Editorial
Special Issue on Deep Reinforcement Learning and Adaptive Dynamic Programming

IN THE first issue of *Nature* 2015, Google DeepMind published a paper "Human-level control through deep reinforcement learning." Furthermore, in the first issue of *Nature* 2016, it published a cover paper "Mastering the game of Go with deep neural networks and tree search" and proposed the computer Go program, AlphaGo. In March 2016, AlphaGo beat the world's top Go player Lee Sedol by 4:1. This becomes a new milestone in artificial intelligence history, the core of which is the algorithm of deep reinforcement learning (RL).

Deep RL is able to output control signal directly based on input images, which incorporates both the advantages of the perception of deep learning (DL) and the decision making of RL or adaptive dynamic programming (ADP). This mechanism makes the artificial intelligence much closer to human thinking modes. Deep RL/ADP has achieved remarkable success in terms of theory and applications since it was proposed. Successful applications cover video games, Go, robotics, smart driving, healthcare, and so on.

However, it is still an open problem to perform the theoretical analysis on deep RL/ADP, e.g., the convergence, stability, and optimality analyses. The learning efficiency needs to be improved by proposing new algorithms or combined with other methods. More practical demonstrations are encouraged to be presented. Therefore, the aim of this special issue is to call for the most advanced research and state-of-the-art works in the field of deep RL/ADP.

We received a large number of submissions from different countries. Finally, 16 papers were accepted for publication in this special issue which can be roughly categorized into two groups. The first group contains one review papers and other nine contributions focusing on the theoretical and implementation issues of RL/ADP. The second group, also starting with a review paper, comprises six papers, which are aimed to combine RL with DL to form new algorithms. The 16 contributions are briefly described as follows.

B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis present a comprehensive review on state-of-the-art RL-based feedback control solutions to optimal regulation and tracking of single and multi-agent systems. Both optimal H_2 and H_∞ control problems, as well as graphical games using existing RL solutions, will be reviewed. The RL-based solution to optimal control and game problems is learned using online measured data along the system trajectories. For discrete-time (DT) and continuous-time (CT) systems, they discuss Q-learning and integral RL algorithms as core algorithms, respectively. Furthermore, a new direction of off-policy RL for both CT and DT systems is pointed out. Finally, several applications are presented and discussed.

RL in environments with many action-state pairs is challenging. I. J. Sledge, M. S. Emigh, and J. C. Prıncipe propose an uncertainty-based information-theoretic approach for performing guided stochastic searches. They present the value of information as a criterion for the optimal tradeoff between expected costs and the granularity of the search process. This criterion is further augmented with a state-transition uncertainty factor, to guide the search process into previously unexplored regions of the policy space. The performance of the proposed uncertainty-based value-of-information policy evaluated on the games Centipede and Crossy Road is better in fewer episodes than stochastic-based exploration strategies.

For data-based-constrained optimal control problems in the case of nonaffine nonlinear DT systems, B. Luo, D. Liu, and H. N. Wu develop an adaptive optimal control approach. The constrained optimal control problem is first transformed to an unconstrained optimal control problem. Then, the value iteration-based Q-learning (VIQL) algorithm is proposed to learn the optimal Q-function. With the help of a handy initial condition, the convergence of the VIQL algorithm is guaranteed. For easy implementation, the critic-only structure which requires only one neural network (NN) to approximate the Q-function is developed. Finally, three simulation examples verify the effectiveness of the proposed solution.

For constrained-input nonlinear systems with matched and unmatched disturbances, H. Zhang, Q. Qu, G. Xiao, and Y. Cui propose a novel sliding mode control based on the ADP theory to guarantee the optimal cost. The optimal sliding mode control problem can be thought of as the optimal control problem of a reformulated auxiliary system with a modified cost function, if the system moves on the sliding surface. They adopt the ADP algorithm based on single critic NN to solve the approximate optimal control law for the auxiliary system. The convergence of the NN weight errors with uniform ultimate boundness is proved with Lyapunov techniques. Moreover, the proposed approximate optimal control also guarantees the stability of the sliding mode dynamics with uniform ultimate boundedness. The feasibility of the proposed control scheme is finally demonstrated by some experimental results.

