Connecting Model-Based and Model-Free Control With Emotion Modulation in Learning Systems

Xiao Huang, Wei Wu, and Hong Qiao, *Fellow, IEEE*

***Abstract*—This article proposes a novel decision-making framework that bridges a gap between model-based (MB) and model-free (MF) control processes through only adjusting the planning horizon. Specifically, the output policy is obtained by solving a model predictive control problem with a locally optimal state value as terminal constraints. When the planning horizon decreases to zero, the MB control will transform into the MF control smoothly. Meanwhile, inspired by the neural mechanism of emotion modulation on decision-making, we build a biologically plausible computational model of emotion processing. This model can generate an uncertainty-related emotional response on the basis of the state prediction error and reward prediction error, and then dynamically modulates the planning horizon in the tasks. The simulation results demonstrate that the proposed decision-making framework can produce better policies than traditional methods. Emotion modulation can shift the MB and MF control well to improve the learning efficiency and the speed of decision-making.**

***Index Terms*—Brain-inspired computing, decision-making, emotion modulation, emotion-cognition interactions, reinforcement learning.**

I. Introduction

HUMAN behaviors are often organized into goal-directed and habitual processes, which are studied as model-based (MB) and model-free (MF) decision-making systems, respectively. MB decision-making involves building a model of the environment to predict the outcomes, and searching an action that maximizes the return via dynamic programming [1]. This process is generally data-efficient but expensive in terms of computation time and working memory. On the contrary, MF decision-making is to learn to estimate a long-term utility and form a fixed state-action mapping based on previous experience with reinforcing consequences during interacting with the environment [2]. Compared with the MB approach, this process usually has a faster speed of decision-making, but requires a lot of practical training. Cooperative working of these two processes are very important for survival of animals. For robotics, a huge challenge is to learn how to reason about the underlying environmental dynamics and learn how to make decisions quickly and accurately. One feasible approach is to develop novel computational models based on a coherent and comprehensive mapping of the key neural mechanisms in human brain, such as [3]–[5].

In recent years, many computational approaches to decision-making have been developed in the framework of reinforcement learning and adaptive dynamic programming. For robotic control, a series of MF algorithms have been proposed to perform complex decision-making tasks. A prevalent architecture is actor-critic methods that consist of a parameterized policy function as an actor and a parameterized value function as a critic. The goal of these methods is to learn an actor that maximizes the critic, which is usually drawn into optimal control [6]–[8]. At the same time, many prominent methods integrated with deep learning have emerged, such as trust-region policy optimization [9], deterministic policy gradients (DPGs) [10], and hierarchical deep reinforcement learning [11]. However, most MF methods suffer from a low data efficiency and need a lot of interactions with the environment, which is unrealistic for general robotic systems.

In order to improve the data efficiency, some researchers adopt MB algorithms to control robotic systems, such as PILCO [12] and guided policy search (GPS) [13], [14]. As for PILCO, the uncertainty of model is explicitly incorporated into long-term planning, which facilitates learning with a high data efficiency. Unfortunately, at each iteration, large nonlinear optimizations require expensive resources of computation and memory. GPS can search deep visuomotor policies through end-to-end supervised training with guide of MB dynamic programming. This approach shows good performance in a range of real-world manipulation tasks with visual input. In fact, both above algorithms are related to the architecture of model

Manuscript received May 31, 2019; accepted July 19, 2019. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1300200 and Grant 2017YFB1300203, in part by the National Natural Science Foundation of China under Grant 91648205, Grant 61627808, and Grant 61702516, in part by the Strategic Priority Research Program of Chinese Academy of Science under Grant XDB32050100, and in part by the Development of Science and Technology of Guangdong Province Special Fund Project under Grant 2016B090910001. This article was recommended by Associate Editor W. Sun. *(Corresponding author: Hong Qiao.)*

X. Huang and W. Wu are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the Beijing Key Laboratory of Research and Application for Robotic Intelligence of Hand-Eye-Brain Interaction, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China.

H. Qiao is with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the Beijing Key Laboratory of Research and Application for Robotic Intelligence of Hand-Eye-Brain Interaction, Beijing 100190, China, also with the University of Chinese Academy of Sciences, Beijing 100049, China, also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with the Cloud Computing Center, Chinese Academy of Sciences, Beijing 100190, China (e-mail: hong.qiao@ia.ac.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2019.2933152

predictive control (MPC), especially differential dynamic programming (DDP) [15] or iterative LQR (iLQR) [16]. As a typical MB method, MPC has been successfully applied to the robotics and control systems. For example, an output feedback MPC with the integration of an extended state observer is proposed to suppress disturbances and increase the robustness against various model uncertainties in hydraulic systems [17]. In the work [18], an online MPC is designed handle the robotic food-cutting task, learning controllers directly from data. However, the error of model prediction has a huge impact on searching the optimal policy, and MB planning is usually time-consuming and only obtains a suboptimal solution. In addition, most of above algorithms have ignored that MB and MF learning systems operate in parallel among humans [19].

Emotion recently has become more attractive to improve the learning efficiency of the agent, especially in the community of computer science and developmental robot. For example, Belkaid *et al.* [20] introduced a conceptual model named eMODEL that integrates emotional modulation into the cognitive robots. Moerland *et al.* [21] reviewed the functions of emotion in reinforcement learning agents and robots. That work investigated various methods of affective modeling for improving the learning efficiency of the agent, and compared different evaluation methods from the aspects of emotion elicitation, emotion type, emotion function, and test scenario. Meantime, the authors consider that emotion may influence the learning loop in four main aspects: 1) reward modification; 2) state modification; 3) meta-learning; and 4) action selection. For instance, Huang *et al.* [22] built a novel model of emotion generation to adjust the parameters of learning adaptively, as a part of meta-learning. However, most methods reviewed in that work are inspired by psychological theories of emotion, which is not biologically plausible.

The motivation of this article is to improve the efficiency of learning and speed up the decision-making in the robotic control tasks. For this purpose, a unified decision-making framework is built to bridge a gap between MB and MF control processes through adjusting the planning horizon. In this framework, the greedy output policy is obtained by solving an MPC problem with a locally optimal state value as terminal constraints. During the optimization process, the sequence of actions is guided by the output of parameterized policy network, which ensures the stability of policy search. When the planning horizon decreases to zero, the MB control will transform into the MF control smoothly. Additionally, inspired by the neural mechanism of emotion modulation between the MB and MF decision-making, we build a biologically plausible computational model of amygdala to produce uncertainty-related emotional responses that can dynamically influence the planning horizon in the course of task. Specifically, if the state prediction error (SPE) is large and the reward prediction error (RPE) is small, the intensity of uncertainty-related emotional response will increase such that more short-term MB planning or MF policy will be adopted for a faster decision-making, and vice versa. As a result, the proposed algorithm is significant for robots to acquire the skills for performing some complex tasks quickly and accurately, such as robotic manipulation, locomotion and navigation. Besides, this article can promote the interdisciplinary integration of neuroscience and artificial intelligence.

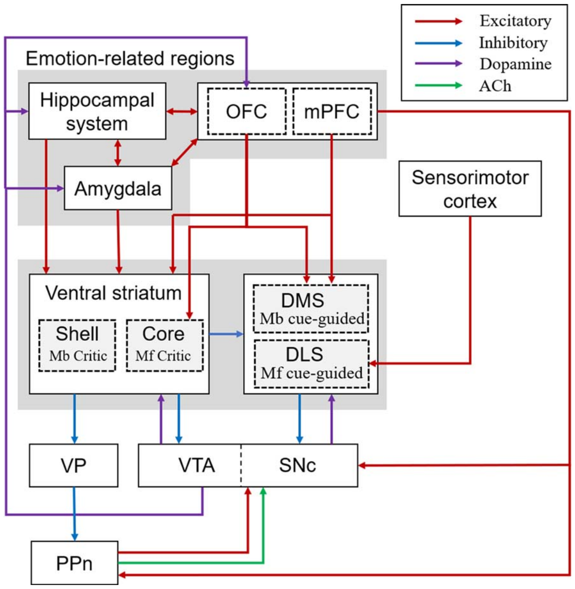

Fig. 1. Substrates of MF and MB decision-making. Abbreviation: Mb, model-based; Mf, model-free; OFC, orbitofrontal cortex; mPFC, medial prefrontal cortex; DMS, dorso-medial striatum; DLS, dorso-lateral striatum; VP, ventral pallidum; VTA, ventral tegmental area; SNc, substantia nigra pars compacta; and PPn, pedunculopontine nucleus.

The rest of this article is organized as follows. First of all, the neural mechanisms of decision-making systems and emotion processing are investigated in Section II. In Section III, some existing approaches are described to learn a probabilistic dynamic model and an MF guiding policy. Then the unified framework for decision-making and the computational model of emotion processing is described in Section IV. In Section V, the proposed algorithms are implemented in the inverted pendulum swing-up task and the robotic arm reaching task, respectively. The experimental results are discussed further. Finally, the conclusions are given in Section VI.

II. Biological Background

A. *Neural Substrates of Model-Free and Model-Based Decision-Making*

It has long been known that human behavior is controlled by multiple competing systems in operant conditioning: a goal-directed system and a habitual system [23], which are often involved in a theory for distinct forms of decision-making, namely MB and MF control. Neural substrates of the two systems have been studied deeply for a long time and fruitful results are achieved. Recently, some neuroscientists try to integrate these two systems into a unified framework [24], [25]. According to the findings of these works, a rough picture of neural computation for these two processes is drawn in Fig. 1.

A large amount of neuroscience literatures indicate that the region of striatum plays a key role in the goal-directed and habitual process. As a growing consensus, the striatum is divided functionally into dorso-lateral (DLS), dorso-medial (DMS), and ventral striatum (VS), each of which plays distinct functional roles within those broader behavioral computation [24]. For example, the DLS is considered as a central substrate for MF learning and expression. During the course of a conditioning task, habit behavior can make the single neuron activity in DLS reorganize [26], [27], and lesions of the DLS make animals obtain more goal-oriented behaviors [28]. Meanwhile, the DMS is mainly involved in MB learning and expression [24], [29]–[31]. Animals with lesions of this region maintain habitual behavior from the outset. The VS is associated with computation of rewards for MF and MB processes, and also with learning the probability of action selection during the transition of states, which acts like value function in the machine learning theory. It is reflected in two aspects: the core of VS may control the influence of reward values in Pavlovian conditioning that learns to associate outcomes to different stimuli directly [32]; the shell of VS may control the influence of reward values in Pavlovian-instrumental transfer that uses those learned experience to infer instrumental actions [32].

Besides striatum, the implementation of these two processes is associated with many other regions, including OFC, mPFC, and the natural reward circuitry, such as VTA, SNc, etc. During the decision-making process, the assessment of the outcomes derives from several aspects of cognition, especially rational assessment and emotional valence [33]. OFC is able to integrate stimulus-reward and context-reward information (from amygdala and hippocampal system) to provide expected reward information. Then, these information is projected into downstream structures, such as VS, VTA and PPn [34]–[36]. In a similar way, the mPFC plays a significant role in emotional decisions, which seemingly uses the emotional reactions to model human behavior in certain social situations [2]. An important function of this pathway is to compute the subjective value of choices [2]. Then this information will be integrated in other brain areas associated with basal ganglia, such as striatum, VTA, and SN, which perform further processing of reward signals. The striatum and the VTA, as critical components of the motor and reward systems, coordinate multiple aspects of motivation/reward cognition and reinforcement learning. Specifically, neurons in these brain areas cannot only integrate reward information (such as RPE) and code them into movement activities, but also change their reward-related activities during learning [37].

B. Modulatory Neural Circuits Between Emotion and Decision-Making

According to the related researches in the field of affective neuroscience and neuroeconomics, emotion play a modulatory role throughout the decision-making process [2]. Actually, in Fig. 1, the amygdala, the PFC, and other structures have been thought as dominant neural substrates of emotion. Affect's modulation on decision-making is mainly reflected in two aspects. The first one is that a specific affective state may alter decisions through modulating some intermediate processes of decision-making. For example, stress usually leads to a shift from goal-directed to habitual behavior in human, which is due to the fact that even mild stress can impair function of the prefrontal cortex (PFC) and enhance amygdala function [38]–[40]. The second one is that emotional reactions may be incorporated into the computation of subjective value during making decisions. For instance, the amygdala, a central component of emotion generation, contributes to value coding in the striatum and the OFC/mPFC, and modulates learning from reinforcement [41], [42]. This region plays a critical role in associating aversive, threatening events with neural cues (i.e., Pavlovian fear conditioning).

Notably, emotion may influence a shift between MB and MF control. For example, some studies have demonstrated that stress can influence the balance between PFC and striatal contributions to decision-making, which are associated with MB and MF reinforcement learning, respectively [19]. When stress occurs, one tends to attenuate MB actions and increase habitual MF actions [38], [39]. Some authors also conjecture that the brain uses a range of pre-programmed control algorithms for survival, including MB and MF control. The output of them may link to a low-dimensional core affective space, such as utility or valence (mediating approach or withdrawal) and arousal (mediating invigoration and inhibition) [1]. Particularly, in the work [19], [43], it has been hypothesized that the uncertainty of state and reward prediction can arbitrate the goal-directed and habitual systems directly. While, empirical evidence indicates that the amygdala plays a significant role in creating uncertainty-related emotional responses [44].

III. Preliminaries

In this section, some existing approaches are described to learn a probabilistic dynamic model and generate an MF guiding policy. The learned model of environment is used to predict the coming states, which contributes to MB planning and emotion processing. The MF learning is mainly used for training a global state-action value function offline and producing a guiding policy that maps the state to the action directly. They are necessary ingredients for the proposed MB control in Section IV.

