SCIENCE CHINA
Information Sciences

• LETTER •

May 2021, Vol. 64 159203:1–159203:3
https://doi.org/10.1007/s11432-018-9779-6

Learning continuous coupled multi-controller coefficients based on actor-critic algorithm for lower-limb exoskeleton

Guangkui SONG[1], Rui HUANG[1], Hong CHENG[1*], Jing QIU[1], Qiming CHENG[1] & Shuai FAN[2]

[1]*Center for Robotics, University of Electronic Science and Technology of China, Chengdu* 611731*, China;*
[2]*School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu* 611731*, China*

Received 23 August 2018/Revised 30 September 2018/Accepted 1 February 2019/Published online 22 May 2020

Citation Song G K, Huang R, Cheng H, et al. Learning continuous coupled multi-controller coefficients based on actor-critic algorithm for lower-limb exoskeleton. Sci China Inf Sci, 2021, 64(5): 159203, https://doi.org/10.1007/s11432-018-9779-6

Dear editor,

Human-powered lower exoskeletons are widely studied by academia and industry with regard to human locomotion and strength augmentation. Technological developments have boosted the use of machine learning toward improving the control performance of exoskeleton systems. Here, reinforcement learning (RL) is used to adapt to changes to pilots and walking patterns [1–3]. The continuous observation space is discretized to facilitate the use of RL in real-life applications. Current methods using discretized observation spaces cannot store the exact value functions or policies for each separate state or state-action pair. In real-life applications, almost all tasks of interest, and most notably physical control tasks, have continuous high-dimensional observation spaces. Therefore, we present a novel algorithm called interactive learning strategy based on actor-critic (ILAC) to learn the continuous coupled coefficients of the proposed multi-controller in continuous high-dimensional observation spaces. The actor-critic algorithm [4, 5] is used for learning the coefficients of the multi-controller so as to improve the interaction performance of exoskeletons in continuous high-dimensional observation spaces. The proposed ILAC is similar to that in our previous studies, wherein the multi-controller consists of a model-based controller and an auxiliary controller (Figure 1(a)). The multi-controller adapts to inaccurate dynamic models and changes to pilots and walking patterns. We introduce the auxiliary controller to handle a case where the values of the dynamic model's parameters obtained by system identification are less than those of the dynamic model of the actual exoskeleton. As the coefficients of the two controllers are related to each other, excessive compensating forces through the amplification of the model-based controller can lead to stronger interaction forces. The relationship between the model-based controller and the auxiliary controller is managed by one learning algorithm. Thus, the coefficients are called coupled coefficients in this study. The coefficients of the two controllers are learned simultaneously, not only to adapt to changes to different pilots and walking patterns and the inaccurate dynamic models, but also to conform to the coupling relationship of the coefficients. Experiments on a human-powered augmentation lower exoskeleton (HUALEX) and a single degree-of-freedom (DOF) exoskeleton are conducted and described in this study. The experimental results show that the proposed ILAC provides better performance in interactive learning environments.

Our algorithm. As seen in Figure 1(a), the control structure is detailed with a single DOF exoskeleton platform. The single DOF exoskeleton consists of a thigh and a shank, and they are connected by a joint, which is powered by a bi-directional linear hydraulic actuator. The torque τ_e gained from the model-based controller and the auxiliary controller results in torque τ_c. A pilot leg is attached to the exoskeleton leg by compliant connections. The pilot can impose forces causing equivalent torque τ_h on the exoskeleton leg through the compliant connections. The dynamic model of the single DOF exoskeleton is written as

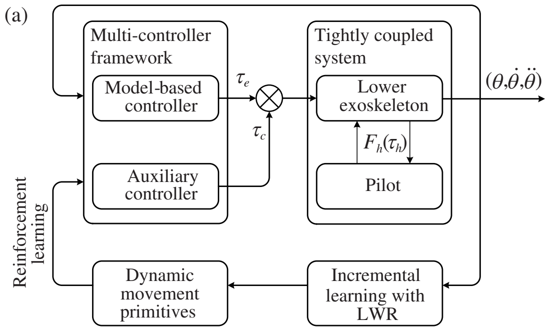
$$J\ddot{\theta}_e + B\dot{\theta}_e + mgl \cdot \sin\theta_e = \tau_e + \tau_c + \tau_h, \tag{1}$$

where J, B, m, and l represent the inertial moment, viscous friction coefficient, exoskeleton shank mass, and length of the single DOF exoskeleton, respectively. The joint states θ_e, $\dot{\theta}_e$, and $\ddot{\theta}_e$ represent the angle, angular velocity, and angular acceleration of the knee joint, respectively, and the gravitational constant is represented as g.

