. The results are shown in Figure 10. When CB or DAER is removed, the average performance decreases dramatically and fluctuates widely. With the joint promotion of all components, the learning process of the bio-inspired integration model converges faster, becomes more stable, and provides more rewards. Therefore, all the components are of great significance to improving motion performance.

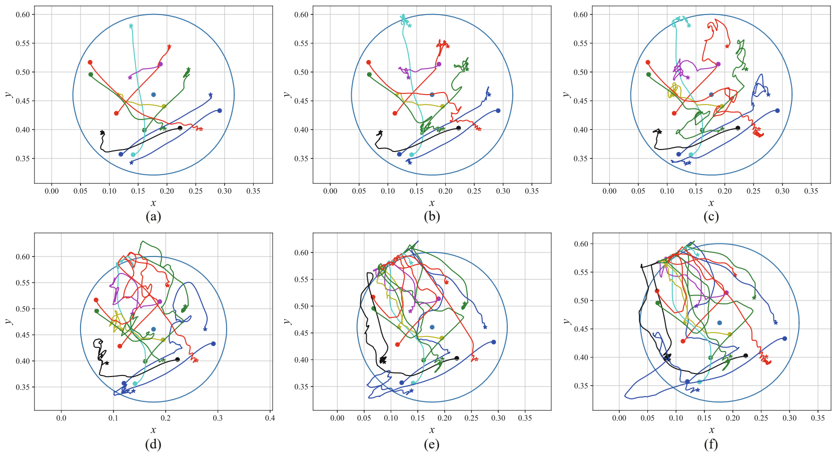


(a) (b)

Figure 10 Ablation study of bio-inspired integration model for musculoskeletal robot. (a) and (b) are the curves of reward and distance error during the learning process respectively. If the CB or DAER component is removed, the performance in terms of stability and motion accuracy will be significantly reduced. Ablation experiments demonstrate that all components of the integration model are necessary

4.4 Robustness of Controller

In the actual environment, the controller has a high probability of suffering from noise disturbances from other electronic devices. Thereby it is necessary and meaningful to study the anti-interference ability of the bio-inspired model. The noise is added to the control commands output by the basal ganglia action network π_ϕ. The noise signal obeys a uniform distribution. The noise duration is from 10 ms to 100 ms. As shown in Figure 11, (a) is the case of no noise, and (b)–(f) correspond to noise distributions of $\mathcal{N}(0, 0.5)$, $\mathcal{N}(0, 0.75)$, $\mathcal{N}(0, 1)$, $\mathcal{N}(0, 1.5)$, and $\mathcal{N}(0, 2)$ respectively. The average distance errors of (a)–(f) are 2.372 ± 0.123 mm, 2.271 ± 0.014 mm, 3.148 ± 0.818 mm, 2.368 ± 0.144 mm, 4.706 ± 2.295 mm, and 5.547 ± 2.347 mm respectively. When the noise amplitude $|\imath|$ is not greater than the maximum of the action space, that is, $|\imath| \leq 1$, the end-effector can still maintain the motion trajectories toward the targets and obtain higher motion accuracy. On the contrary, that is, $|\imath| > 1$, the motion trajectories are greatly deflected by noise disturbances, but most motion tasks still reach the targets with high accuracy.



(a) (b) (c) (d) (e) (f)

Figure 11 Motion trajectories of end-effector under noise disturbance. (a) is a situation without noise, and (b)–(f) are situations under noises with $\mathcal{N}(0, 0.5)$, $\mathcal{N}(0, 0.75)$, $\mathcal{N}(0, 1)$, $\mathcal{N}(0, 1.5)$, and $\mathcal{N}(0, 2)$ respectively. During the application of noise disturbances, the larger the noise amplitude is, the more serious the deviation of motion trajectories is. After noise disturbances end, the end-effector can still reach the targets with high accuracy. The integration model is robust

The noise disturbance experiments are analyzed as follows. Firstly, the reason why the robot can still achieve high-precision movement despite noise disturbances is as follows. Noise disturbances cause more or less fluctuations in the motion of the end-effector. But when the noises are terminated, the integration model starts from the current environment state to do motion control from random points to random points, so that the robot can accurately reach the targets. Secondly, the reason why the trajectories are widely deviated to the upper left when the noise disturbances are too large is as follows. Under the action of large noises, the original control signals may easily reach the maximum or minimum value of the action space, which causes muscles to be fully contracted or relaxed. Hence the musculoskeletal upper limb moves to its limit positions, that is, the trajectories are biased to the upper left.

The bio-inspired model can still accomplish motion tasks with high accuracy and reliability under the influence of noise disturbances, so it has strong robustness.