A. Probabilistic Dynamic Model Learning

Assume the dynamic system is probabilistic with uncertainty, where the next state can be represented as a conditional distribution given the current state and action. Here, as in [45], an ensemble of forward probabilistic neural networks is used to model the uncertain dynamics instead of Gaussian regression model that has an expansive computational cost. Assume there are N probabilistic neural networks with the same structure. Each of them, parameterized by ϕ, encodes a Gaussian distribution for capturing aleatoric uncertainty $f_{\phi_n}(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t) = \mathcal{N}(\boldsymbol{\mu}_{\phi_n}(\mathbf{x}_t, \mathbf{u}_t), \boldsymbol{\Sigma}_{\phi_n}(\mathbf{x}_t, \mathbf{u}_t))$. Aleatoric uncertainty is a kind of random noise, such as observation and process noise, which arises from inherent stochasticity of a system [46]. During training, the mean of the negative log-likelihood loss of each subnetwork is minimized as follows:

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\Big[\big(\boldsymbol{\mu}_{\phi_n}(\mathbf{x}_t,\mathbf{u}_t)-\mathbf{y}\big)^T\boldsymbol{\Sigma}_{\phi_n}^{-1}\big(\boldsymbol{\mu}_{\phi_n}(\mathbf{x}_t,\mathbf{u}_t)-\mathbf{y}\big) + \log\det\boldsymbol{\Sigma}_{\phi_n}(\mathbf{x}_t,\mathbf{u}_t)\Big] \tag{1}$$

where $\mathbf{y}$ represents the true next state $\mathbf{x}_{t+1}$. $\boldsymbol{\mu}_{\phi_n}(\cdot)$ and $\boldsymbol{\Sigma}_{\phi_n}(\cdot)$, as the mean and covariance of next state, are computed by nth probabilistic neural network.

Single subnetwork can model aleatoric uncertainty successfully, but cannot model epistemic uncertainty that represents subjective uncertainty about the dynamic function. Fortunately, this kind of uncertainty can be estimated through analyzing the activity of an ensemble of many networks. Due to limited experience, different subnetworks may output different predictions of the next state. We first create M particles from the distribution of current state $p(\mathbf{x}_t)$, and assign them to each subnetwork to predict plausible distribution of the next state. Then the prediction is approximated further as a Gaussian distribution whose covariance $\bar{\boldsymbol{\Sigma}} = \mathrm{diag}(\bar{\boldsymbol{\sigma}}^2)$ and mean $\bar{\boldsymbol{\mu}}$ are computed by

$$\bar{\boldsymbol{\mu}}(\mathbf{x}_t,\mathbf{u}_t) = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\mu}_{\phi_n}\big(\mathbf{x}_{t+1}^m|\mathbf{x}_t^m,\mathbf{u}_t^m\big)\right] \tag{2}$$

$$\bar{\boldsymbol{\sigma}}^2(\mathbf{x}_t,\mathbf{u}_t) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{1}{M}\sum_{m=1}^{M}\Big[\boldsymbol{\mu}_{\phi_n}^2\big(\mathbf{x}_{t+1}^m|\mathbf{x}_t^m,\mathbf{u}_t^m\big) + \boldsymbol{\sigma}_{\phi_n}^2\big(\mathbf{x}_{t+1}^m|\mathbf{x}_t^m,\mathbf{u}_t^m\big)\Big]\right\} - \bar{\boldsymbol{\mu}}^2(\mathbf{x}_t,\mathbf{u}_t). \tag{3}$$

The prediction of the next state is obtained by sampling from the learned probabilistic dynamics model $\mathcal{F} : \tilde{\mathbf{x}}_{t+1} \sim \mathcal{N}(\bar{\boldsymbol{\mu}}(\mathbf{x}_t,\mathbf{u}_t), \bar{\boldsymbol{\Sigma}}(\mathbf{x}_t,\mathbf{u}_t))$.

B. Model-Free Learning

The actor-critic algorithm is adopted to train a global state-action value function and produce a guiding policy that maps the observation to the action directly. At each timestep t, the agent receives an observation $\mathbf{x}_t \in \mathcal{X}$, takes an action $\mathbf{u}_t \in \mathcal{U}$ and receives a scalar reward r_t. The agent's output of decision-making is determined by a policy π, which maps states to a probability distribution over the actions $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{U})$. Assume the environment is stochastic and subject to Gaussian distribution. It is usually modeled as a Markov decision process (MDP) with an initial state distribution $p(\mathbf{x}_1)$, transition dynamics $p(\mathbf{x}_{t+1}|\mathbf{x}_t,\mathbf{u}_t)$, and reward function $r(\mathbf{x}_t,\mathbf{u}_t)$. The state value function is $V(\mathbf{x}_t) = \sum_{k=t}^{\infty}\gamma^{k-t}r(\mathbf{x}_k,\mathbf{u}_k)$, where $\gamma \in [0, 1]$ is a discount factor. The state-action value function is usually introduced to associate with policy $\mathbf{u} = \pi(\mathbf{x})$ directly as

$$Q^{\pi}(\mathbf{x}_t,\mathbf{u}_t) = r(\mathbf{x}_t,\mathbf{u}_t) + \gamma V^{\pi}(\mathbf{x}_{t+1}). \tag{4}$$

The objective of reinforcement learning is to select a policy that maximizes the value function, which obtains

$$V^*(\mathbf{x}_t) = \max_{\pi(\cdot)}\sum_{k=t}^{\infty}\gamma^{k-t}r(\mathbf{x}_k,\pi(\mathbf{x}_k)). \tag{5}$$

The corresponding optimal Q function is defined as follows:

$$Q^*(\mathbf{x}_t,\mathbf{u}_t) = r(\mathbf{x}_t,\mathbf{u}_t) + \gamma V^*(\mathbf{x}_{t+1}). \tag{6}$$

The expected return of MF learning method is defined as

$$\mathbb{E}_{p(\mathbf{x}_1),\pi(\mathbf{u}_t|\mathbf{x}_t)_{t\geq1},p(\mathbf{x}_{t+1}|\mathbf{x}_t,\mathbf{u}_t)_{t\geq1}}\left[\sum_{t=1}^{\infty}\gamma^{t-1}r(\mathbf{x}_t,\mathbf{u}_t)\right] = \mathbb{E}_{p^{\chi}(\mathbf{x}),\pi(\mathbf{u}|\mathbf{x})}\big[Q^{\pi}(\mathbf{x},\mathbf{u})\big] \tag{7}$$

where χ represents the state distribution that derives from the policy $\pi(\mathbf{u}|\mathbf{x})$ and the true system dynamics. In actor-critic framework, the policy is usually parameterized by a policy network with parameter θ, and the state-action value function is usually estimated by value/critic network denoted by $Q_{\varphi}(\mathbf{x},\mathbf{u})$, whose parameter φ is learned such that $Q_{\varphi}(\mathbf{x},\mathbf{u}) \approx Q^{\pi_\theta}(\mathbf{x},\mathbf{u})$. And the parameter θ can be optimized through solving the following problem:

$$\theta^* = \arg\max_{\theta}\mathbb{E}_{p^{\chi}(\mathbf{x}),\pi_\theta(\mathbf{u}|\mathbf{x})}[Q(\mathbf{x},\mathbf{u})]. \tag{8}$$

The DPGs method [10] assumes the actor π_θ is deterministic for continuous control tasks, such that the parameters can be updated as follows:

$$\theta \leftarrow \theta + \eta_{\pi}\mathbb{E}_{p^{\chi}(\mathbf{x})}\big[\nabla_\theta\pi_\theta(\mathbf{x})\nabla_u Q(\mathbf{x},\mathbf{u})|_{\mathbf{u}=\pi_\theta(\mathbf{x})}\big] \tag{9}$$

where the first-order information of critic is used to modify the actor's parameter, and η_π is the learning rate. As for the critic network, the parameter φ is updated by minimizing the squared Bellman error, which is formalized as follows:

$$\varphi \leftarrow \varphi - \eta_Q\nabla_\varphi\mathbb{E}_{p^{\chi}(\mathbf{x}),\pi_\theta(\mathbf{u}|\mathbf{x}),p(\mathbf{x}'|\mathbf{x},\mathbf{u})}\Big[\big(Q_\varphi(\mathbf{x},\mathbf{u}) - y\big)^2\Big] \tag{10}$$

where η_Q is the learning rate. The expected value $y = r(\mathbf{x},\mathbf{u}) + \gamma\mathbb{E}_{\pi_\theta(\mathbf{u}'|\mathbf{x}')}[Q_{\varphi'}(\mathbf{x}',\mathbf{u}')]$ is obtained by the target critic function $Q_{\varphi'}(\mathbf{x}',\mathbf{u}')$, whose parameter φ' is updated by running average $\varphi' \leftarrow \tau\varphi + (1-\tau)\varphi'$. As suggested in [10], the target critic network can improve the learning stability.

IV. Proposed Approach

In this section, a unified decision-making framework is proposed to link the MB and MF control, and some important properties of this method are analyzed. Then the computational model of emotion-processing network is described, which is used for producing the uncertainty-related emotional response based on SPE and RPE, and modulating the planning horizon in the decision-making process. The architecture is shown in Fig. 2.

A. Model-Based Control

MB control has to build a set of beliefs about the structure of the environment and predict the outcome of each plan. Based on these predictions, it can search a series of choices in the form of dynamic programming [37]. MPC is a typical MB theory in robotics and control systems. In MPC, a locally optimal policy $\hat{\pi}(\mathbf{u}|\mathbf{x})$ is computed based on the knowledge

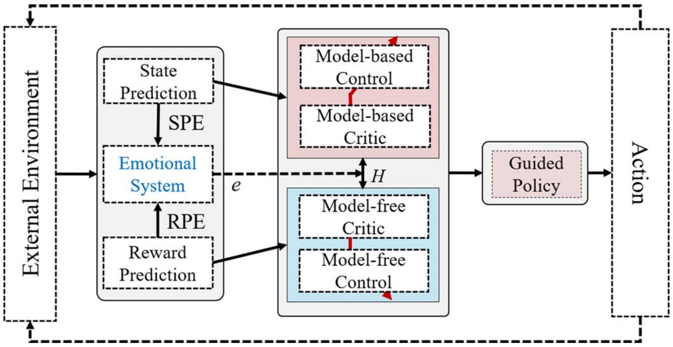

Fig. 2. Architecture of decision-making with modulation of emotion.

of the learned dynamics model. Generally, the optimization objective of MPC problem is

$$\begin{aligned}\max_{\hat{\pi}} \quad & \mathbb{E}_{p^{\nu}(\mathbf{x}),\hat{\pi}(\mathbf{u}|\mathbf{x})}\left[\sum_{t=0}^{H-1}\gamma^{t} r(\mathbf{x}_t,\mathbf{u}_t)+\gamma^{H} r_f(\mathbf{x}_H)\right] \\ \text{s.t.} \quad & \mathbf{x}_{t+1}\sim\mathcal{F}(\mathbf{x}_t,\mathbf{u}_t)\end{aligned} \tag{11}$$

where ν denotes the state distribution that derives from the learned dynamics $\mathcal{F}$ and the locally optimal policy $\hat{\pi}$. The short-term discounted return is the sum of running reward r and terminal reward r_f. The running and terminal reward function is often defined as a quadratic equation

$$r(\mathbf{x}_t,\mathbf{u}_t)=\left(\mathbf{x}_t-\mathbf{x}_t^{g}\right)^{T}\mathbf{Q}\left(\mathbf{x}_t-\mathbf{x}_t^{g}\right)+\mathbf{u}_t^{T}\mathbf{R}\mathbf{u}_t \tag{12}$$

where $\mathbf{x}_t^{g}$ is the target state. For a continuous MDP, iLQR or DDP is an effective approach to optimize the local policy $\hat{\pi}:\mathbf{u}_{0:H-1}$ in a closed loop as in [15], where a local dynamic model is created based on a first or second-order linear approximation of the learned dynamic model $\mathcal{F}$. It is very efficient for trajectory optimization in many robotic learning tasks. However, the traditional iLQR or DDP uses a finite short-term prediction of model to compute a locally optimal sequence of actions, which generally leads to a suboptimal policy. While, ADP adopts a value function to approximate a infinite-horizon discounted return, which can generally obtain a better solution. Whereas, compared with DDP, this method is not enough data-efficient. Some works [47], [48] proposed to substitute the terminal reward $r_f(\mathbf{x}_H)$ with an approximate value function $V(\mathbf{x}_H)$. But it is still not efficient enough for using a learned value function to compute the optimal policy directly.

In this article, to bridge the gap between the MB and MF learning processes, we propose two improvements: 1) the locally optimal policy should be searched near the MF guiding policy of actor network and 2) the terminal reward in MPC is replaced by locally optimal value $V(\mathbf{x}_H)$ that is computed by optimizing the corresponding state-action value function $Q(\mathbf{x}_H,\cdot)$. Assume the greedy policy $\hat{\pi}$ with respect to Q is deterministic, then $V(\mathbf{x}_H)=Q(\mathbf{x}_H,\hat{\pi}(\mathbf{x}_H))$. Formally, the optimization problem can be expressed as follows:

$$\begin{aligned}\max_{\hat{\pi}} \quad & \mathbb{E}_{p^{\nu}(\mathbf{x}),\hat{\pi}(\mathbf{u}|\mathbf{x})}\left[\sum_{t=0}^{H-1}\gamma^{t} r(\mathbf{x}_t,\mathbf{u}_t)+\gamma^{H} Q\big(\mathbf{x}_H,\hat{\pi}(\mathbf{x}_H)\big)\right] \\ \text{s.t.} \quad & \mathbf{x}_{t+1}\sim\mathcal{F}(\mathbf{x}_t,\mathbf{u}_t), \\ & d\big(\hat{\pi}(\mathbf{u}|\mathbf{x}),\pi_{\theta}(\mathbf{u}|\mathbf{x})\big)\leq\epsilon\end{aligned} \tag{13}$$

where $d(\cdot,\cdot)$ is a function that measures the closeness of two policies, and ϵ is the neighborhood size.