For a class of output-constrained CT unknown nonlinear systems, B. Fan, Q. Yang, X. Tang, and Y. Sun present a novel optimal control strategy based on robust ADP (RADP). An equivalent nonlinear system is first generated with an

Digital Object Identifier 10.1109/TNNLS.2018.2818878

2162-237X © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

error transformation technique, with the same asymptotic stability and the satisfaction of the output restriction as the original system. Then, they propose RADP algorithms to solve the transformed nonlinear optimal control problem with completely unknown dynamics to guarantee the closed-loop systems stable. Theoretical analysis of the asymptotic stability of the original and transformed nonlinear systems is given. The merit of the proposed control policy is that the output of the plant is always within user-defined bounds, compared to usual optimal control results, which is finally verified by suitable computer simulations.

For the distributed output synchronization problem of multiagent systems with an active leader, Y. Yang, H. Modares, D. C. Wunsch II, and Y. Yin present optimal control protocols. The leader can act independently with a distributed observer designed for the leader's state estimation, which also generates the reference signal for each follower. Then, they formulate the output synchronization problem of the leader–follower system as a distributed optimal tracking problem. The inhomogeneous algebraic Riccati equations are solved online with an off-policy RL algorithm without requiring system dynamics. The effectiveness of the proposed algorithm is verified with a simulation example. The results show that both the steady-state error and the transient response of the agents can be minimized effectively.

For an infinite-horizon optimal regulation problem of an affine deterministic system, P. Deptula, J. A. Rosenfeld, R. Kamalapurkar, and W. E. Dixon present a new online method to approximate the value function of RL. To approximate the value function, the traditional regional model-based RL (R-MBRL) method is suitable over a large compact set, while the local state following (StaF) kernel approach is valid in a local neighborhood within a compact set. They adopt a state-dependent convex combination of the StaF-based and the R-MBRL-based method for approximation. The value function approximation is switched from the StaF approach to the R-MBRL approach, if the state comes into a neighborhood including the origin. A Lyapunov-based analysis is used to proof the semiglobally uniformly ultimately bounded convergence of the system. To show the scalability and performance of the developed method, multiple-state dynamical systems are designed in simulation results.

For switched systems with autonomous subsystems and CT dynamics, T. Sardarmehni and A. Heydari present two optimal control solutions. The first solution adopts recursive least squares to formulate a policy iteration algorithm. While the traditional policy iteration algorithm suffers from the computational burden, they further present a second solution with a single-loop policy iteration. In the second algorithm, the Lyapunov equation is not derived to generate the evolving policies which are not necessarily stabilizing so supervisory stabilizing policy for online training is required. For the implementation of each solution, both online and concurrent training algorithms are designed. Finally, some computer simulations have been performed to evaluate the effectiveness of the presented algorithms.

For a class of unknown nonlinear DT systems with actuator fault, Z. Wang, L. Liu, Y. Wu, and H. Zhang present an optimal fault tolerant control (FTC) based on adaptive critic design (ACD). It is a typical causal problem, so they first transform the original unknown nonlinear system into a new system using the diffeomorphism theory. To approximate a predefined unknown function in the backstepping design procedure, the action NN is adopted. Then, a RL algorithm is proposed to achieve an optimal FTC, by integrating the strategic utility function and the ACD technique, where the cost function is approximated by the critic NN. The stability of the systems and optimal control performance can be obtained. The effectiveness of the proposed optimal FTC strategy is finally verified with two simulation examples.

For the efficient distributed economic dispatch problem of random nonlinear distributed generation (DG) units and loads in microgrids, W. Liu, P. Zhuang, H. Liang, J. Peng, and Z. Huang propose a cooperative RL algorithm. Learning algorithms can help to solve the problem of stochastic modeling difficulty and high computational complexity. The function approximation is suitable for the large and continuous state spaces in the cooperative RL algorithm. They incorporate a diffusion strategy for the cooperation of the actions of DG units and energy storage devices. The proposed algorithm enables each node in microgrids to communicate only with its local neighbors, so that any centralized controllers are not necessary. The convergence of the proposed algorithm is analyzed. The performance is further validated with simulations on real-world meteorological and load data.

The second group starts with a comprehensive survey paper by M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli on the application of DL, RL, and deep RL techniques, especially with biological data. Recent achievements of artificial intelligence are greatly promoted by the development on DL, RL, and their combination (deep RL) and also benefit from the highly efficient computational power and increased big data. Moreover, the performances of DL techniques are further compared in detail with different data sets for various application domains. In the end, open issues in this challenging research area are pointed out and future development perspectives are discussed.