5 Discussion

This research involves an integration method of the basal ganglia and cerebellum (i.e., the research method) and a musculoskeletal robot (i.e., the research object). In both respects, this research differs from previous papers[31–38, 43–47] in method designs and experimental tasks. Those studies have been introduced in Section 1, so only a brief overview of these papers is given below.

The differences from other integration methods of brain regions are as follows. Firstly, designs of network structures are different. In [43–47], various types of basal ganglia networks (such as reservoir actor-critic network[43]) and cerebellum networks (such as motivated developmental network[44–46]) were designed. But in this paper, by simulating the connection patterns of typical cell nuclei in the brain regions, the basal ganglia network and cerebellum networks with series and parallel structures are constructed, which is biologically plausible. Secondly, connections of basal ganglia and cerebellum are different. In [43–47], these two brain regions were equivalent to two parallel controllers, and their outputs were added as the total control signal. But in this paper, the two networks have two-way interactive communication modes, which improves the capabilities of state feature extraction and network learning. Thirdly, designs of exploration-exploitation trade-off in network learning are different. In [43, 45, 47], there was no approach of designing exploration-exploitation trade-off. The ACC neuromodulatory system[44] and the curiosity indicator[46] were proposed to realize the exploration-exploitation trade-off. But in this paper, both adaptive adjustment of target entropy and dopaminergic experience replay can achieve the exploration-utilization trade-off. Finally, robot systems are different. In [43–47], wheeled robots were used to study tasks such as navigation, obstacle avoidance, and self-balancing. But in this paper, a highly nonlinear complex musculoskeletal robot is used to verify the effectiveness of methods.

The differences from other neuro-inspired motion control methods of musculoskeletal robots are as follows. Firstly, control frameworks are different. In [31–37], these control frameworks are essentially open-loop. But in this paper, the control system is closed-loop, that is, the environmental state feeds back to the basal ganglia network at each timestep. Secondly, neural mechanisms are different. In [31–38], these methods were basically designed by simulating the neural mechanism of a single brain region, such as the cerebral cortex, basal ganglia, amygdala, etc. But in this paper, the motion learning method is designed inspired by the integration regulation mechanism of multiple brain regions. Thirdly, motion tasks of musculoskeletal robots are different. In [31–38], a similar musculoskeletal robot was used to complete reaching random target positions from a fixed starting location. But in this paper, the experimental task is generalized. The musculoskeletal robot can be controlled to reach random target positions from random starting locations.

6 Conclusions and Future Work

Inspired by the structures and functions of basal ganglia and cerebellum, the bio-inspired integration model of these two brain regions is proposed, which can effectively control the musculoskeletal robot to complete random reaching tasks. Based on several neural properties of the basal ganglia, it is modeled as an actor-critic architecture. According to the anatomy of the cerebellum, it is modeled as a multi-layer forward network. Different from other methods that dynamically combine the outputs of basal ganglia and cerebellum, the method of this integration model simulates the subcortical interconnections of these two brain regions from a new perspective. The basal ganglia critic network and the cerebellum network are responsible for reward prediction estimation and sensory prediction evaluation respectively, and the basal ganglia actor network outputs control commands. Meanwhile, for the sake of realizing the exploration-exploitation trade-off in the learning process, the automating adjustment method of policy target entropy and the dopaminergic experience replay method are proposed by using the dynamic dopamine proportion hyperparameter. The integration model is biologically interpretable and credible, and its learning efficiency and stability are pretty high. It can control the musculoskeletal end-effector from random initial positions to random desired positions with high precision and high robustness. Therefore, the bio-inspired integration model improves the ability of motion learning and generalization of musculoskeletal robots. The proposed method provides a reference idea for the interdisciplinary research of information science and neuroscience, which is valuable for further carrying out robotics research by drawing on neural mechanisms.

The limitations of this research are analyzed as follows. On the one hand, there is a lack of reference of motion behavior to enable the musculoskeletal robot to produce more human-like movements. On the other hand, the movement task of the musculoskeletal robot is relatively simple. In future work, the methods of this paper will be further improved from the following aspects. Firstly, draw lessons from the motion behavior research to optimize the movement pattern of the musculoskeletal upper limb system. The multiple process model of goal-directed reaching[75, 76] indicates that there exist two types of online regulation sequentially during a single rapid aiming movement, namely impulse control and limb-target control. By introducing these schemes into the state feedback of the integration model, it is not only conducive to achieving dexterous and flexible robot operations, but also expected to realize speed-accuracy-energy optimization. Secondly, perform diverse motion tasks to verify the effectiveness of the methods. It is of great research value that controlling musculoskeletal robots to complete tasks such as grasping, handling, and assembly. In order to accomplish these tasks, issues such as trajectory planning[77, 78] and dual-arm cooperation[79, 80] need to be considered, which requires adapting and improving certain components of the bio-inspired model in terms of different tasks. It is of great significance for the future development and application of musculoskeletal platforms.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Duan G and Liu G P, Attitude and orbit optimal control of combined spacecraft via a fully-actuated system approach, *Journal of Systems Science & Complexity*, 2022, **35**(2): 623–640.