On the first point, we suggest the initial action sequence comes from the actor's policy $\mathbf{u}_{0:H}^{0}\sim\pi_{\theta}(\mathbf{u}|\mathbf{x}_{0:H})$, instead of being initialized randomly in the traditional methods. The MB optimization can modify the original habitual policy. On the second point, the terminal value $V(\mathbf{x}_H)$ is estimated by optimizing the corresponding state-action value function around the MF policy $\pi_{\theta}(\mathbf{x}_H)$. Since the terminal greedy policy $\hat{\pi}(\mathbf{x}_H)$ is not taken into account in traditional algorithms, this policy needs to be computed by maximizing the Q function. However, it is impossible to globally search the optimal Q value that is represented by a critic network in continuous space. Here, the terminal value function is approximated through optimizing the critic network with several information-theoretic constraints. Assume the greedy policy is guided by MF policy, then the optimization problem is formalized as follows.

$$\begin{aligned}\max_{\hat{\pi}} \quad & \mathbb{E}_{p^{\nu}(\mathbf{x}_H),\hat{\pi}(\mathbf{u}|\mathbf{x}_H)}[Q(\mathbf{x}_H,\mathbf{u})] \\ \text{s.t.} \quad & \mathbb{E}_{p^{\nu}(\mathbf{x}_H)}\big[\mathrm{KL}\big(\hat{\pi}(\mathbf{u}|\mathbf{x}_H)\|\pi_{\theta}(\mathbf{u}|\mathbf{x}_H)\big)\big]\leq\epsilon \\ & \mathbb{E}_{p^{\nu}(\mathbf{x}_H)}\big[\mathrm{H}\big(\hat{\pi}(\mathbf{u}|\mathbf{x}_H)\big)\big]\geq\kappa \\ & \mathbb{E}_{p^{\nu}(\mathbf{x}_H),\hat{\pi}(\mathbf{u}|\mathbf{x}_H)}=1\end{aligned} \tag{14}$$

where $p^{\nu}(\mathbf{x}_H)$ is the state distribution that complies with the policy $\hat{\pi}(\mathbf{u}|\mathbf{x}_H)$ and the learned dynamics. $\pi_{\theta}(\mathbf{u}|\mathbf{x}_H))$ is the guiding policy created by actor network parameterized by θ. In order to make the optimization problem feasible, several information-theoretic constraints are added. The first constraint is the Kullback–Leibler (KL) divergence between new policy and MF guiding policy, which can be used for limiting the loss of information between policy updates and preventing unstable learning. The second constraint is the entropy constraint of new policy that is crucial for controlling exploration and exploitation. The final constraint represents the integral of joint state-action probability is one.

1) Computation of the Terminal Value: The similar optimization problem is described in [49] and [50]. This optimization problem allows for a closed-form solution through the method of Lagrangian multipliers. The solution is given by

$$\hat{\pi}(\mathbf{u}|\mathbf{x}_H)\propto\pi_{\theta}(\mathbf{u}|\mathbf{x}_H)^{\frac{\eta^{*}}{\eta^{*}+\omega^{*}}}\exp\left[\frac{Q(\mathbf{x}_H,\mathbf{u})}{\eta^{*}+\omega^{*}}\right] \tag{15}$$

where $\eta^{*}\geq 0$ and $\omega^{*}\geq 0$ are optimal dual variables of KL and entropy constraints, respectively. These dual variables are computed by minimizing the following dual function:

$$\begin{aligned}g(\eta,\omega)=\ & \eta\epsilon-\omega\kappa+(\eta+\omega) \\ & \times\mathbb{E}_{p^{\nu}(\mathbf{x}_H)}\left[\log\int\pi_{\theta}(\mathbf{u}|\mathbf{x}_H)^{\frac{\eta}{\eta+\omega}}\exp\left(\frac{Q(\mathbf{x}_H,\mathbf{u})}{\eta+\omega}\right)d\mathbf{u}\right].\end{aligned} \tag{16}$$

The proof is given in the Appendix. When the greedy terminal action $\hat{\mathbf{u}}_H\sim\hat{\pi}(\mathbf{u}|\mathbf{x}_H)$ is obtained, the terminal value function can be computed as follows:

$$V(\mathbf{x}_H) = Q\big(\mathbf{x}_H, \hat{\mathbf{u}}_H\big) \tag{17}$$

$$V^x(\mathbf{x}_H) = \nabla_x Q\big(\mathbf{x}, \hat{\mathbf{u}}_H\big)|_{\mathbf{x}=\mathbf{x}_H} \tag{18}$$

$$V^{xx}(\mathbf{x}_H) = \nabla_x^2 Q\big(\mathbf{x}, \hat{\mathbf{u}}_H\big)|_{\mathbf{x}=\mathbf{x}_H}. \tag{19}$$

Then the general iLQR algorithm is used to optimize the initial sequences of actions $\mathbf{u}_{0:H}^0$ further. When the control's bounds are taken into account, the control can be computed by solving a quadratic program (QP) subject to the box constraints, as mentioned in [15].

2) Impact of Value Approximation Error on the Model-Based Policy: In this section, we aim to analyze how much impact the approximation error in Q value has on the quality of MB policy. Assume that the learned dynamics $\mathcal{F}$ is ideal, then the following theorem bounds the performance of the policy.

Lemma 1: An approximate state-action value function Q is obtained, and assume the approximate error is defined as: $||Q - Q^*||_\infty \le \epsilon$. Let the terminal reward $r_f(\mathbf{x}_H) = V(\mathbf{x}_H)$, and $\hat{\pi}$ be the greedy policy with respect to Q. Then for all states $\mathbf{x}$, the performance of the MPC policy can be bounded as

$$V^*(\mathbf{x}) - V(\mathbf{x}) \le \frac{2\gamma^H \epsilon}{1 - \gamma^H} \tag{20}$$

where $||Q||_\infty = \max_{x,u} |Q(x, u)|$.

Proof: Let $\hat{\tau}$ and τ^* represents the trajectories generated by greedy policy $\hat{\pi}$ and optimal policy π^*, respectively, on the MDP with planning horizon H. Define an operator $\mathcal{B}_\tau^H Q(\mathbf{x}_H, \pi(\mathbf{x}_H)) = \mathbb{E}_\tau[\sum_{t=0}^{H-1} \gamma^t r_t + \gamma^H Q(\mathbf{x}_H, \pi(\mathbf{x}_H))]$. Since $\hat{\pi}$ is the greedy policy with respect to Q and $\hat{\tau}$ is the corresponding trajectory, then there is

$$\mathcal{B}_{\hat{\tau}}^H Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) \ge \mathcal{B}_{\tau^*}^H Q\big(\mathbf{x}_H, \pi^*(\mathbf{x}_H)\big). \tag{21}$$

Starting from state $\mathbf{x}$ and using this condition, there is

$$\begin{aligned} &V^*(\mathbf{x}) - Q^*\big(\mathbf{x}, \hat{\pi}(\mathbf{x})\big) \\ &= \mathcal{B}_{\tau^*}^H V^*(\mathbf{x}_H) - \mathcal{B}_{\hat{\tau}}^H Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x})\big) \\ &= \mathcal{B}_{\tau^*}^H V^*(\mathbf{x}_H) - \mathcal{B}_{\hat{\tau}}^H Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) \\ &\quad + \mathcal{B}_{\hat{\tau}}^H Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - \mathcal{B}_{\hat{\tau}}^H Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) \\ &\le \mathcal{B}_{\tau^*}^H V^*(\mathbf{x}_H) - \mathcal{B}_{\hat{\tau}}^H Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) + \gamma^H \epsilon \\ &= \mathcal{B}_{\tau^*}^H Q^*\big(\mathbf{x}_H, \pi^*(\mathbf{x}_H)\big) - \mathcal{B}_{\hat{\tau}}^H Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) + \gamma^H \epsilon \\ &\le \mathcal{B}_{\tau^*}^H Q^*\big(\mathbf{x}_H, \pi^*(\mathbf{x}_H)\big) - \mathcal{B}_{\tau^*}^H Q\big(\mathbf{x}_H, \pi^*(\mathbf{x}_H)\big) + \gamma^H \epsilon \\ &\le 2\gamma^H \epsilon. \end{aligned} \tag{22}$$

Then there is

$$\begin{aligned} V^*(\mathbf{x}) - V(\mathbf{x}) &= \mathcal{B}_{\tau^*}^H V^*(\mathbf{x}_H) - \mathcal{B}_{\hat{\tau}}^H V(\mathbf{x}_H) \\ &= \mathcal{B}_{\tau^*}^H V^*(\mathbf{x}_H) - \mathcal{B}_{\hat{\tau}}^H Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) \\ &\quad + \mathcal{B}_{\hat{\tau}}^H Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - \mathcal{B}_{\hat{\tau}}^H V(\mathbf{x}_H) \\ &\le 2\gamma^H \epsilon + \mathcal{B}_{\hat{\tau}}^H V^*(\mathbf{x}_H) - \mathcal{B}_{\hat{\tau}}^H V(\mathbf{x}_H) \\ &= 2\gamma^H \epsilon + \gamma^H \mathbb{E}_{\hat{\tau}}\big[V^*(\mathbf{x}_H) - V(\mathbf{x}_H)\big] \\ &\le 2\gamma^H \epsilon \Big(1 + \gamma^H + \gamma^{2H} + \cdots\Big) \\ &\le \frac{2\gamma^H \epsilon}{1 - \gamma^H}. \end{aligned} \tag{23}$$

Thus, the discount factor and the planning horizon are the key variables for the performance of MB control if the learned dynamics is accurate. ■

3) Accelerating Convergence of the Value Function: The proposed MB control can accelerate convergence of the global value function, which improves the data efficiency. Assume that the learned dynamics $\mathcal{F}$ is ideal, then the value function has the following contraction property.

Lemma 2: An approximate state-action value function Q is obtained. Define an operator $\mathcal{B}_\tau^H Q(\mathbf{x}_H, \pi(\mathbf{x}_H)) = \mathbb{E}_\tau[\sum_{t=0}^{H-1} \gamma^t r_t + \gamma^H Q(\mathbf{x}_H, \pi(\mathbf{x}_H))]$. Let $\hat{\pi}$ is the greedy policy with respect to Q and $\hat{\tau}$ is the corresponding trajectory. Then the value function satisfies the following contraction property:

$$\begin{aligned} &||\mathcal{B}_{\hat{\tau}}^H Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - \mathcal{B}_{\hat{\tau}}^H Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big)||_\infty \\ &\quad \le \gamma^H ||Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big)||_\infty. \end{aligned} \tag{24}$$

Proof: Assume that $\mathcal{B}_{\hat{\tau}}^H Q(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)) \ge \mathcal{B}_{\hat{\tau}}^H Q^*(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H))$, then

$$\begin{aligned} &\mathcal{B}_{\hat{\tau}}^H Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - \mathcal{B}_{\hat{\tau}}^H Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) \\ &\quad = \gamma^H \big[Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big)\big] \\ &\quad \le \gamma^H ||Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big)||_\infty. \end{aligned} \tag{25}$$

In the same way, repeating this argument in case of $\mathcal{B}_{\hat{\tau}}^H Q(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)) \le \mathcal{B}_{\hat{\tau}}^H Q^*(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H))$

$$\begin{aligned} &\mathcal{B}_{\hat{\tau}}^H Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - \mathcal{B}_{\hat{\tau}}^H Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) \\ &\quad = \gamma^H \big[Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big)\big] \\ &\quad \le \gamma^H ||Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big)||_\infty. \end{aligned} \tag{26}$$

Thus there is

$$\begin{aligned} &|\mathcal{B}_{\hat{\tau}}^H Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - \mathcal{B}_{\hat{\tau}}^H Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big)| \\ &\quad \le \gamma^H ||Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big)||_\infty. \end{aligned} \tag{27}$$

Since the learned dynamic is assumed to be ideal, $\mathbf{x}_H$ can also represent the real state that is in the feasible region. For all $\mathbf{x}_H$, taking the supremum over $\mathbf{x}_H$, the following result is obtained further:

$$\begin{aligned} &||\mathcal{B}_{\hat{\tau}}^H Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - \mathcal{B}_{\hat{\tau}}^H Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big)||_\infty \\ &\quad \le \gamma^H ||Q\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big) - Q^*\big(\mathbf{x}_H, \hat{\pi}(\mathbf{x}_H)\big)||_\infty. \end{aligned} \tag{28}$$

In this article, iLQR with the constraint of terminal value function provides an efficient method to compute the term $\mathcal{B}_{\hat{\tau}}^H Q$, which improves the performance of the policy due to the smaller value error. Meanwhile, the good control results cause a faster convergence of value function. ■

4) Impact of Model Approximation Error on the Value Function: Both results mentioned above have a premise that the learned dynamics $\mathcal{F}$ is ideal. But there must be errors between the learned model and the true dynamics during learning. Inspired by the work [51], we find that the MB value using the learned dynamics gradually tends to the global optimal value as the model approximation error decreases. Specifically, the following conclusion is drawn.