To accomplish a range of tasks with RL, a meta-policy can be learned over a set of training tasks from the same distribution. How to identify unrelated or even opposite tasks and how to relate meta-policy to task features are two major obstacles. Y. Yu, S. Y. Chen, Q. Da, and Z. H. Zhou propose a MAPLE approach by introducing the shallow trail to overcome the two difficulties. MAPLE not only groups similar tasks with the rewards of the shallow trail, but also runs a roughly trained policy to probe a task. Furthermore, the rewards of the shallow trail can be directly served as task features, if the task parameters are unknown. MAPLE is verified with empirical studies on several controlling tasks to obtain better meta-policies with a higher reward.

For training deep RL, Z. Ren, D. Dong, H. Li, C. Chen propose a new paradigm using self-paced prioritized curriculum learning with coverage penalty, called deep curriculum RL (DCRL). DCRL selects appropriate transitions from replay memory adaptively to take the most advantage of experience replay. The self-paced priority and coverage penalty

are derived as the criteria of complexity in DCRL. Atari 2600 games are used for comparison, and the experimental results show that DCRL outperforms deep Q-network (DQN) and prioritized experience replay methods on most of these games. They also present promising results for the proposed curriculum training paradigm to be applicable and effective to other memory-based deep RL approaches, such as double DQN and dueling network. In conclusion, the training efficiency and robustness for deep RL can be improved with DCRL.

To solve the problem of action overestimation in DQN, J. Pan, X. Wang, Y. Cheng, and Q. Yu propose a multisource transfer double DQN (MTDDQN) to eliminate the error accumulation. The advantage of integrating the transfer learning technique with deep RL facilitates the RL agent to collect, summarize, and transfer action knowledge. Moreover, a multisource transfer learning mechanism is established for ensuring strong correlations between source and target tasks to avoid negative knowledge transfer. The feasibility and performance of MTDDQN are evaluated on the Atari2600 games. Compared to mainstream approaches, such as DQN and double DQN, experimental results show that MTDDQN achieves not only better learning transfer capability but also higher learning efficiency and testing accuracy.

S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi propose a deep NN-based visual tracker to directly capture the target object in a video with a bounding box. Various training video sequences are used to pretrain the proposed deep NN-based visual tracker, which is further fine-tuned with online adaptation. Deep RL and supervised learning are used for pretraining, while RL is used for semisupervised learning with even partially labeled data. The proposed tracker is validated against the object tracking benchmark dataset, to show a highly competitive performance in speed than existing deep network-based trackers. The fast version of the proposed method, operating on GPU in real time, outperforms state-of-the-art real-time trackers showing a relevant tracking accuracy improvement.

Choosing a sequence of actions for an active camera will help to discriminate between the objects; in this brief, H. Liu, Y. Wu, and F. Sun develop a new deep RL method to actively recognize objects. The method is realized using trust region policy optimization. The policy is learned with an extreme learning machine to make the optimization algorithm efficient. The advantages of the developed extreme trust region optimization method are finally verified with experimental results on a publicly available data set.

The above briefly introduces 16 papers in this special issue. Thanks again to the authors for sharing their valuable work with us. Last but not least, we want to express our sincere gratitude to IEEE TNNLS EiC Prof. H. He and editorial staff Prof. J. Yan for their patience and precious support!

DONGBIN ZHAO, *Guest Editor*

DERONG LIU, *Guest Editor*

F. L. LEWIS, *Guest Editor*

JOSE C. PRINCIPE, *Guest Editor*

STEFANO SQUARTINI, *Guest Editor*

Dongbin Zhao (M'06–SM'10) has been a Professor with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, since 2002, and with the University of Chinese Academy of Sciences, Beijing. From 2007 to 2008, he was a Visiting Scholar at the University of Arizona, Tucson, AZ, USA. He has authored or co-authored four books and over 60 international journal papers. His current research interests include deep reinforcement learning, computational intelligence, adaptive dynamic programming, autonomous driving, robotics, intelligent transportation systems, and smart grids.

Dr. Zhao has served as an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS since 2012 and IEEE COMPUTATION INTELLIGENCE MAGAZINE since 2014. He was the Chair of the Adaptive Dynamic Programming and Reinforcement Learning Technical Committee and Multimedia Subcommittee of the IEEE Computational Intelligence Society from 2015 to 2016. He is the Chair of Beijing Chapter. He works as a Guest Editor of several renowned international journals. He is involved in organizing many international conferences.