[2] Hu Y, Guo J, Meng W, et al., Longitudinal control for balloon-borne launched solar powered uavs in near-space, *Journal of Systems Science & Complexity*, 2022, **35**(2): 802–819.

[3] Kong X, Xia Y, Hu R, et al., Trajectory tracking control for under-actuated hovercraft using differential flatness and reinforcement learning-based active disturbance rejection control, *Journal of Systems Science & Complexity*, 2022, **35**(2): 502–521.

[4] Wang B, Zhang Y, and Zhang W, A composite adaptive fault-tolerant attitude control for a quadrotor UAV with multiple uncertainties, *Journal of Systems Science & Complexity*, 2022, **35**(1): 81–104.

[5] Qiao H, Chen J, and Huang X, A survey of brain-inspired intelligent robots: Integration of vision, decision, motion control, and musculoskeletal systems, *IEEE Transactions on Cybernetics*, 2021, **52**(10): 11267–11280.

[6] Qiao H, Wu Y, Zhong S, et al., Brain-inspired intelligent robotics: Theoretical analysis and systematic application, M*achine Intelligence Research*, 2023, **20**(1): 1–18.

[7] Qiao H, Zhong S, Chen Z, et al., Improving performance of robots using human-inspired approaches: A survey, *Science China Information Sciences*, 2022, **65**(12): 221201.

[8] Kurumaya S, Suzumori K, Nabae H, et al., Musculoskeletal lower-limb robot driven by multifilament muscles, *Robomech Journal*, 2016, **3**: 1–15.

[9] Wittmeier S, Alessandro C, Bascarevic N, et al., Toward anthropomimetic robotics: Development, simulation, and control of a musculoskeletal torso, *Artificial Life*, 2013, **19**(1): 171–193.

[10] Asano Y, Okada K, and Inaba M, Design principles of a human mimetic humanoid: Humanoid platform to study human intelligence and internal body system, *Science Robotics*, 2017, **2**(13): eaaq0899.

[11] Narioka K and Hosoda K, Motor development of an pneumatic musculoskeletal infant robot, 2011 *IEEE International Conference on Robotics and Automation*, Shanghai, 2011, 963–968.

[12] Boblan I and Schulz A, A humanoid muscle robot torso with biologically inspired construction, *ISR* 2010 (41*st International Symposium on Robotics*) *and ROBOTIK* 2010 (6*th German Conference on Robotics*), Munich, 2010, 1–6.

[13] Yip M C and Niemeyer G, High-performance robotic muscles from conductive nylon sewing thread, 2015 *IEEE International Conference on Robotics and Automation* (*ICRA*), Seattle, 2015, 2313–2318.

[14] Wu Y, Chen J, and Qiao H, Anti-interference analysis of bio-inspired musculoskeletal robotic system, *Neurocomputing*, 2021, **436**: 114–125.

[15] Zhong S, Zhang J, and Nie X, Redundancy reduction of musculoskeletal model for robots with group sparse neural network, 2021 6*th International Conference on Control and Robotics Engineering* (*ICCRE*), Beijing, 2021, 39–43.

[16] Zhong S, Chen J, Niu X, et al., Reducing redundancy of musculoskeletal robot with convex hull vertexes selection, *IEEE Transactions on Cognitive and Developmental Systems*, 2019, **12**(3): 601–617.

[17] Thelen D G, Anderson F C, and Delp S L, Generating dynamic simulations of movement using computed muscle control, *Journal of Biomechanics*, 2003, **36**(3): 321–328.

[18] Jäntsch M, Wittmeier S, Dalamagkidis K, et al., Computed muscle control for an anthropomimetic elbow joint, 2012 *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vilamoura-Algarve, 2012, 2192–2197.

[19] Stanev D and Moustakas K, Simulation of constrained musculoskeletal systems in task space, *IEEE Transactions on Biomedical Engineering*, 2017, **65**(2): 307–318.

[20] Jäntsch M, Wittmeier S, Dalamagkidis K, et al., Adaptive neural network dynamic surface control: An evaluation on the musculoskeletal robot anthrob, 2015 *IEEE International Conference on Robotics and Automation* (*ICRA*), Seattle, 2015, 4347–4352.