Theorem 1: An approximate state-action value function Q is obtained. Define an operator $\mathcal{B}^H_{\hat{\tau}|\hat{M}}Q(\hat{\mathbf{x}}_H, \hat{\pi}(\hat{\mathbf{x}}_H)) = \mathbb{E}_{\hat{\tau}}[\sum_{t=0}^{H-1}\gamma^t r_t + \gamma^H Q(\hat{\mathbf{x}}_H, \hat{\pi}(\hat{\mathbf{x}}_H))]$, where $\hat{\pi}$ is the greedy policy with respect to Q and $\hat{\tau}$ is the corresponding trajectory based on the learned dynamics $\hat{M}$. Similarly, define an operator $\mathcal{B}^H_{\hat{\tau}|M^*}Q(\mathbf{x}^*_H, \hat{\pi}(\mathbf{x}^*_H))$ whose optimal trajectory is based the true dynamics M^*. Let the reward function r be L_r-Lipschitz over states $\mathbf{x}$, and the value function Q be L_Q-Lipschitz over states $\mathbf{x}$. Assume the model approximation errors satisfy $\max_{t\in[H]}\mathbb{E}[||\hat{\mathbf{x}}_t - \mathbf{x}^*_t||^2] \leq \epsilon^2$ in an H-step rollout. Then

$$\mathbb{E}\Big(\mathcal{B}^H_{\hat{\tau}|\hat{M}}Q - \mathcal{B}^H_{\hat{\tau}|M^*}Q\Big)^2 \leq \Big(c^2L_r^2 + 2\gamma^H cL_rL_Q + \gamma^{2H}L_Q^2\Big)\epsilon^2. \tag{29}$$

Proof: According to the definition above, there is

$$\begin{aligned}&\mathbb{E}\Big[\mathcal{B}^H_{\hat{\tau}|\hat{M}}Q\big(\hat{\mathbf{x}}_H, \hat{\pi}(\hat{\mathbf{x}}_H)\big) - \mathcal{B}^H_{\hat{\tau}|M^*}Q\big(\mathbf{x}^*_H, \hat{\pi}(\mathbf{x}^*_H)\big)\Big]^2\\ &\quad= \mathbb{E}\Big[\Big(\hat{R} - R^*\Big) - \gamma^H\big(Q\big(\hat{\mathbf{x}}_H, \hat{\pi}(\hat{\mathbf{x}}_H)\big) - Q\big(\mathbf{x}^*_H, \hat{\pi}(\mathbf{x}^*_H)\big)\big)\Big]^2\end{aligned} \tag{30}$$

where $\hat{R} - R^* = \sum_{t=0}^{H}\gamma^t\hat{r}_t - \sum_{t=0}^{H}\gamma^t r^*_t$.

For any L^2 random variables A and B, $\mathbb{E}[(A-B)^2]$ satisfies the following inequality by using the Cauchy–Schwarz inequality:

$$\begin{aligned}\mathbb{E}\Big[(A-B)^2\Big] &= \mathbb{E}A^2 - 2\mathbb{E}[AB] + \mathbb{E}B^2\\ &\leq \mathbb{E}A^2 + 2\mathbb{E}^{\frac{1}{2}}A^2\mathbb{E}^{\frac{1}{2}}B^2 + \mathbb{E}B^2.\end{aligned} \tag{31}$$

Thus there is

$$\begin{aligned}&\mathbb{E}\Big[\mathcal{B}^H_{\hat{\tau}|\hat{M}}Q\big(\hat{\mathbf{x}}_H, \hat{\pi}(\hat{\mathbf{x}}_H)\big) - \mathcal{B}^H_{\hat{\tau}|M^*}Q\big(\mathbf{x}^*_H, \hat{\pi}(\mathbf{x}^*_H)\big)\Big]^2\\ &\leq \mathbb{E}\Big(\hat{R} - R^*\Big)^2\\ &\quad+ 2\gamma^H\sqrt{\mathbb{E}\Big(\hat{R} - R^*\Big)^2\mathbb{E}\big[Q\big(\hat{\mathbf{x}}_H, \hat{\pi}(\hat{\mathbf{x}}_H)\big) - Q\big(\mathbf{x}^*_H, \hat{\pi}(\mathbf{x}^*_H)\big)\big]^2}\\ &\quad+ \gamma^{2H}\mathbb{E}\big[Q\big(\hat{\mathbf{x}}_H, \hat{\pi}(\hat{\mathbf{x}}_H)\big) - Q\big(\mathbf{x}^*_H, \hat{\pi}(\mathbf{x}^*_H)\big)\big]^2.\end{aligned} \tag{32}$$

Then all MB terms are bounded as follows:

$$\mathbb{E}\Big(\hat{R} - R^*\Big)^2 \leq \sum_{i,j}\gamma^{2(i+j)}\sqrt{\mathbb{E}\big(\hat{r}_i - r^*_i\big)^2\mathbb{E}\Big(\hat{r}_j - r^*_j\Big)^2}. \tag{33}$$

Since the reward function is L_r-Lipschitz over states $\mathbf{x}$, $||\hat{r}_i - r^*_i|| \leq L_r||\hat{\mathbf{x}}_i - \mathbf{x}^*_i||$. So

$$\mathbb{E}\Big(\hat{R} - R^*\Big)^2 \leq c^2L_r^2\epsilon^2 \tag{34}$$

where $c := \sum_{i,j}\gamma^{2(i+j)} \leq \min(H^2, (1-\gamma^2)^{-2})$.

According to the above assumption $d(\hat{\pi}, \pi_\theta) \leq \delta$ and the value function is L_Q-Lipschitz over states $\mathbf{x}$, there is

$$\begin{aligned}&\mathbb{E}\big[Q\big(\hat{\mathbf{x}}_H, \hat{\pi}(\hat{\mathbf{x}}_H)\big) - Q\big(\mathbf{x}^*_H, \hat{\pi}(\mathbf{x}^*_H)\big)\big]^2\\ &\leq \mathbb{E}\Big(L_Q^2||\hat{\mathbf{x}}_H - \mathbf{x}^*||^2\Big) = L_Q^2\epsilon^2.\end{aligned} \tag{35}$$

Taking all these results into consideration, the following conclusion can be drawn:

$$\begin{aligned}&\mathbb{E}\Big[\mathcal{B}^H_{\hat{\tau}|\hat{M}}Q\big(\hat{\mathbf{x}}_H, \hat{\pi}(\hat{\mathbf{x}}_H)\big) - \mathcal{B}^H_{\hat{\tau}|M^*}Q\big(\mathbf{x}^*_H, \hat{\pi}(\mathbf{x}^*_H)\big)\Big]^2\\ &\leq \Big(c^2L_r^2 + 2\gamma^H cL_rL_Q + \gamma^{2H}L_Q^2\Big)\epsilon^2.\end{aligned} \tag{36}$$

Obviously, with the model approximation error decreasing, MB value using the learned dynamics gradually tends to the optimal value.

It is worth noting that the MB control gradually transforms into the MF control with the decrease of planning horizon H. If H is relatively large, the MB control will obtain a high learning speed, but the meantime, it will spend more time to compute the policy at each step. Conversely, a smaller H leads to a faster computing time for each decision-making step, but causes a lower data-using efficiency. Particularly, when $H = 0$, the policy search of MB control transform into MF control called guide actor-critic (GAC) [50], where the greedy policy is searched by maximizing the global state-action value function around a guiding policy produced by the policy network:

$$\begin{aligned}\max_{\hat{\pi}}\quad & \mathbb{E}_{p^\chi(\mathbf{x}),\hat{\pi}(\mathbf{u}|\mathbf{x})}[Q(\mathbf{x}, \mathbf{u})] \qquad (37)\\ \text{s.t.}\quad & \mathbb{E}_{p^\chi(\mathbf{x})}\big[\mathrm{KL}\big(\hat{\pi}(\mathbf{u}|\mathbf{x})\|\pi_\theta(\mathbf{u}|\mathbf{x})\big)\big] \leq \epsilon\\ & \mathbb{E}_{p^\chi(\mathbf{x})}\big[\mathrm{H}\big(\hat{\pi}(\mathbf{u}|\mathbf{x})\big)\big] \geq \kappa\\ & \mathbb{E}_{p^\chi(\mathbf{x}),\hat{\pi}(\mathbf{u}|\mathbf{x})} = 1.\end{aligned}$$

As for the stability, the parameters of controlled plant play a crucial role of the stability of the system. Because the output policy is guided by the MF policy that is closely related to the plant parameters [52]. If the system itself is uncontrollable, it will be hard to control stably. From the point of basic algorithms, both iLQR and reinforcement learning can guarantee the stability under certain conditions [52], [53]. One of the key factors is the design of reward function, where the matrices of **Q** and **R** can influence the stability of the whole closed-loop system directly.

The pseudo-code description of H-step MB control is presented in Algorithm 1, where *backward* and *forward* passes of iLQR refer to [15]. ■

B. Computational Model of Emotion Processing

In this section, a computational model of amygdala is built to simulate emotion processing. As mentioned above, amygdala and PFC are the core regions in emotion processing, which integrate lots of sensory information to acquire, maintain, and regulate a variety of emotions. Anatomically, the amygdala consists of four major components: 1) lateral amygdala (LA); 2) basal amygdala (BA); 3) central amygdala (CeM); and 4) the intercalated (ITC) cell clusters [54], [55], whose connections are shown in Fig. 3. These four nuclei have different properties and serve distinct roles in emotion processing. More specifically, the LA can receive and process some conditioned or unconditioned stimuli at the beginning, and then project them to the BA and dorsal ITC neurons (ITCd) that connect to ventral ITC cells (ITCv) further. The BA cells mainly send excitatory inputs to the ITC neurons and

```
Algorithm 1 $H$-Step MB Control
Inputs: Observation $\mathbf{x}_0 \leftarrow \mathbf{x}_t$, the dynamic model $\mathcal{F}_\phi(\cdot)$,
    the reward function $r(\cdot)$, the critic function $Q_\varphi(\cdot)$ and the
    policy function $\pi_\theta(\cdot)$.
Outputs: Action $\mathbf{u}_t$.
    for $k = 0, \ldots, H-1$ do
        Select action $\mathbf{u}_k = \pi_\theta(\mathbf{x}_k)$.
        Predict the next state $\mathbf{x}_{k+1} \sim \mathcal{F}_\phi(\mathbf{x}_k, \mathbf{u}_k)$.
        Store transition $(\mathbf{x}_k, \mathbf{u}_k) \rightarrow \boldsymbol{\tau}_k$
    end for
    for $i = 0, \ldots, N_{iter}$ do
        Compute the difference set $\mathcal{D}$, where the terminal value
        is obtained by optimize the critic function $Q_\varphi$:
        $\mathcal{D} = diff(\boldsymbol{\tau}_{0:H-1}, r, Q_\varphi, \pi_\theta)$.
        Run the backward pass of iLQR:
        $\mathbf{I}, \mathbf{K} = backward(\boldsymbol{\tau}_{0:H-1}, \mathcal{D})$.
        Run the forward pass of iLQR to update the trajectory:
        $\hat{\boldsymbol{\tau}}_{0:H-1} = forward(\tau_{0:H-1}, \mathbf{I}, \mathbf{K})$
    end for
    return $\mathbf{u}_t \leftarrow \hat{\mathbf{u}}_0$
```

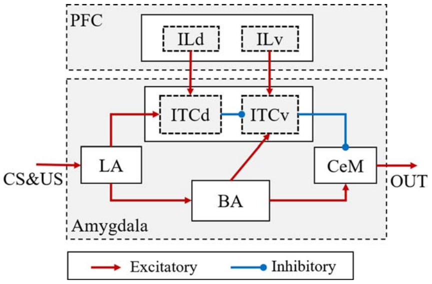

Fig. 3. Amygdala circuitry for processing conditioned fear [54].

CeM neurons, while the cells located in ITC are responsible for generating inhibitory singles to depress the activities of the CeM cells. These two paths serve an opponent processing to balance the response of emotion. Meanwhile, the amygdala receives some mediating information from the infralimbic (IL) cortex located in vmPFC that is implicated in extinction of conditioned fear responses. The studies of neuroscience show that the IL mainly projects these excitatory signals to the ITC.

Based on the anatomical structure of amygdala and its extended circuits, an emotion-processing network is presented to simulate simple cognitive-emotional interactions following [56]. The schematic is shown in Fig. 4. The activity of each neuron follows the shunting dynamics, which was developed based on the dynamics of membrane voltage proposed by Hodgkin and Huxley [57] and Grossberg [58]. The activity of each node is analogous to short-term memory (STM) and adaptive weights between nodes are regarded as long-term memory (LTM). Formally, the dynamics of the ith neuron can be written as the following shunting STM equation:

$$\tau_x \frac{dx_i}{dt} = -Ax_i + (B - x_i)S_i^+ - (C + x_i)S_i^- \tag{38}$$

where each STM trace is bounded within an interval $[-C, B]$, and S_i^+ and S_i^- correspond to the excitatory and inhibitory inputs, respectively. A is a passive decay rate, and τ_x is the time constant of neural integration.

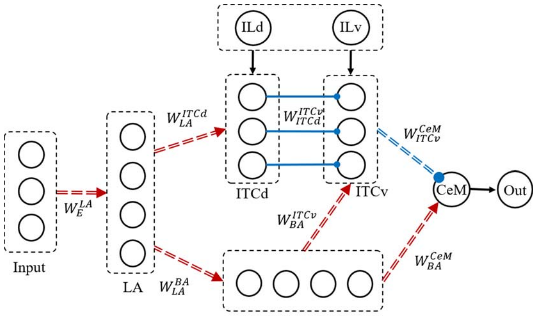

Fig. 4. Schematic of emotion-processing network based on the amygdala circuitry and the interaction between amygdala and vmPFC.