Derong Liu (S'91–M'94–SM'96–F'05) received the Ph.D. degree in electrical engineering from the University of Notre Dame, Notre Dame, IN, USA, in 1994.

He became a Full Professor of electrical and computer engineering and of computer science at the University of Illinois at Chicago, Chicago, IL, USA, in 2006. He has authored or co-authored 19 books.

Dr. Liu is a fellow of the International Neural Network Society and the International Association of Pattern Recognition. He was selected for the "100 Talents Program" by the Chinese Academy of Sciences in 2008, and he served as the Associate Director of The State Key Laboratory of Management and Control for Complex Systems at the Institute of Automation, from 2010 to 2015. He is the Editor-in-Chief of the *Artificial Intelligence Review* (Springer). He was the Editor-in-Chief of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS from 2010 to 2105.

F. L. Lewis received the Ph.D. degree from the Georgia Institute of Technology, Atlanta, GA, USA.

He is currently a Distinguished Visiting Professor with the Nanjing University of Science and Technology, Nanjing, China, and a Project 111 Professor with Northeastern University, Shenyang, China. He is also a Qian Ren Thousand Talents Consulting Professor with Northeastern University, Shenyang, China. He is involved in feedback control, intelligent systems, cooperative control systems, and nonlinear systems. He has authored numerous journal special issues, 320 journal papers, and 20 books, including *Optimal Control, Aircraft Control, Optimal Estimation*, and *Robot Manipulator Control* which are used as university textbooks worldwide. He holds six U.S. patents.

Dr. Lewis is a member of the National Academy of Inventors, a fellow of IFAC and the U.K. Institute of Measurement and Control, PE Texas, a U.K. Chartered Engineer, and the Moncrief-O'Donnell Chair at the University of Texas at Arlington Research Institute. He was a recipient of the Fulbright Research Award, the NSF Research Initiation Grant, the ASEE Terman Award, the International Neural Network Society Gabor Award, the U.K. Institute of Measurement and Control Honeywell Field Engineering Medal, and the IEEE Computational Intelligence Society Neural Networks Pioneer Award. He received the Outstanding Service Award from the Dallas IEEE Section and the Texas Regents Outstanding Teaching Award 2013. He is a Founding Member of the Board of Governors of the Mediterranean Control Association. He was selected Engineer of the Year by the Ft. Worth IEEE Section.

Jose C. Principe (M'83–SM'90–F'00) is currently a Distinguished Professor of electrical and computer engineering and biomedical engineering with the University of Florida, Gainesville, FL, USA, where he teaches advanced signal processing, machine learning, and artificial neural networks modeling. He is the Eckis Professor of ECE and the Founder and the Director with the Computational NeuroEngineering Laboratory (CNEL), University of Florida, Gainesville, FL, USA. The CNEL Lab has been studying signal and pattern recognition principles based on information theoretic criteria (entropy and mutual information). He has authored or co-authored over 800 publications. He directed 95 Ph.D. dissertations and 65 master's theses. He wrote in 2000 an interactive electronic book entitled *Neural and Adaptive Systems* (John Wiley and Sons) and more recently co-authored several books on *Brain Machine Interface Engineering* (Morgan and Claypool), *Information Theoretic Learning* (Springer), and *Kernel Adaptive Filtering* (Wiley). His current research interests include processing of time varying signals with adaptive neural models.

Dr. Principe is a member of the Advisory Board of the University of Florida Brain Institute and the Advisory Board of the University of Florida Brain Institute. He was the Chair of the Technical Committee on Neural Networks of the IEEE Signal Processing Society, the President of the International Neural Network Society, and the Editor-in-Chief of the IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING.

Stefano Squartini (SM'12) was born in Ancona, Italy, in 1976. He received the Italian Laurea degree (Hons.) in electronic engineering and the Ph.D. degree from the Polytechnic University of Marche (UnivPM), Ancona, Italy, in 2002 and 2005, respectively.

He joined the Department of Information Engineering as an Assistant Professor of electrical circuit theory with UnivPM, where he was a Post-Doctoral Researcher from 2006 to 2007. He has been an Associate Professor with UnivPM since 2014. He has authored or co-authored over 170 scientific peer-reviewed articles. His current research interests include computational intelligence and digital signal processing, with a special focus on speech/audio/music processing and energy management.

Dr. Squartini is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, and the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE. He joined the Organizing and the Technical Program Committees of more than 70 international conferences and Workshops in the recent past.