[21] Li J, Li Z, Li X, et al., Skill learning strategy based on dynamic motion primitives for human-robot cooperative manipulation, *Transactions on Cognitive and Developmental Systems*, 2020, **13**(1): 105–117.

[22] Lu Z, Wang N, Li M, et al., Incremental motor skill learning and generalization from human dynamic reactions based on dynamic movement primitives and fuzzy logic system, *IEEE Transactions on Fuzzy Systems*, 2021, **30**(6): 1506–1515.

[23] Zeng C, Su H, Li Y, et al., An approach for robotic leaning inspired by biomimetic adaptive control, *IEEE Transactions on Industrial Informatics*, 2021, **18**(3): 1479–1488.

[24] Li Z, Huang Z, He W, et al., Adaptive impedance control for an upper limb robotic exoskeleton using biological signals, *IEEE Transactions on Industrial Electronics*, 2016, **64**(2): 1664–1674.

[25] Li Z, Li X, Li Q, et al., Human-in-the-loop control of soft exosuits using impedance learning on different terrains, *IEEE Transactions on Robotics*, 2022, **38**(5): 2979–2993.

[26] Lillicrap T P, Hunt J J, Pritzel A, et al., Continuous control with deep reinforcement learning, 2015, arXiv: 1509.02971.

[27] Schulman J, Levine S, Abbeel P, et al., Trust region policy optimization, *Proceedings of the International Conference on Machine Learning, PMLR*, 2015, **37**: 1889–1897.

[28] Schulman J, Wolski F, Dhariwal P, et al., Proximal policy optimization algorithms, 2017, arXiv: 1707.06347.

[29] Kidziński Ł, Mohanty S P, Ong C F, et al., Learning to run challenge solutions: Adapting reinforcement learning methods for neuromusculoskeletal environments, *The NIPS'*17 *Competition: Building Intelligent Systems*, Springer, Cham, 2018, 121–153.

[30] Kidziński Ł, Ong C, Mohanty S P, et al., Artificial intelligence for prosthetics: Challenge solutions, *The NeurIPS'*18 *Competition: From Machine Learning to Intelligent Conversations*, Springer, Cham, 2020, 69–128.

[31] Chen J, Chen Z, Yao C, et al., Neural manifold modulated continual reinforcement learning for musculoskeletal robots, *IEEE Transactions on Cognitive and Developmental Systems*, 2022, DOI: 10.1109/TCDS.2022.3231055.

[32] Chen J and Qiao H, Motor-cortex-like recurrent neural network and multitask learning for the control of musculoskeletal systems, *IEEE Transactions on Cognitive and Developmental Systems*, 2020, **14**(2): 424–436.

[33] Chen J and Qiao H, Muscle-synergies-based neuromuscular control for motion learning and gen-

eralization of a musculoskeletal system, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, **51**(6): 3993–4006.

[34] Wang X, Chen J, and Wu W, Motion learning for musculoskeletal robots based on cortex-inspired motor primitives and modulation, *IEEE Transactions on Cognitive and Developmental Systems*, 2023, DOI: 10.1109/TCDS.2023.3293097.

[35] Zhang J, Chen J, Wu W, et al., A cerebellum-inspired prediction and correction model for motion control of a musculoskeletal robot, *IEEE Transactions on Cognitive and Developmental Systems*, 2023, **15**(3): 1209–1223.

[36] Zhong S, Chen Z, and Zhou J, Structure transforming for constructing constraint force field in musculoskeletal robot, *Assembly Automation*, 2021, **42**(2): 169–180.

[37] Zhong S L and Wu W, Motion learning and generalization of musculoskeletal robot using gain primitives, *IEEE Transactions on Automation Science and Engineering*, 2023, DOI: 10.1109/TASE.2023.3249228.

[38] Zhou J, Zhong S, and Wu W, Hierarchical motion learning for goal-oriented movements with speed-accuracy tradeoff of a musculoskeletal system, *IEEE Transactions on Cybernetics*, 2021, **52**(11): 11453–11466.

[39] Joel D, Niv Y, and Ruppin E, Actor-critic models of the basal ganglia: New anatomical and computational perspectives, *Neural Networks*, 2002, **15**(4–6): 535–547.

[40] Kaplan A, Mizrahi-Kliger A D, Israel Z, et al., Dissociable roles of ventral pallidum neurons in the basal ganglia reinforcement learning network, *Nature Neuroscience*, 2020, **23**(4): 556–564.

[41] Takahashi Y, Schoenbaum G, and Niv Y, Silencing the critics: Understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model, *Frontiers in Neuroscience*, 2008, **2**(1): 86–99.