TABLE I
EXCITATORY AND INHIBITORY INPUTS TO EACH NETWORK COMPONENT

Cell	Excitation term S^+	Inhibition term S^-
x_i^{LA}	$f_L \sum W_{E_k}^{LA_i} E_k$	0
x_i^{BA}	$\sum W_{LA_k}^{BA_i} [x_k^{LA}]^+$	0
x_i^{ITCd}	$\sum W_{LA_k}^{ITCd_i} [x_k^{LA}]^+ + E_{ILd}$	0
x_i^{ITCv}	$\sum W_{BA_k}^{ITCv_i} [x_k^{BA}]^+ + E_{ILv}$	$W_{ITCd_i}^{ITCv_i} [x_i^{ITCd}]^+$
x^{CeM}	$\sum W_{BA_k}^{CeM} [x_k^{BA}]^+$	$f_C \sum W_{ITCv_k}^{CeM} [x_k^{ITCv}]^+$

Table I specifies the excitatory and inhibitory inputs to each network component. N denotes the number of cells, $\sum(\cdot) = \sum_{k=1}^{N}(\cdot)$, and $[\,\cdot\,]^+$ is rectified linear unit. W is a synaptic weight between two neurons. To train the synaptic weights to form LTM, a novel Oja-like learning rule is proposed to integrate active forgetting and reinforcement processes, which demonstrates effectiveness in classical fear-conditioning simulation. Specifically, all weights but $W_{\text{ITCd}}^{\text{ITCv}}$ can be modified based on reinforcing signal, and presynaptic and postsynaptic neural activities. There are two cases to be discussed.

Case 1: When the weights is updated at each time step, the learning rule is defined as follows:

$$\Delta W_t = \frac{\Delta t}{\tau_W}\left(-DW_t + R_t\left[x_t^{\text{post}}\right]^+\left(\left[x_t^{\text{pre}}\right]^+\right)^T - R_t^2 W_t\right) \tag{39}$$

where D is a passive decay rate, and $-DW_t$ term drives the active forgetting process. Δt is the step size in simulation, and τ_W is the constant of integration. $R_t[x_t^{\text{post}}]^+([x_t^{\text{pre}}]^+)^T - R_t^2 W_t$ term is inspired by Oja learning rule. If x^{pre} and x^{post} are activated simultaneously, and the teaching signal R_t is positive, the $R_t[x_t^{\text{post}}]^+([x_t^{\text{pre}}]^+)^T$ term will be strengthened. The $R_t^2 W_t$ is a reward-related forgetting term, which prevents the weights from growing unlimitedly.

Case 2: If the teaching signal is sparse, and the reward is only received at the end of each episode, the weights is updated via the following learning rule:

$$\triangle W_n = \frac{\Delta t}{\tau_W T}\left(-DW_n + \left(R_n - \bar{R}_n\right)\sum_{t=1}^{T}\left[x_t^{\text{post}}\right]^+\left(\left[x_t^{\text{pre}}\right]^+\right)^T - \left(R_n - \bar{R}_n\right)^2 W_n\right) \tag{40}$$

where $\bar{R}$ represents the average reward baseline at the nth episode. A simple approach to estimate this baseline is to compute a moving average of actual rewards:

$$\bar{R}_n = (1-\alpha)\bar{R}_{n-1} + \alpha R_n \tag{41}$$

where $0 < \alpha < 1$.

To evaluate this computational network of amygdala, the classical fear conditioning is used as a test case. This behavioral paradigm is a type of classical conditioning that involves associating an aversive unconditional stimulus (US, such as an electric shock) with either a conditional stimulus (CS, such as a tone). It exhibits flexible acquisition and extinction of stimulus-triggered emotional responses (such as fear), and shows an implicit emotional learning process. The learning process usually includes four epochs: 1) fear acquisition; 2) fear extinction; 3) fear retrieval; and 4) extinction retrieval. At the first epoch, the CS always co-occurs with the US, and the animal will express conditional fear (such as the freezing response in rodents) in case of CS after training. But at the second epoch, only CS is repeated in absence of US, which causes fear extinction gradually. The third epoch is the same as the first epoch, the conditional fear will be aroused rapidly. The last epoch is the same as the second epoch, the fear response will extinct again.

In the simulation, the parameters of neural dynamics are set to the same value in each subnetwork, with $N = 20, A = 10, B = 2, C = 0$, and $\tau_x = 0.1$. The simulation is run for 10 000 time steps, with step size of $\Delta t = 0.001$. Each epoch lasts for 2500 time steps. In fact, the characteristics of the synaptic weights have a significant impact on the performance of network. Assume that synaptic decay is much faster in the near-output subnetwork. Specifically, we choose $\tau_W = 0.4, D_E^{\text{LA}} = 0.01, D_{\text{LA}}^{\text{BA}} = 0.01, D_{\text{LA}}^{\text{ITCd}} = 0.8, D_{\text{BA}}^{\text{ITCv}} = 0.8, D_{\text{BA}}^{\text{CeM}} = 1.2, D_{\text{ITCv}}^{\text{CeM}} = 1.2, f_L = 5$, and $f_C = 10$. During training, when the US is present, the weights $W_{\text{LA}}^{\text{BA}}$, $W_{\text{LA}}^{\text{ITCd}}$, and $W_{\text{BA}}^{\text{CeM}}$ receive a positive reinforcing signal with $R^+ = 1$. While $W_{\text{BA}}^{\text{ITCv}}$ and $W_{\text{ITCv}}^{\text{CeM}}$ is assumed to receive a complementary reinforcing signal $R^- = (1 - R^+)$, due to the inhibition from the activation of ITCv. The weights are updated following (39).

Fig. 5 shows the output of cell x^{CeM}. The blue line represents the change of conditional fear without high-level cognitive regulation ($E_{ILd} = 0$ and $E_{ILv} = 0$), while the red line shows the emotional response with $E_{ILd} = 0$ and $E_{ILv} = 10$. Obviously, similar emotional changes occur in both cases. After repeated pairings of CS and US between 0–2.5 s, the animal gradually learns to fear the CS signal. As the US disappears at the second epoch, the fear fades away due to the active forgetting of synaptic weights. When the US co-occurs with the CS again between 5–7.5 s, the fear memory is recalled faster than one at the first epoch. Meanwhile, high-level modulation from *ILv* can bias the intensity of fear. A high E_{ILv} input increases inhibition onto the activity of *CeM*, which leads to a relatively low response in fear acquisition and a faster extinction of fear. Hence, it may be an important path for top-down cognitive control of emotional response and learning. According to the neural dynamics mentioned above, the ITCd, ITCv, and CeM cell reach the following equilibrium value in response of CS when $C = 0$:

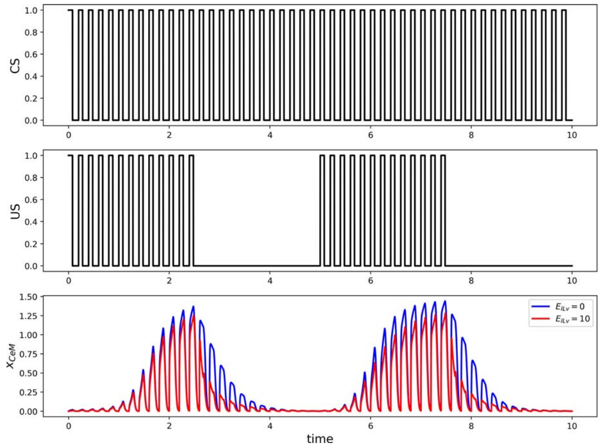

Fig. 5. Simulation of classical fear conditioning, including fear acquisition, extinction, fear retrieval, and extinction. The first and second panels show the signal of conditioned and unconditioned stimulus, respectively. The activity of output cell is recorded in the last panel.

$$\bar{x}_i^{\text{ITCd}} = \frac{B\left(\sum W_{LA_k}^{ITCd_i}\left[x_k^{\text{LA}}\right]^+ + E_{ILd}\right)}{A + \sum W_{LA_k}^{\text{ITCd}_i}\left[x_k^{\text{LA}}\right]^+ + E_{ILd}} \tag{42}$$

$$\bar{x}_i^{\text{ITCv}} = \frac{B\left(\sum W_{BA_k}^{\text{ITCv}_i}\left[x_k^{\text{BA}}\right]^+ + E_{ILv}\right)}{A + \sum W_{BA_k}^{\text{ITCv}_\text{i}}\left[x_k^{\text{BA}}\right]^+ + E_{ILv} + W_{\text{ITCd}_i}^{\text{ITCv}_i}\left[x_i^{\text{ITCd}}\right]^+} \tag{43}$$

$$\bar{x}^{\text{CeM}} = \frac{B\sum W_{BA_k}^{\text{CeM}}\left[x_k^{\text{BA}}\right]^+}{A + \sum W_{BA_k}^{CeM}\left[x_k^{\text{BA}}\right]^+ + f_C\sum W_{\text{ITCv}_k}^{\text{CeM}}\left[x_k^{\text{ITCv}}\right]^+} \tag{44}$$

where $\sum(\cdot) = \sum_{k=1}^{N}(\cdot)$.

It is easy to see that the increasing excitatory input E_{ILd} can enhance the fear-related responses to some extent due to the inhibitory effect of *ILd* to ITCv. On the contrary, the input E_{ILv} plays an important role to suppress the emotional responses and accelerate the extinction of emotion.

In this article, the uncertainty-related response mainly results from three basic factors: 1) SPE; 2) RPE; and 3) episodic reinforcing signal (ERS). The SPE is fed into the network as input, and the RPE is used as high-level regulatory factors E_{ILv}. The ERS is used to change the synaptic weights at the end of each episode. SPE directly reflects the accuracy of understanding the external environment, which is highly related to some affective states such as surprise and curiosity. The smaller the SPE is, the more reliable long-horizon MB control is. The RPE is usually reported by dopamine neurons in animal learning theory, which can trigger a series of reward-related emotional reactions such as positive/negtive emotion (pleasure and desire). This term is strongly associated with the performance of value network that estimates the discount returns in the future. If the RPE is smaller, more short-term MB or MF planning is encouraged. Except for these immediate signals, some long-term feedback also bias the emotional state, and then influence decision-making. Herein, the ERS is introduced to modify the weights of emotion processing network.

In this article, SPE and PRE are defined based on the predictions of state and reward. Assume each state is subject to Gaussian distribution $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ and the predicted states $p(\hat{\mathbf{x}}) = \mathcal{N}(\hat{\boldsymbol{\mu}}_x, \hat{\boldsymbol{\Sigma}}_x)$. The KL divergence between the predicted state and real state is applied to quantify the SPE

$$\mathrm{SPE} = \mathrm{KL}\big(p(\hat{\mathbf{x}})\|p(\mathbf{x})\big). \tag{45}$$

The SPE signal is further bounded in (0, 1) through a sigmoid-type function. In like manner, the RPE is the difference between a return that is being received and the return that is predicted to be received. The loss of critic network is used to compute this value, which is also constrained within (0, 1) by a sigmoid-type function

$$\mathrm{RPE} = \big[Q(\mathbf{x}, \mathbf{u}) - \big(r(\mathbf{x}, \mathbf{u}) + \gamma Q'\big(\mathbf{x}', \pi_\theta(\mathbf{x}')\big)\big)\big]^2. \tag{46}$$

The ERS is the cumulative rewards $R = \sum_{t=1}^{T} r_t$ after each episode, and the average reward baseline is updated by according to (40). The episodic reward error $R - \bar{R}$ is activated by tanh-type function so that it is segmented to form excitatory or inhibitory signal.

Finally, the activation of output cell x_{CeM} represents the intensity of emotional response that implies the uncertainty of long-term MB policy (or the certainty of short-term policy). The planning horizon is a deceasing function of the this emotional variable. If the SPE is large and the RPE is small, the uncertainty of the long-term planning is high, which lead to the decrease of H. Conversely, if the SPE is small and the RPE is large, the uncertainty is low so that more long-term planning is encouraged. Formally, the horizon is defined as follows:

$$H = \min\left(0, <H_{\max} - ke>\right) \tag{47}$$

where e denotes the intensity of uncertainty-related emotional response, and k is a gain. $< \cdot >$ is the round function.

V. Simulations

The proposed algorithm is integrated into the control architecture of two simulated systems. First, the swing-up task of inverted pendulum is conducted to verify the effectiveness. Then, the reaching task is performed with a simulated robotic arm that has higher state dimensions.

A. Inverted Pendulum Swing-Up Task

In this task, the cart slides freely along a rod bounded in range $[-1, 1]$ as shown in Fig. 6, aiming to swing up the pole with as little energy as possible. There are five dimensions of the sensory observation $\mathbf{x} = [x, v, \dot{\theta}, \cos\theta, \sin\theta]$, that refer to the position and velocity of cart, the swinging angle and angular speed of pole, respectively. The control command includes one continuous variable $\mathbf{u} = [u]$, where $u \in (-1, 1)$, that represents the input of motor actuator.

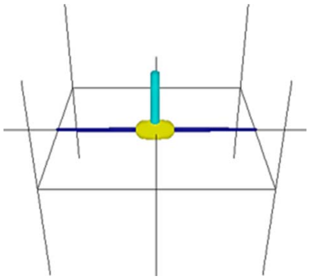

Fig. 6. Scene of the inverted pendulum swing-up task.