In order to reduce the complexity of the exoskeletons sensor system, a model-based controller, sensitivity amplification control (SAC), is utilized to control the exoskeleton. This controller is represented as

$$\tau_e = mgl \cdot \sin\theta_e + (1-\alpha^{-1})(\hat{J}\ddot{\theta}_e + \hat{B}\dot{\theta}_e), \tag{2}$$

[*] Corresponding author (email: hcheng@uestc.edu.cn)



(b)

Time (s)	Pilot A	Pilot B	Pilot C
Actor	23	24	32
Critic	4	3.5	3

Figure 1 (a) Control diagram of multi-controller consisting of a model-based controller and an auxiliary controller; (b) compensation of convergence time of the cost function

where $\hat{J}$ and $\hat{B}$ are the estimated inertial moment and viscous friction coefficient of the exoskeletons, respectively. In order to manage the changes to pilots and walking patterns, the sensitivity factors of the SAC are adapted to the changing of the physical human-robot interaction (pHRI) between the pilot and exoskeleton [1–3]. However, the sensitivity factors are limited to between zero and one to meet robustness requirements [6], which makes the strategies powerless for cases in which the values of the dynamic model's parameters obtained by the system identification method are less than those of the dynamic model of the actual exoskeleton. Thus, an auxiliary controller is introduced to manage such cases. The auxiliary controller is designed as

$$\tau_c = P(\theta_e - \theta_h) + D(\dot{\theta}_e - \dot{\theta}_h), \tag{3}$$

where P is a proportion gain, D is a derivative gain, and θ_h and $\dot{\theta}_h$ are the angle and angular velocity of the pilot's knee joint, respectively. The angle and angular velocity of the pilot are predicted using the dynamic movement primitives (DMPs) model, and a locally weighted regression (LWR) method is used to update the parameters of the DMPs [2].

In order to adapt the changes to pilots and walking patterns, we choose an actor-critic algorithm to learn the coefficients of the multi-controller. Here, the learning agent is split into two separate entities: the actor and the critic (policy and value function). The actor is responsible for generating a control input a, given the current state s. The critic is responsible for processing the rewards it receives [5]. Thus, an actor network $\mu(s|\theta^\mu)$ specifies the current policy by mapping states to a specific action with the weights θ^μ. A critic network $Q(s,a|\theta^Q)$ indicating the value function with the weights θ^Q is learned by using the Bellman equation, as in Q-learning. The actor network is updated by applying the chain rule to the expected return from the start distribution J with respect to the actor's parameters:

$$\begin{aligned}&\nabla_{\theta^\mu} J \approx \mathbb{E}_{s_t\sim\rho^\beta}[\nabla_{\theta^\mu} Q(s,a|\theta^Q)|_{s=s_t,a=\mu(s_t|\theta^\mu)}]\\ &=\mathbb{E}_{s_t\sim\rho^\beta}[\nabla_a Q(s,a|\theta^Q)|_{s=s_t,a=\mu(s_t)}\nabla_{\theta_\mu}\mu(s|\theta^\mu)|_{s=s_t}],\end{aligned} \tag{4}$$

where ρ^β denotes the discounted state visitation distribution for a policy β. To update the critic function $Q(s,a|\theta^Q)$, a copy of the actor and critic networks, $\mu'(s|\theta^\mu)$ and $Q'(s,a|\theta^Q)$, are used for calculating the target values. The weights of the target networks are updated by tracking the learned networks $\theta' \leftarrow \tau\theta+(1-\tau)\theta'$ with $\tau \ll 1$. This means that the target values are constrained to change slowly. The temporal difference (TD) error is used to update the critic. The TD error is estimated as