[42] Caligiore D, Arbib M A, Miall R C, et al., The super-learning hypothesis: Integrating learning processes across cortex, cerebellum and basal ganglia, *Neuroscience & Biobehavioral Reviews*, 2019, **100**: 19–34.

[43] Dasgupta S, Wörgötter F, and Manoonpong P, Neuromodulatory adaptive combination of correlation-based learning in cerebellum and reward-based learning in basal ganglia for goal-directed behavior control, *Frontiers in Neural Circuits*, 2014, **8**: 126.

[44] Wang D, Chen S, Hu Y, et al., Behavior decision of mobile robot with a neurophysiologically motivated reinforcement learning model, *IEEE Transactions on Cognitive and Developmental Systems*, 2020, **14**(1): 219–233.

[45] Wang D, Hu Y, and Ma T, Mobile robot navigation with the combination of supervised learning in cerebellum and reward-based learning in basal ganglia, *Cognitive Systems Research*, 2020, **59**: 1–14.

[46] Zhu J, Wang D, and Si J, Flexible behavioral decision making of mobile robot in dynamic environment, *IEEE Transactions on Cognitive and Developmental Systems*, 2022, **15**(1): 134–149.

[47] Ruan X, Chen J, and Yu N, Thalamic cooperation between the cerebellum and basal ganglia with a new tropism-based action-dependent heuristic dynamic programming method, *Neurocomputing*, 2012, **93**: 27–40.

[48] Calabresi P, Picconi B, Tozzi A, et al., Direct and indirect pathways of basal ganglia: A critical reappraisal, *Nature Neuroscience*, 2014, **17**(8): 1022–1030.

[49] Jin X, Tecuapetla F, and Costa R M, Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences, *Nature Neuroscience*, 2014, **17**(3): 423–430.

[50] Gazzaniga M, Ivry R, and Mangun G, *Cognitive Neuroscience: The Biology of the Mind,* W.W. Norton & Company, New York, 2019.

[51] Stephenson-Jones M, Yu K, Ahrens S, et al., A basal ganglia circuit for evaluating action outcomes, *Nature*, 2016, **539**(7628): 289–293.

[52] Abadía I, Naveros F, Ros E, et al., A cerebellar-based solution to the nondeterministic time delay problem in robotic control, *Science Robotics*, 2021, **6**(58): eabf2756.

[53] Bouvier G, Aljadeff J, Clopath C, et al., Cerebellar learning using perturbations, *Elife*, 2018, **7**: e31599.

[54] Manto M, Bower J M, Conforto A B, et al., Consensus paper: Roles of the cerebellum in motor control — The diversity of ideas on cerebellar involvement in movement, *The Cerebellum*, 2012, **11**: 457–487.

[55] Bostan A C, Dum R P, and Strick P L, The basal ganglia communicate with the cerebellum, *Proceedings of the National Academy of Sciences*, 2010, **107**(18): 8452–8456.

[56] Bostan A C and Strick P L, The basal ganglia and the cerebellum: Nodes in an integrated network, *Nature Reviews Neuroscience*, 2018, **19**(6): 338–350.

[57] Hoshi E, Tremblay L, Féger J, et al., The cerebellum communicates with the basal ganglia, *Nature Neuroscience*, 2005, **8**(11): 1491–1493.

[58] Wagner M J, Kim T H, Savall J, et al., Cerebellar granule cells encode the expectation of reward, *Nature*, 2017, **544**(7648): 96–100.

[59] Chen C H, Fremont R, Arteaga-Bracho E E, et al., Short latency cerebellar modulation of the basal ganglia, *Nature Neuroscience*, **17**(12): 1767–1775.

[60] Yoshida J, Oñate M, Khatami L, et al., Cerebellar contributions to the basal ganglia influence motor coordination, reward processing, and movement vigor, *Journal of Neuroscience*, 2022, **42**(45): 8406–8415.

[61] Ohmae S and Medina J F, Climbing fibers encode a temporal-difference prediction error during cerebellar learning in mice, *Nature Neuroscience*, 2015, **18**(12): 1798–1803.

[62] Haarnoja T, Zhou A, Abbeel P, Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, *International Conference on Machine Learning, PMLR*, Stockholm, 2018, 1861–1870.

[63] Haarnoja T, Zhou A, Hartikainen K, et al., Soft actor-critic algorithms and applications, 2018, arXiv: 1812.05905.

[64] Humphries M D, Khamassi M, and Gurney K, Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia, *Frontiers in Neuroscience*, 2012, **6**: 16922.