During interacting with the environment, the agent learns a probabilistic transition dynamic model from sequential sensory observations and actions. Here, a ensemble of five probabilistic neural networks is used to predict the next state. Each neural network is fed with twenty particles sampling from the distribution of current state $p(\mathbf{x})$, which is for capturing the aleatoric uncertainty. Additionally, each subnetwork is a multilayer perceptron with two layers, and each of them consists of 200 neurons. In all simulations, the actor network has two layers, and each of them consists of 128 neurons. The critic network has the same structure but with a larger learning rate $\eta_Q = 1 \times 10^{-3}$ than the one in actor network $\eta_\pi = 1 \times 10^{-4}$. The target state is set as $x^g = [0, 0, 0, 1, 0]$ that implies the swinging angle is zero. In the reward function, $\mathbf{Q}$ is a matrix that is filled with 0 but $\mathbf{Q}(4, 4) = 1$, and $\mathbf{R} = [0.01]$.

For comparing the performance of the different approaches, four sets of experiments are designed. The first one is using traditional iLQR control. Then our proposed H-step MB control is used to control the cart, where the planning horizon is set to 0, 2, 4, and 6, respectively. When $H = 0$, the controller corresponds to the GAC algorithm [50]. The auto-differentiation function in PyTorch is adopted to obtain the gradient of learned dynamics model with respect to state $\mathbf{x}$ and action $\mathbf{u}$. During optimization, The parameter of KL constraint is chosen as $\delta = 1 \times 10^{-4}$. The lower bound of the policy entropy is set as $\kappa = 0.05$. *L-BFGS-B* algorithm is used to optimize the dual function for computing the dual variables η^* and ω^*. The final one is integrating MB and MF control with emotion modulation. The parameters are the same as in simulation above, expect $A = 20$ for a faster responding. During training, the weights $W_{\mathrm{LA}}^{\mathrm{BA}}$, $W_{\mathrm{LA}}^{\mathrm{ITCd}}$, and $W_{\mathrm{BA}}^{\mathrm{CeM}}$ receive a positive reinforcing signal with $dR^+ = R - \bar{R}$. While $W_{\mathrm{BA}}^{\mathrm{ITCv}}$ and $W_{\mathrm{ITCv}}^{\mathrm{CeM}}$ is assumed to receive a complementary reinforcing signal $dR^- = -dR^+$.

In the experiment, the whole learning process lasts 50 episodes, and each of them contains 200 time steps. The planning horizon of iLQR is chosen as 25 here, and a shorter horizon results in a very poor performance. The cumulative rewards are summarized in Fig. 7. Obviously, the iLQR has the highest data efficiency in the early stage, which benefits from a short-term planning with the certain reward function. However, this algorithm usually leads to a suboptimal strategy in the later stage so that the cumulative reward is relatively low. By

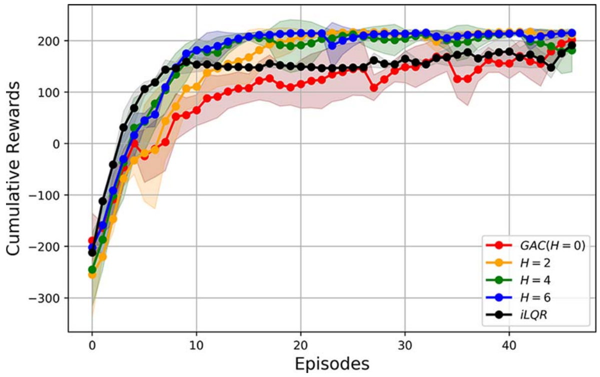

Fig. 7. Cumulative rewards of iLQR and our proposed MB control with different horizons.

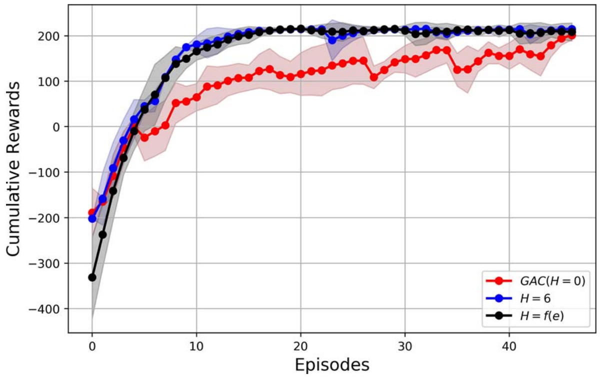

Fig. 8. Cumulative rewards of MF, six-step MB, and emotion-modulated control.

contrast, the proposed method aims to maximize the infinite-horizon discount returns so that a better policy is obtained. But different planning time has a significant impact on the performance in H-step control. Theoretically, the longer the planning horizon, the better the performance in terms of learning efficiency. But, generally, the accuracy of model prediction decreases with the increasing planning horizon. Thus, the longer planning horizon can improve learning efficiency within a limited range. This is also verified by the results in Fig. 7. When the $H = 0$, the MB control transforms into the MF GAC control that is not data-efficient obviously. As the horizon H increases, the proposed MB control enables a faster learning.

After that, the performance of the MB, MF and emotion-modulated control are compared further. The cumulative rewards are shown in Fig. 8. Obviously, the emotion-modulated method reaches to a higher reward value with less time, which shows a higher data efficiency. Emotion-modulated control can change the planning horizon adaptively according to the uncertainty of long-term MB planning. The 6th, 26th, and 46th epoch are selected, respectively, to show the change of emotional response as in Fig. 9. At each epoch, if the SPE is large and the RPE is small, the intensity of uncertainty-related emotional response will increase, such that more short-term MB planning or MF policy will be adopted for a faster decision-making. On the contrary, if the SPE is small and the RPE is large, the intensity of emotional response is low such that more long-term planning is encouraged to produce more accurate control.

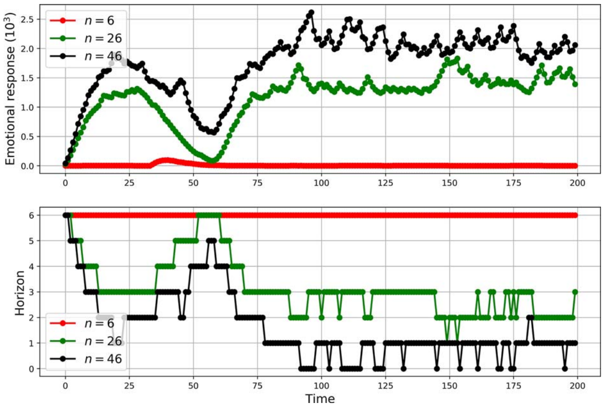

Fig. 9. Change of emotional response and the corresponding planning horizon.

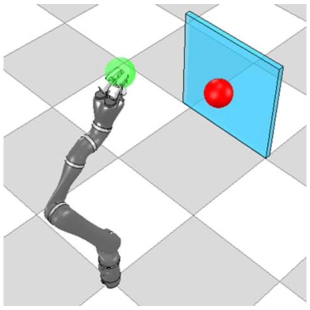

Fig. 10. Scene of the Jaco reaching task in V-REP.

As skills become more proficient, certainty of the short-term MB planning or MF policy increases due to the reinforcing effect of positive rewards to the weights of emotion-processing network. In the early stage, with the accumulation of environmental knowledge, the agent makes more decisions from long-term MB planning. While, in the later stage, more short-term controls are accepted and MF habitual controls come into being. According to (40) and (41), the average reward baseline represents the expected reward based on previous experience. If the received reward is better than this baseline, the synaptic weights will be strengthen, and vice versa. This kind of plasticity regulates the slow change of certainty-related emotional response, which is significant to accelerate the speed of decision-making.

B. Jaco Reaching Task

For more complex control problems, the effectiveness of our proposed method is investigated in a reaching task with a simulated robotic arm in V-REP (Fig. 10). The physical Jaco assistive robotic arm, developed by Kinova Robotics, is assisting people with limited or no upper limb mobility to do something safely. This arm has six joints that can be controlled by a position, velocity or torque controller through ROS packages. Herein, torque control is adopted to implement more flexible and faster manipulations, where each motor command represents the desired torque of the corresponding joint.

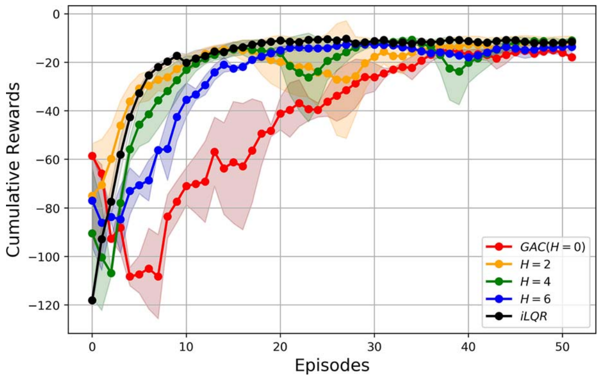

Fig. 11. Cumulative rewards of iLQR and our proposed MB control with different horizons in the reaching task.

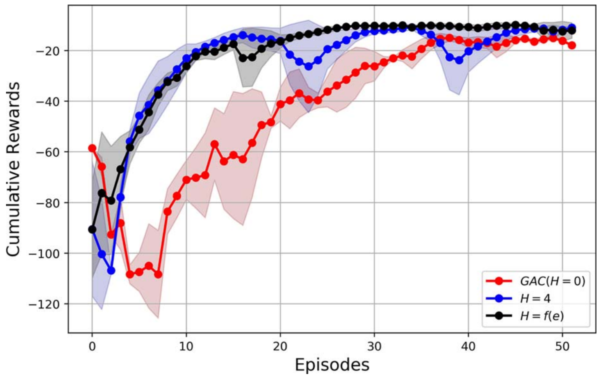

Fig. 12. Cumulative rewards of MF, four-step MB, and emotion-modulated control in the reaching task.

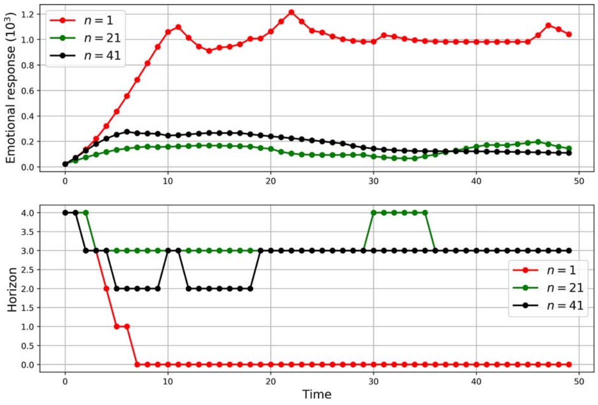

Fig. 13. Change of emotional response and the corresponding planning horizon in the reaching task.

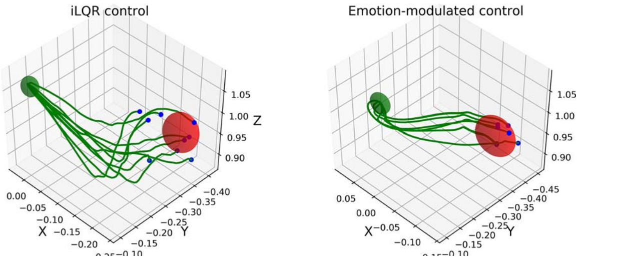

Fig. 14. Trajectories of end effector produced by iLQR and emotion-modulated control.

Reaching process is very important in many robotic manipulations, such as variant grasp and assembly tasks. In the experiment, the robotic arm aims to control its hand to reach the red goal located at $[-0.2, -0.37, 0.95]$ from the initial position, as shown in Fig. 10. There are 21 dimensions of the sensory observation $\mathbf{x} \in \mathbb{R}^{21}$, including angle, angular velocity of all joints and position of end effector. The motor command consists of six continuous variables $\mathbf{u} \in \mathbb{R}^{6}$, which represents the desired torque of motor actuators.

All settings of this experiment is similar to the above inverted pendulum swing-up task expect for the size of networks and parameters of emotion module. More specifically, the structure of probabilistic ensemble neural network for model learning is also composed of five subnetworks with three layers, each of which consists of 200 neurons. The size of both actor and critic network is set as [128, 128]. In the simulation, the robotic arm interacts with environment for totally 50 times, and each trial includes 50 time steps. After five tests, the cumulative rewards are shown in Figs. 11 and 12.

As a result, all algorithms have no problem to find an optimal sequence of actions to control the actuators. Traditional iLQR policy has the highest learning efficiency, meanwhile, the suboptimality of the strategy does not show up in this task. However, the planning horizon of iLQR is chosen as 25 here, and a shorter horizon results in a very poor performance. Hence, the higher data efficiency results from long-term dynamic programming at the expense of computing time. On the contrary, MF control is able to make decisions quickly, but learns slowly obviously. Additionally, the increasing H can improve the learning efficiency, but reduce the accuracy of model prediction as well. As reflected in Fig. 11, due to the complexity of real dynamics, the error of six-step model prediction has a worse impact on the performance of learning than two- or four-step predictions.

In the emotion-modulated control, the hyper-parameters are set as $H_{\max} = 4$ and $k = 5\times10^{-3}$. The learning result is shown in Fig. 12, where the emotion-modulated control has a better efficiency of learning. Emotion modulation can shift the MB and MF processes adaptively in the course of skill acquisition. Early in training, the large errors of state and reward prediction awaken an uncertainty-related emotional response that facilitates more short-term planning or MF control. As the training continues, the decreasing SPE strengthens the certainty-related emotion, which is conducive to improve the accuracy through a long-term planning. With over-training, the RPE also decreases so that certainty of the short-term decision gradually increases. More short-term planning or MF control becomes dominant again and the habitual behavior is gradually formed. The change of the emotional response and planning horizon are drawn in Fig. 13 at the 1st, 21st, and 41st epoch, respectively.

After training, the trajectories of end effector are drawn in Fig. 14, where the traditional iLQR control and emotion-modulated control are used to generate the strategies of motion, respectively. As shown in the figure, the emotion-modulated decision system outperforms the pure iLQR control in terms of uncertainty of learned policy, which may result from the suboptimality of the locally MB strategy.