$$y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'}), \tag{5}$$

where $\gamma \in [0,1]$ is a discounting factor, $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$, and $\mathcal{N}_t$ is an exploring noise. The critic network is updated by minimizing the loss:

$$L = \frac{1}{N}\sum_i (y_i - Q(s_i,a_i|\theta^Q))^2. \tag{6}$$

In the proposed ILAC, the controller of the exoskeleton observes the current state s_t, executes an action a_t, receives an immediate reward r_t, and then observes the next state s_{t+1}. The immediate reward r_t is written as

$$r_t = -[k_1(\theta_e - \theta_h) + k_2(\dot{\theta}_e - \dot{\theta}_h)], \tag{7}$$

where k_1 and k_2 are weighted parameters. In the proposed ILAC, the states of the pilot are obtained from the DMPs models using the parameters that have been learned [2]. The states of the exoskeleton leg s_t at time t, and the actions of the exoskeleton leg a_t are described as follows:

$$s_t \to (\theta_e, \dot{\theta}_e, \ddot{\theta}_e), \quad a_t \to (\alpha, P, D). \tag{8}$$

Experimental results. Our algorithm is validated on a HUALEX system and a single DOF exoskeleton simulation platform. The hip joints and knee joints of the HUALEX are active joints. Each of them is activated by a hydraulic cylinder. The ankle joints are designed as an energy storage mechanism, which stores energy in the stance phase and releases it in the swing phase. Many compliant connections are used to connect the pilot to the HUALEX.

In the HUALEX experiments, three pilots (A, B, and C) of different heights (168 cm, 170 cm, and 180 cm, respectively) are chosen to operate the system in different environments (flat, sloping, and stairs) with different walking speeds. Each pilot operates the HUALEX system in the three environments in different orders. The weights of the actor network and critic network are initialized randomly before each pilot begins operating the HUALEX system.

Figure 1(b) shows the time spent in the learning process of the proposed ILAC for the HUALEX system with different pilots in different environments, and the times are roughly the same and acceptable in physical applications. After the online learning, the HUALEX system shows a good performance with regard to dealing with different pilots and walking patterns in varied environments.

In the single DOF exoskeleton simulation experiments, we use different frequencies and amplitudes to simulate various pilots and walking patterns. The results show the control performance of the ILAC for different dynamic model errors of the exoskeleton and the comparison with the strategies using discretized observation spaces (nMSE: 0.004 rad of continuous policies compared with 0.010 rad of discrete policies) with 20% model error. We can conclude that the proposed algorithm can ensure that the exoskeleton system follows the pilot's motion with fewer tracking errors than discrete policies.

The experimental results of the HUALEX system and single DOF exoskeleton show that the proposed ILAC can identify an optimal continuous policy in continuous high-dimensional domains subjected to different pilots and walking patterns.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2017YFB1302300) and National Natural Science Foundation of China (Grant Nos. 61673088, 61503060). Moreover, the authors gratefully acknowledge the contribution of an anonymous reviewer's comments.

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Huang R, Cheng H, Chen Q, et al. Interactive learning for sensitivity factors of a human-powered augmentation lower exoskeleton. In: Proceedings of IEEE International Conference on Intelligent Robots and Systems (IROS), Hamburg, 2015. 6409–6415

2 Huang R, Cheng H, Guo H, et al. Hierarchical interactive learning for a human-powered augmentation lower exoskeleton. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), Stockholm, 2016. 257–263

3 Huang R, Cheng H, Guo H, et al. Hierarchical learning control with physical human-exoskeleton interaction. Inf Sci, 2018, 432: 584–595

4 Lillicrap T P, Hunt J J, Pritzel A. Continuous control with deep reinforcement learning. 2015. ArXiv: 1509.02971

5 Grondman I, Busoniu L, Lopes G A D, et al. A survey of actor-critic reinforcement learning: standard and natural policy gradients. IEEE Trans Syst Man Cybern C, 2012, 42: 1291–1307

6 Kazerooni H, Racine J L, Huang L, et al. On the control of the Berkeley lower extremity exoskeleton (BLEEX). In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), Barcelona, 2005. 4353–4360