[65] Budzillo A, Duffy A, Miller K E, et al., Dopaminergic modulation of basal ganglia output through coupled excitation-inhibition, *Proceedings of the National Academy of Sciences*, 2017, **114**(22): 5713–5718.

[66] Zhao F, Zeng Y, Wang G, et al., A brain-inspired decision making model based on top-down biasing of prefrontal cortex to basal ganglia and its application in autonomous UAV explorations, *Cognitive Computation*, 2018, **10**: 296–306.

[67] Fujimoto S, Hoof H, and Meger D, Addressing function approximation error in actor-critic methods, *International Conference on Machine Learning, PMLR,* Stockholm, 2018, 1587–1596.

[68] Haarnoja T, Ha S, Zhou A, et al., Learning to walk via deep reinforcement learning, 2018, arXiv: 1812.11103.

[69] Mnih V, Kavukcuoglu K, Silver D, et al., Human-level control through deep reinforcement learn-

ing, *Nature*, 2015, **518**(7540): 529–533.

[70] Katharopoulos A and Fleuret F, Not all samples are created equal: Deep learning with importance sampling, *International Conference on Machine Learning, PMLR*, Stockholm, 2018, 2525–2534.

[71] Wang C, Wu Y, Vuong Q, et al., Striving for simplicity and performance in off-policy drl: Output normalization and non-uniform sampling, *International Conference on Machine Learning, PMLR,* 2020, 10070–10080.

[72] Constantinidis C and Klingberg T, The neuroscience of working memory capacity and training, *Nature Reviews Neuroscience*, 2016, **17**(7): 438–449.

[73] Cools R and D'Esposito M, Inverted-u-shaped dopamine actions on human working memory and cognitive control, *Biological Psychiatry*, 2011, **69**(12): e113–e125.

[74] Todorov E, Erez T, and Tassa Y, Mujoco: A physics engine for model-based control, 2012 *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vilamoura-Algarve, 2012, 5026–5033.

[75] Elliott D, Hansen S, Grierson L E, et al., Goal-directed aiming: Two components but multiple processes, *Psychological Bulletin*, 2010, **136**(6): 1023–1044.

[76] Elliott D, Lyons J, Hayes S J, et al., The multiple process model of goal-directed reaching revisited, *Neuroscience & Biobehavioral Reviews*, 2017, **72**: 95–110.

[77] Huang X, Wu W, and Qiao H, Connecting model-based and model-free control with emotion modulation in learning systems, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, **51**(8): 4624–4638.

[78] Liu Z, Lu Z, Zhao Z, et al., Single parameter adaptive neural network control for multi-agent deployment with prescribed tracking performance, *Automatica*, 2023, **156**: 111207.

[79] Li Z, Li G, Wu X, et al., Asymmetric cooperation control of dual-arm exoskeletons using human collaborative manipulation models, *IEEE Transactions on Cybernetics*, 2021, **52**(11): 12126–12139.

[80] Wang Y, Li H, Zhao Y, et al., A fast coordinated motion planning method for dual-arm robot based on parallel constrained DDP, *IEEE/ASME Transactions on Mechatronics*, 2023, DOI: 10.1109/TMECH.2023.3323798.

Appendix

A.1 Proof of Lemma 3.1

Proof Define the entropy augmented reward in the basal ganglia model as follows:

$$r_\pi(\boldsymbol{s}_t,\boldsymbol{a}_t) \triangleq r(\boldsymbol{s}_t,\boldsymbol{a}_t) + \gamma\alpha\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\left[\mathcal{H}\left(\pi\left(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1}\right)\right)\right]. \tag{A.1}$$

Rewrite the soft state-action value as follows:

$$Q(\boldsymbol{s}_t,\boldsymbol{a}_t) \leftarrow r(\boldsymbol{s}_t,\boldsymbol{a}_t) + \gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\left[V\left(\boldsymbol{s}_{t+1}\right)\right]$$

$$\begin{aligned}
&\leftarrow r(\boldsymbol{s}_t,\boldsymbol{a}_t) + \gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p,\boldsymbol{a}_{t+1}\sim\pi}\left[Q\left(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1}\right) - \alpha\ln\pi\left(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1}\right)\right]\\
&\leftarrow r(\boldsymbol{s}_t,\boldsymbol{a}_t) + \gamma\alpha\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p,\boldsymbol{a}_{t+1}\sim\pi}\left[-\ln\pi\left(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1}\right)\right] + \gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p,\boldsymbol{a}_{t+1}\sim\pi}\left[Q\left(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1}\right)\right]\\
&\leftarrow r(\boldsymbol{s}_t,\boldsymbol{a}_t) + \gamma\alpha\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\left[\mathcal{H}\left(\pi\left(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1}\right)\right)\right] + \gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p,\boldsymbol{a}_{t+1}\sim\pi}\left[Q\left(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1}\right)\right]\\
&\leftarrow r_\pi(\boldsymbol{s}_t,\boldsymbol{a}_t) + \gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p,\boldsymbol{a}_{t+1}\sim\pi}\left[Q\left(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1}\right)\right].
\end{aligned} \tag{A.2}$$