The reaching process is a part of robotic manipulation, which is relatively easy if the goal is fixed. In the future, some new brain-inspired integration framework of MB and MF control will be required to perform more complex decision-making tasks, such as tracking [59], grasp [60], and assembly [61]. Especially, it is a significant research direction of high-level decision-making that using the underlying environmental knowledge to reason out control strategies. Therein, more key neural mechanisms in human brain will be incorporated into the computational models of perception, cognition, and decision for intelligent robots.

VI. Conclusion

According to the accumulating neural evidence, human behaviors are often separated into by two decision systems: 1) a deliberative MB system for guiding goal-directed behaviors and 2) a reflexive MF system for driving habitual behaviors. And emotion is one of the most important factors in modulating these two paths. In this article, a computational model of decision-making process with modulation of emotion is proposed to improve the efficiency of learning and speed up the decision-making in the robotic control tasks. Two main contributions have been made.

1) A novel decision-making framework is proposed to build a bridge between MB and MF decision-making processes through adjusting the planning horizon.
2) A biologically plausible computational model is built to simulate emotion processing and modulating the planning horizon based on the intensity of uncertainty-related emotional response. The emotional response is aroused by the SPE and RPE in the course of tasks.

The experimental results show two conclusions.

1) The new proposed MB reasoning can improve efficiency and stability of learning compared with MF approaches. Meanwhile, it allows for better policies beyond local MB solutions.
2) Emotion modulation can shift these two parts of decision-making systems well in terms of the learning efficiency and the speed of decision-making.

Appendix

The closed-form solution of the terminal policy is obtained by optimizing the problem as (14). The Lagrangian function of this problem is

$$\begin{aligned}\mathcal{L}\big(\hat{\pi}, \eta, \omega, \nu\big) = {} & \mathbb{E}_{p^{\nu}(\mathbf{x}_H), \hat{\pi}(\mathbf{u}|\mathbf{x}_H)}[Q(\mathbf{x}_H, \mathbf{u})] \\ & + \eta\big(\epsilon - \mathbb{E}_{p^{\nu}(\mathbf{x}_H)}\big[\mathrm{KL}\big(\hat{\pi}(\mathbf{u}|\mathbf{x}_H) \| \pi_{\theta}(\mathbf{u}|\mathbf{x}_H)\big)\big]\big) \\ & + \omega\big(\mathbb{E}_{p^{\nu}(\mathbf{x}_H)}\big[\mathrm{H}\big(\hat{\pi}(\mathbf{u}|\mathbf{x}_H)\big)\big] - \kappa\big) \\ & + \nu\big(\mathbb{E}_{p^{\nu}(\mathbf{x}_H), \hat{\pi}(\mathbf{u}|\mathbf{x}_H)} - 1\big). \end{aligned} \tag{48}$$

The partial derivative with respect to $\hat{\pi}$ is

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \hat{\pi}} = \mathbb{E}_{p^{\nu}(\mathbf{x}_H)}\bigg[\int & Q(\mathbf{x}_H, \mathbf{u}) - (\eta + \omega)\log \hat{\pi}(\mathbf{u}|\mathbf{x}_H) \\ & - \eta \log \pi_{\theta}(\mathbf{u}|\mathbf{x}_H) + (\nu - \eta - \omega) d\mathbf{u}\bigg]. \end{aligned} \tag{49}$$

Let $(\partial \mathcal{L}/\partial \hat{\pi}) = 0$, we have

$$\begin{aligned}\hat{\pi}(\mathbf{u}|\mathbf{x}_H) &= \pi_{\theta}(\mathbf{u}|\mathbf{x}_H)^{\frac{\eta}{\eta+\omega}} \exp\bigg(\frac{Q(\mathbf{x}_H, \mathbf{u})}{\eta + \omega}\bigg) \exp\bigg(-\frac{\eta + \omega - \nu}{\eta + \omega}\bigg) \\ &\propto \pi_{\theta}(\mathbf{u}|\mathbf{x}_H)^{\frac{\eta}{\eta+\omega}} \exp\bigg(\frac{Q(\mathbf{x}_H, \mathbf{u})}{\eta + \omega}\bigg). \end{aligned} \tag{50}$$

In order to obtain the solution in a closed form, assume that the MF policy is subject to a Gaussian distribution

$$\pi_{\theta}(\mathbf{u}|\mathbf{x}_H) = \mathcal{N}\big(\mathbf{u}|\boldsymbol{\mu}_{\theta}(\mathbf{x}_H), \boldsymbol{\Sigma}_{\theta}(\mathbf{x}_H)\big). \tag{51}$$

Additionally, the critic $Q(\mathbf{x}_H, \mathbf{u})$ can be locally estimated through Taylor series expansion up to the second order. The Taylor's approximation at an arbitrary action $\mathbf{u}_0$ is given by

$$\begin{aligned}Q(\mathbf{x}_H, \mathbf{u}) \approx {} & Q(\mathbf{x}_H, \mathbf{u}_0) + (\mathbf{u} - \mathbf{u}_0)^T \mathbf{g}_0(\mathbf{x}_H) \\ & + \frac{1}{2}(\mathbf{u} - \mathbf{u}_0)^T \mathbf{H}_0(\mathbf{x}_H)(\mathbf{u} - \mathbf{u}_0) + \mathcal{O}\Big(\|\mathbf{u}\|^3\Big) \end{aligned} \tag{52}$$

where $\mathbf{g}_0(\mathbf{x}) = \nabla_u Q(\mathbf{x}_H, \mathbf{u})|_{\mathbf{u}=\mathbf{u}_0}$ is the gradient of the critic w.r.t $\mathbf{u}$ at $\mathbf{u}_0$, and $\mathbf{H}_0(\mathbf{x}_H) = \nabla_u^2 Q(\mathbf{x}_H, \mathbf{u})|_{\mathbf{u}=\mathbf{u}_0}$ is the Hessian. The auto-differentiation function in PyTorch is adopted to obtain this gradient and Hessian matrix.

While assume the higher-order term $\mathcal{O}(\|\mathbf{u}\|^3)$ is too small to be ignored, the Taylor's approximation is rewritten as a quadratic form as follows:

$$Q_0(\mathbf{x}_H, \mathbf{u}) = \frac{1}{2}\mathbf{u}^T \mathbf{H}_0 \mathbf{u} + \mathbf{u}^T \mathbf{G}_0(\mathbf{x}_H) + \mathbf{B}_0(\mathbf{x}_H) \tag{53}$$

where $\mathbf{G}_0(\mathbf{x}_H) = \mathbf{g}_0(\mathbf{x}_H) - \mathbf{H}_0(\mathbf{x}_H)\mathbf{u}_0$ and $\mathbf{B}_0(\mathbf{x}_H) = (1/2)\mathbf{u}_0^T \mathbf{H}_0(\mathbf{x}_H)\mathbf{u}_0 + \mathbf{u}_0^T \mathbf{g}_0(\mathbf{x}_H) + Q(\mathbf{x}_H, \mathbf{u}_0)$. The local policy is also a Gaussian distribution

$$\hat{\pi}(\mathbf{u}|\mathbf{x}_H) = \mathcal{N}\Big(\mathbf{u}|\hat{\boldsymbol{\mu}}(\mathbf{x}_H), \hat{\boldsymbol{\Sigma}}(\mathbf{x}_H)\Big) \tag{54}$$

where the mean and covariance are given by

$$\hat{\boldsymbol{\mu}}(\mathbf{x}_H) = \mathbf{F}(\mathbf{x}_H)\mathbf{L}(\mathbf{x}_H), \quad \hat{\boldsymbol{\Sigma}}(\mathbf{x}_H) = \big(\eta^* + \omega^*\big)\mathbf{F}(\mathbf{x}_H) \tag{55}$$

where $\mathbf{F}(\mathbf{x}_H) = [\eta^* \boldsymbol{\Sigma}_{\theta}^{-1}(\mathbf{x}_H) - \mathbf{H}_0(\mathbf{x}_H)]^{-1}$ and $\mathbf{L}(\mathbf{x}_H) = \eta^* \boldsymbol{\Sigma}_{\theta}^{-1}(\mathbf{x}_H)\boldsymbol{\mu}_{\theta}(\mathbf{x}_H) + \mathbf{G}_0(\mathbf{x}_H)$,

We substitute the solution to the Lagrangian function. Then there is

$$\begin{aligned}\mathcal{L}(\eta, \omega) &= \eta\epsilon - \omega\kappa + \mathbb{E}_{p^{\nu}(\mathbf{x}_H)}(\eta + \omega - \nu) \\ &= \eta\epsilon - \omega\kappa + \mathbb{E}_{p^{\nu}(\mathbf{x}_H)}\big[-(\eta + \omega)\log \hat{\pi}(\mathbf{u}|\mathbf{x}_H) \\ &\quad + \eta \log \pi_{\theta}(\mathbf{u}|\mathbf{x}_H) + Q(\mathbf{x}_H, \mathbf{u})\big] \\ &= \eta\epsilon - \omega\kappa + \mathbb{E}_{p^{\nu}(\mathbf{x}_H)}\big[\eta \log \pi_{\theta}(\mathbf{u}|\mathbf{x}_H) \\ &\quad - (\eta + \omega)\log \hat{\pi}(\mathbf{u}|\mathbf{x}_H)\big] + c \\ &= \hat{g}(\eta, \omega) \end{aligned} \tag{56}$$

where c is a constant.

Thus, the dual variables η^* and ω^* are obtained by minimizing the following dual function:

$$\begin{aligned}\hat{g}(\eta, \omega) = {} & \eta\epsilon - \omega\kappa + (\eta + \omega)\mathbb{E}_{p^{\nu}(\mathbf{x}_H)}\left[\log \sqrt{\frac{|2\pi(\eta + \omega)\mathbf{F}(\mathbf{x}_H)|}{|2\pi \boldsymbol{\Sigma}_{\theta}(\mathbf{x}_H)|^{\frac{\eta}{\eta+\omega}}}}\right] \\ & + \frac{1}{2}\mathbb{E}_{p^{\nu}(\mathbf{x}_H)}\Big[\mathbf{L}(\mathbf{x}_H)^T \mathbf{F}(\mathbf{x}_H)\mathbf{L}(\mathbf{x}_H) - \eta \boldsymbol{\mu}_{\theta}(\mathbf{x}_H)^T \boldsymbol{\Sigma}_{\theta}^{-1}(\mathbf{x}_H)\boldsymbol{\mu}_{\theta}(\mathbf{x}_H)\Big] \\ & + c. \end{aligned} \tag{57}$$

This dual function is rewritten as

$$\begin{aligned}\hat{g}(\eta,\omega) = \eta\epsilon - \omega\kappa + \frac{1}{2}\mathbb{E}_{p^{\nu}(\mathbf{x}_H)}\Big[&(\eta+\omega)\log|2\pi(\eta+\omega)\mathbf{F}(\mathbf{x}_H)| \\ &- \eta\log|2\pi\boldsymbol{\Sigma}_\theta(\mathbf{x}_H)| + \mathbf{L}(\mathbf{x}_H)^T\mathbf{F}(\mathbf{x}_H)\mathbf{L}(\mathbf{x}_H) \\ &- \eta\boldsymbol{\mu}_\theta(\mathbf{x}_H)^T\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{x}_H)\boldsymbol{\mu}_\theta(\mathbf{x}_H)\Big] + c. \end{aligned} \tag{58}$$

We have

$$\begin{aligned}\nabla_\eta\mathbf{F}(\mathbf{x}_H) &= -\mathbf{F}(\mathbf{x}_H)\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{x}_H)\mathbf{F}(\mathbf{x}_H) \\ \nabla_\eta\mathbf{L}(\mathbf{x}_H) &= \boldsymbol{\Sigma}_\theta^{-1}(\mathbf{x}_H)\boldsymbol{\mu}_\theta(\mathbf{x}_H). \end{aligned} \tag{59}$$

Then there is

$$\begin{aligned}\nabla_\eta\hat{g} = \epsilon + \frac{1}{2}\mathbb{E}_{p^{\nu}(\mathbf{x}_H)}\Big[&\log|2\pi(\eta+\omega)\mathbf{F}(\mathbf{x}_H)| \\ &+ \Big(N_u - (\eta+\omega)tr\Big(\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{x}_H)\mathbf{F}(\mathbf{x}_H)\Big)\Big) - \log|2\pi\boldsymbol{\Sigma}_\theta(\mathbf{x}_H)| \\ &+ \mathbf{L}(\mathbf{x}_H)^T\nabla_\eta\mathbf{F}(\mathbf{x}_H)\mathbf{L}(\mathbf{x}_H) + 2\mathbf{L}(\mathbf{x}_H)^T\mathbf{F}(\mathbf{x}_H)\nabla_\eta\mathbf{L}(\mathbf{x}_H) \\ &- \boldsymbol{\mu}_\theta(\mathbf{x}_H)^T\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{x}_H)\boldsymbol{\mu}_\theta(\mathbf{x}_H)\Big] \end{aligned} \tag{60}$$

$$\nabla_\omega\hat{g} = -\kappa + \frac{1}{2}\big[1 + \log|2\pi(\eta+\omega)\mathbf{F}(\mathbf{x}_H)|\big] \tag{61}$$

where N_u represents the dimension of $\mathbf{u}$.

References

[1] D. R. Bach and P. Dayan, "Algorithms for survival: A comparative perspective on emotions," *Nat. Rev. Neurosci.*, vol. 18, no. 5, pp. 311–319, 2017.