For (A1), the reward term $r(\boldsymbol{s}_t, \boldsymbol{a}_t)$ is bounded, and the assumption $\|\mathcal{A}\|_\infty < \infty$ guarantees that the entropy term $\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\left[\mathcal{H}\left(\pi\left(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1}\right)\right)\right]$ is bounded, so that $r_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$ is bounded. Leveraging the standard convergence results for policy evaluation, $Q^k \to Q^\pi$ as $k \to \infty$. ▮

A.2 Proof of Lemma 3.2

Proof For any policy $\pi_{\text{old}} \in \mathit{\Pi}$, its soft state value and soft state-action value are $V^{\pi_{\text{old}}}$ and $Q^{\pi_{\text{old}}}$, respectively. According to the policy update rules for the integration model of basal ganglia and cerebellum, obtaining a new policy π_{new} can be equivalently transformed into solving the minimization problem as follows:

$$\begin{aligned}\pi_{\text{new}}\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right) &= \arg\min_{\pi'\in\mathit{\Pi}} D_{\text{KL}}\left(\pi'\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\left\|\frac{\exp\left(\frac{1}{\alpha}Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)\right)}{Z^{\pi_{\text{old}}}\left(\boldsymbol{s}_t\right)}\right.\right) + \mu(\nu(\boldsymbol{s}_t)) \\ &= \arg\min_{\pi'\in\mathit{\Pi}} J_{\pi_{\text{old}}}\left(\pi'\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right) + \mathcal{F}(\boldsymbol{s}_t),\end{aligned} \tag{A.3}$$

where

$$\begin{aligned}J_{\pi_{\text{old}}}\left(\pi'\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right) &\triangleq D_{\text{KL}}\left(\pi'\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\left\|\frac{\exp\left(\frac{1}{\alpha}Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)\right)}{Z^{\pi_{\text{old}}}\left(\boldsymbol{s}_t\right)}\right.\right) \\ &= D_{KL}\left(\pi'\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\left\|\exp\left(\frac{1}{\alpha}Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) - \ln Z^{\pi_{\text{old}}}\left(\boldsymbol{s}_t\right)\right)\right.\right) \\ &= \int \pi'\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\ln\frac{\pi'\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)}{\exp\left(\frac{1}{\alpha}Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) - \ln Z^{\pi_{\text{old}}}\left(\boldsymbol{s}_t\right)\right)}d\boldsymbol{a}_t \\ &= \int \pi'\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\left(\ln\pi'\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right) - \frac{1}{\alpha}Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) + \ln Z^{\pi_{\text{old}}}\left(\boldsymbol{s}_t\right)\right)d\boldsymbol{a}_t \\ &= \mathbb{E}_{\boldsymbol{a}_t\sim\pi'}\left[\ln\pi'\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right) - \frac{1}{\alpha}Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) + \ln Z^{\pi_{\text{old}}}\left(\boldsymbol{s}_t\right)\right]\end{aligned} \tag{A.4}$$

and

$$\mathcal{F}(\boldsymbol{s}_t) \triangleq \mu(\nu(\boldsymbol{s}_t)). \tag{A.5}$$

There must exist π_{new} that is not worse than π_{old} in $\mathit{\Pi}$. The worst case is to let $\pi_{\text{new}} = \pi_{\text{old}}$. Thus,

$$J_{\pi_{\text{old}}}\left(\pi_{\text{new}}\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right) + \mathcal{F}(\boldsymbol{s}_t) \leq J_{\pi_{\text{old}}}\left(\pi_{\text{old}}\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right) + \mathcal{F}(\boldsymbol{s}_t). \tag{A.6}$$

Substitute (A.4) and (A.5) into the above equation (A.6):

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{a}_t\sim\pi_{\text{new}}}\left[\ln\pi_{\text{new}}\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right) - \frac{1}{\alpha}Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) + \ln Z^{\pi_{\text{old}}}\left(\boldsymbol{s}_t\right)\right] + \mu(\nu(\boldsymbol{s}_t)) \\ \leq &\mathbb{E}_{\boldsymbol{a}_t\sim\pi_{\text{old}}}\left[\ln\pi_{\text{old}}\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right) - \frac{1}{\alpha}Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right) + \ln Z^{\pi_{\text{old}}}\left(\boldsymbol{s}_t\right)\right] + \mu(\nu(\boldsymbol{s}_t)).\end{aligned} \tag{A.7}$$

$Z^{\pi_{\text{old}}}(\boldsymbol{s}_t)$ and $\mu(\nu(\boldsymbol{s}_t))$ depend only on $\boldsymbol{s}_t$, which is generated by π_{old} in (A.3), so they can be directly canceled out. Further, both sides of the inequality are multiplied by α, which facilitates calculation and does not affect the result:

$$\mathbb{E}_{\boldsymbol{a}_t\sim\pi_{\text{new}}}\left[\alpha\ln\pi_{\text{new}}\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right) - Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)\right] \leq \mathbb{E}_{\boldsymbol{a}_t\sim\pi_{\text{old}}}\left[\alpha\ln\pi_{\text{old}}\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right) - Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)\right]. \tag{A.8}$$

Apply the definition (3) of soft state value function to the above inequality (A.8):

$$\mathbb{E}_{\boldsymbol{a}_t\sim\pi_{\text{new}}}\left[Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)-\alpha\ln\pi_{\text{new}}\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right]\geq V^{\pi_{\text{old}}}\left(\boldsymbol{s}_t\right). \tag{A.9}$$

Equation (A.9) connects the trajectory generated by π_{new} with that generated by π_{old}. Next apply (A.9) to the soft Bellman equation:

$$\begin{aligned}&Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)\\ =&r\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)+\gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\left[V^{\pi_{\text{old}}}\left(\boldsymbol{s}_{t+1}\right)\right]\\ \leq&r\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)+\gamma\mathbb{E}_{\boldsymbol{s}_{t+1}\sim p}\left[\mathbb{E}_{\boldsymbol{a}_{t+1}\sim\pi_{\text{new}}}\left[Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_{t+1},\boldsymbol{a}_{t+1}\right)-\alpha\ln\pi_{\text{new}}\left(\boldsymbol{a}_{t+1}|\boldsymbol{s}_{t+1}\right)\right]\right].\end{aligned} \tag{A.10}$$

In the above inequality (A.10), $\boldsymbol{a}_t$ is all generated by π_{new}, and $Q^{\pi_{\text{old}}}$ is pushed to the next time $t+1$. Repeatedly, expand $Q^{\pi_{\text{old}}}$ on the right-hand side of (A.10) by applying the soft Bellman equation, and then replace the $V^{\pi_{\text{old}}}$ with (A.9). Eventually, after all $Q^{\pi_{\text{old}}}$ are replaced, the right-hand side of (A.10) will become soft Bellman equations based on π_{new}. Then leveraging Lemma 3.1, these terms all converge to $Q^{\pi_{\text{new}}}$. Hence

$$Q^{\pi_{\text{old}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)\leq Q^{\pi_{\text{new}}}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right). \tag{A.11}$$

▮

A.3 Proof of Theorem 3.3

Proof Let π_i denote the policy after i iterations of policy evaluation and policy improvement. The soft state-value value sequence Q^{π_i} must satisfy two conditions. Firstly, according to Lemma 3.2, $Q^{\pi_{i+1}}\geq Q^{\pi_i}$ is always true, that is, Q^{π_i} is monotonically increasing. Secondly, under the assumptions that reward and policy entropy are bounded, Q^{π} is bounded, so Q^{π_i} converges to some π^*. For any $\pi\in\mathit{\Pi}$ and $\pi\neq\pi^*$, it must be case that

$$J_{\pi^*}\left(\pi^*\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right)<J_{\pi^*}\left(\pi\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right), \tag{A.12}$$

where

$$J_{\pi^*}\left(\pi'\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right)\triangleq D_{\text{KL}}\left(\pi'\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\|\frac{\exp\left(\frac{1}{\alpha}Q^{\pi^*}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)\right)}{Z^{\pi^*}\left(\boldsymbol{s}_t\right)}\right). \tag{A.13}$$

And then

$$J_{\pi^*}\left(\pi^*\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right)+\mathcal{F}(\boldsymbol{s}_t)<J_{\pi^*}\left(\pi\left(\boldsymbol{a}_t|\boldsymbol{s}_t\right)\right)+\mathcal{F}(\boldsymbol{s}_t). \tag{A.14}$$

Similar to the proof process in Lemma 3.2, it is easily deduced that

$$Q^{\pi^*}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)>Q^{\pi}\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right),\quad\forall\left(\boldsymbol{s}_t,\boldsymbol{a}_t\right)\in\mathcal{S}\times\mathcal{A}. \tag{A.15}$$

Hence π^* is indeed the optimal policy in $\mathit{\Pi}$. ▮