[2] E. A. Phelps, K. M. Lempert, and P. Sokol-Hessner, "Emotion and decision making: Multiple modulatory neural circuits," *Annu. Rev. Neurosci.*, vol. 37, no. 1, pp. 263–287, 2014.

[3] H. Qiao, Y. L. Li, F. F. Li, X. Y. Xi, and W. Wu, "Biologically inspired model for visual cognition achieving unsupervised episodic and semantic feature learning," *IEEE Trans. Cybern.*, vol. 46, no. 10, pp. 2335–2347, Oct. 2016.

[4] P. J. Yin *et al.*, "A novel biologically inspired visual cognition model: Automatic extraction of semantics, formation of integrated concepts, and reselection features for ambiguity," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 2, pp. 420–431, Jun. 2018.

[5] J. Chen, S. Zhong, E. Kang, and H. Qiao, "Realizing human-like manipulation with a musculoskeletal system and biologically inspired control scheme," *Neurocomputing*, vol. 339, pp. 116–129, Apr. 2019.

[6] D. Wang, H. B. He, and D. R. Liu, "Adaptive critic nonlinear robust control: A survey," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3429–3451, Oct. 2017.

[7] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.

[8] Z. H. Huang, X. Xu, H. B. He, J. Tan, and Z. P. Sun, "Parameterized batch reinforcement learning for longitudinal control of autonomous land vehicles," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 4, pp. 730–741, Apr. 2019.

[9] J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1889–1897.

[10] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, 2016.

[11] Z. Yang, K. Merrick, L. Jin, and H. A. Abbass, "Hierarchical deep reinforcement learning for continuous action control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5174–5184, Nov. 2018.

[12] M. P. Deisenroth and C. E. Rasmussen, "PILCO: A model-based and data-efficient approach to policy search," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, Bellevue, WA, USA, 2011, pp. 465–472.

[13] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1334–1373, 2015.

[14] F. Z. Xiong *et al.*, "Guided policy search for sequential multitask learning," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 1, pp. 216–226, Jan. 2019.

[15] Y. Tassa, N. Mansard, and E. Todorov, "Control-limited differential dynamic programming," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2014, pp. 1168–1175.

[16] W. Li and E. Todorov, "Iterative linear quadratic regulator design for nonlinear biological movement systems," in *Proc. ICINCO (1)*, 2004, pp. 222–229.

[17] W. Gu, J. Yao, Z. Yao, and J. Zheng, "Output feedback model predictive control of hydraulic systems with disturbances compensation," *ISA Trans.*, vol. 88, pp. 216–224, May 2019.

[18] I. Lenz, R. A. Knepper, and A. Saxena, "DeepMPC: Learning deep latent features for model predictive control," in *Proc. Robot. Sci. Syst. XI*, Rome, Italy, 2015. [Online]. Available: http://www.roboticsproceedings.org/rss11/p12.pdf

[19] N. D. Daw, Y. Niv, and P. Dayan, "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control," *Nat. Neurosci.*, vol. 8, no. 12, pp. 1704–1711, 2005.

[20] M. Belkaid, N. Cuperlier, and P. Gaussier, "Autonomous cognitive robots need emotional modulations: Introducing the eMODUL model," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 1, pp. 206–215, Jan. 2019.

[21] T. M. Moerland, J. Broekens, and C. M. Jonker, "Emotion in reinforcement learning agents and robots: A survey," *Mach. Learn.*, vol. 107, no. 2, pp. 443–480, 2017.

[22] X. Huang, W. Wu, H. Qiao, and Y. Ji, "Brain-inspired motion learning in recurrent neural network with emotion modulation," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 4, pp. 1153–1164, Dec. 2018.

[23] B. W. Balleine and A. Dickinson, "Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates," *Neuropharmacology*, vol. 37, nos. 4–5, pp. 407–419, 1998.

[24] M. Khamassi and M. D. Humphries, "Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies," *Front. Behav. Neurosci.*, vol. 6, p. 79, Nov. 2012.

[25] J. Zsuga, K. Biro, C. Papp, G. Tajti, and R. Gesztelyi, "The 'proactive' model of learning: Integrative framework for model-free and model-based reinforcement learning utilizing the associative learning-based proactive brain concept," *Behav. Neurosci.*, vol. 130, no. 1, pp. 6–18, 2016.

[26] C. K. Tang, A. P. Pawlak, V. Prokopenko, and M. O. West, "Changes in activity of the striatum during formation of a motor habit," *Eur. J. Neurosci.*, vol. 25, no. 4, pp. 1212–1227, 2007.

[27] E. Y. Kimchi, M. M. Torregrossa, J. R. Taylor, and M. Laubach, "Neuronal correlates of instrumental learning in the dorsal striatum," *J. Neurophysiol.*, vol. 102, no. 1, pp. 475–489, 2009.

[28] H. H. Yin and B. J. Knowlton, "Contributions of striatal subregions to place and response learning," *Learn. Memory*, vol. 11, no. 4, pp. 459–463, 2004.

[29] H. H. Yin and B. J. Knowlton, "The role of the basal ganglia in habit formation," *Nat. Rev. Neurosci.*, vol. 7, no. 6, pp. 464–476, 2006.

[30] C. A. Thorn, H. Atallah, M. Howe, and A. M. Graybiel, "Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning," *Neuron*, vol. 66, no. 5, pp. 781–795, 2010.

[31] A. M. Bornstein and N. D. Daw, "Multiplicity of control in the basal ganglia: Computational roles of striatal subregions," *Current Opin. Neurobiol.*, vol. 21, no. 3, pp. 374–380, 2011.

[32] H. H. Yin, S. B. Ostlund, and B. W. Balleine, "Reward-guided learning beyond dopamine in the nucleus accumbens: The integrative functions of cortico-basal ganglia networks," *Eur. J. Neurosci.*, vol. 28, no. 8, pp. 1437–1448, 2008.

[33] M. Verweij, T. J. Senior, D. J. F. Domínguez, and R. Turner, "Emotion, rationality, and decision-making: How to link affective and social neuroscience with social theory," *Front. Neurosci.*, vol. 9, p. 332, Sep. 2015.

[34] S. S. Cho *et al.*, "Investing in the future: Stimulation of the medial prefrontal cortex reduces discounting of delayed rewards," *Neuropsychopharmacology*, vol. 40, no. 3, pp. 546–553, 2014.

[35] J. W. Kable and P. W. Glimcher, "The neural correlates of subjective value during intertemporal choice," *Nat. Neurosci.*, vol. 10, no. 12, pp. 1625–1633, 2007.

[36] K. Okada, K. Toyama, Y. Inoue, T. Isa, and Y. Kobayashi, "Different pedunculopontine tegmental neurons signal predicted and actual task rewards," *J. Neurosci.*, vol. 29, no. 15, pp. 4858–4870, 2009.

[37] W. Schultz, "Erratum to: Reward functions of the basal ganglia," *J. Neural Transm.*, vol. 124, no. 9, p. 1159, 2017.

[38] A. R. Otto, C. M. Raio, A. Chiang, E. A. Phelps, and N. D. Daw, "Working-memory capacity protects model-based learning from stress," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 52, pp. 20941–20946, 2013.

[39] L. Schwabe and O. T. Wolf, "Stress prompts habit behavior in humans," *J. Neurosci.*, vol. 29, no. 22, pp. 7191–7198, 2009.

[40] A. F. T. Arnsten, “Stress signalling pathways that impair prefrontal cortex structure and function,” *Nat. Rev. Neurosci.*, vol. 10, no. 6, pp. 410–422, 2009.

[41] J. Li, D. Schiller, G. Schoenbaum, E. A. Phelps, and N. D. Daw, “Differential roles of human striatum and amygdala in associative learning,” *Nat. Neurosci.*, vol. 14, no. 10, pp. 1250–1252, Sep. 2011.

[42] P. H. Rudebeck, A. R. Mitz, R. V. Chacko, and E. A. Murray, “Effects of amygdala lesions on reward-value coding in orbital and medial prefrontal cortex,” *Neuron*, vol. 80, no. 6, pp. 1519–1531, 2013.

[43] S. W. Lee, S. Shimojo, and J. P. O’Doherty, “Neural computations underlying arbitration between model-based and model-free learning,” *Neuron*, vol. 81, no. 3, pp. 687–699, 2014.

[44] J. B. Rosen and M. P. Donley, “Animal studies of amygdala function in fear and uncertainty: Relevance to human research,” *Biol. Psychol.*, vol. 73, no. 1, pp. 49–60, 2006.

[45] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6402–6413.

[46] K. Chua, R. Calandra, R. McAllister, and S. Levine, “Deep reinforcement learning in a handful of trials using probabilistic dynamics models,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4754–4765.

[47] K. Lowrey, A. Rajeswaran, S. M. Kakade, E. Todorov, and I. Mordatch, “Plan online, learn offline: Efficient learning and exploration via model based control,” in *Proc. 7th Int. Conf. Learn. Represent. (ICLR)*, 2019.

[48] M. Y. Zhong, M. Johnson, Y. Tassa, T. Erez, and E. Todorov, “Value function approximation and model predictive control,” in *Proc. IEEE Symp. Adapt. Dyn. Program. Reinforcement Learn. (ADPRL), IEEE Symp. Comput. Intell. (SSCI)*, 2013, pp. 100–107.

[49] R. Akrour, A. Abdolmaleki, H. Abdulsamad, J. Peters, and G. Neumann, “Model-free trajectory-based policy optimization with monotonic improvement,” *J. Mach. Learn. Res.*, vol. 19, no. 1, pp. 565–589, 2018.

[50] V. Tangkaratt, A. Abdolmaleki, and M. Sugiyama, “Guide actor–critic for continuous control,” in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, 2018.

[51] V. Feinberg, A. Wan, I. Stoica, M. I. Jordan, J. E. Gonzalez, and S. Levine, “Model based value estimation for efficient model free reinforcement learning,” *CoRR*, 2018.

[52] F. L. Lewis and D. Vrabie, “Reinforcement learning and adaptive dynamic programming for feedback control,” *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, Aug. 2009.

[53] V. Roulet, D. Drusvyatskiy, S. Srinivasa, and Z. Harchaoui, “Iterative linearized control: Stable algorithms and complexity guarantees,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5518–5527.

[54] G. S. Li, “Computational models of the amygdala in acquisition and extinction of conditioned fear,” in *The Amygdala—Where Emotions Shape Perception, Learning and Memories*. London, U.K.: IntechOpen, 2017. [Online]. Available: https://www.intechopen.com/

[55] G. Li, T. Amano, D. Pare, and S. S. Nair, “Impact of infralimbic inputs on intercalated amygdala neurons: A biophysical modeling study,” *Learn. Memory*, vol. 18, no. 4, pp. 226–240, 2011.

[56] Y. J. John, D. Bullock, B. Zikopoulos, and H. Barbas, “Anatomy and computational modeling of networks underlying cognitive-emotional interaction,” *Front. Human Neurosci.*, vol. 7, no. 14, p. 101, 2013.

[57] A. L. Hodgkin and A. F. Huxley, “A quantitative description of membrane current and its application to conduction and excitation in nerve,” *J. Physiol.*, vol. 117, no. 4, pp. 500–544, 1952.

[58] S. Grossberg, “Recurrent neural networks,” *Scholarpedia*, vol. 8, no. 2, p. 1888, 2013.

[59] Q. Zhou, S. Zhao, H. Li, R. Lu, and C. Wu, “Adaptive neural network tracking control for robotic manipulators with dead zone,” *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[60] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” *Int. J. Robot. Res.*, vol. 37, nos. 4–5, pp. 421–436, 2018.

[61] T. Inoue, G. De Magistris, A. Munawar, T. Yokoya, and R. Tachibana, “Deep reinforcement learning for high precision assembly tasks,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Vancouver, BC, Canada, 2017, pp. 819–825.

Xiao Huang received the B.S. degree in guidance, navigation, and control from Central South University, Changsha, China, in 2015. He is currently pursuing the Ph.D. degree in control theory and control engineering with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His current research interests include brain-inspired computing, affective computing, and machine learning.

Wei Wu received the B.Sc. degree in physics and the M.Sc. degree in theoretical physics from Beijing Normal University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computational neuroscience from Johann Wolfgang Goethe University, Frankfurt, Germany, in 2008.

He is currently an Associate Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing.

Hong Qiao (SM’06–F’18) received the B.Eng. degree in hydraulics and control and the M.Eng. degree in robotics from Xi’an Jiaotong University, Xi’an, China, in 1986 and 1989, respectively, the M.Phil. degree in robotics control from the Industrial Control Center, University of Strathclyde, Strathclyde, U.K., in 1992, and the Ph.D. degree in robotics and artificial intelligence from De Montfort University, Leicester, U.K., in 1995.

She was a University Research Fellow with De Montfort University from 1995 to 1997. She was a Research Assistant Professor with the Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong, from 1997 to 2000, where she was an Assistant Professor from 2000 to 2002. Since 2002, she has been a Lecturer with the School of Informatics, University of Manchester, Manchester, U.K. She is currently a Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. She first proposed the concept of the attractive region in strategy investigation, which has successfully been applied by herself in robot assembly, robot grasping, and part recognition. She has authored the book entitled *Advanced Manufacturing Alert* (Wiley, 1999). Her current research interests include information-based strategy investigation, robotics and intelligent agents, animation, machine learning, and pattern recognition.

Prof. Qiao is currently an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS and the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING. She is the Editor-in-Chief of the *Assembly Automation*. She is currently a member of the Administrative Committee of the IEEE Robotics and Automation Society, the IEEE Medal for Environmental and Safety Technologies Committee, the Early Career Award Nomination Committee, the Most Active Technical Committee Award Nomination Committee, and the Industrial Activities Board for RAS.