Machine Learning (2021) 110:2291–2293
https://doi.org/10.1007/s10994-021-06041-3

# Guest editorial: special issue on reinforcement learning for real life

**Yuxi Li[1] · Alborz Geramifard[2] · Lihong Li[3] · Csaba Szepesvari[4] · Tao Wang[5]**

Published online: 2 August 2021


Reinforcement learning (RL) is a general paradigm for learning, predicting, and decision making, with broad applications in sciences, engineering and arts. RL has seen prominent successes in many problems, such as Atari games, AlphaGo, robotics, recommender systems, and AutoML. However, applying RL in the real world remains challenging. A natural question arises: What are the challenges and how to address them?

The main goals of the special issue are to: (1) identify key research problems that are critical for the success of real-world applications; (2) report progress on addressing these critical issues; and (3) have practitioners share their successful stories of applying RL to real-world problems, and the insights gained from the applications.

We received 60 submissions, following an open call for papers successfully applying RL algorithms to real-life problems and/or addressing practically relevant RL issues, with respect to practical RL algorithms, practical issues, and applications. After a rigorous reviewing process, we accepted 11 articles, each of which was assessed by at least three reviewers, with one, mostly two, or three rounds of revisions.

In the article titled "Inverse Reinforcement Learning in Contextual MDPs", the authors Stav Belogolovsky, Philip Korsunsky, Shie Mannor, Chen Tessler, and Tom Zahavy formulate the contextual inverse RL problem as a non-differential convex optimization problem

✉ Yuxi Li
yuxili@gmail.com

Alborz Geramifard
alborz.geramifard@gmail.com

Lihong Li
lihongli.cs@gmail.com

Csaba Szepesvari
csaba.szepesvari@gmail.com

Tao Wang
taowangml@gmail.com

[1] Attain.ai, Edmonton T5R 4R4, Canada

[2] Facebook AI, Menlo Park 94025, USA

[3] Amazon, Seattle 98121, USA

[4] DeepMind & University of Alberta, Edmonton T6G 2E8, Canada

[5] Apple, Cupertino 95014, USA

and adapt descent methods to compute its subgradients, with both theoretical and empirical analysis, in particular, for zero-shot transfer to unseen contexts, and with applications to sepsis treatment.

The authors Ioannis Boukas, Damien Ernst, Thibaut Théate, Adrien Bolland, Alexandre Huynen, Martin Buchwald, Christelle Wynants, and Bertrand Cornélusse in the article titled "A Deep Reinforcement Learning Framework for Continuous Intraday Market Bidding" study the strategic participation of energy storage to maximize profits in a continuous intraday market with a centralized order book, considering operational constraints, using an asynchronous fitted Q-iteration algorithm for improved sample efficiency.

In the article titled "Bandit Algorithms to Personalize Educational Chatbots", the authors William Cai, Josh Grossman, Zhiyuan Jerry Lin, Hao Sheng, Johnny Tian-Zheng Wei, Joseph Jay Williams, and Sharad Goe develop a rule-based chatbot to explain math concepts, provide practice questions, and offer tailored feedback. They show that a contextual bandit approach outperforms A/B testing with lower cost for personalization in a live deployment of an educational conversational agent.

The authors Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester identify nine challenges for RL to be deployed in real-life scenarios, formulating each challenge in MDP, analyzing its effects on recent RL algorithms, presenting existing efforts to address it, and open source a suite of continuous control environments, in the article titled "Challenges of Real-World Reinforcement Learning: Definitions, Benchmarks & Analysis".

The authors Josiah P. Hanna, Siddharth Desai, Haresh Karnan, Garrett Warnell, and Peter Stone in the article titled "Grounded Action Transformation for Sim-to-Real Reinforcement Learning" study how to transfer policies learned in simulation to real-life scenarios with a humanoid robot, in particular, with stochastic state transitions, and demonstrate performance gains in controlled experiments and with a real robot.

In the article titled "Air Learning: A Deep Reinforcement Learning Gym for Autonomous Aerial Robot Visual Navigation", the authors Srivatsan Krishnan, Behzad Boroujerdian, William Fu, Aleksandra Faust, and Vijay Janapa Reddi introduce an open source simulator for resource-constrained aerial robots, and deploy domain randomization and hardware-in-the-loop techniques for an Unmanned Aerial Vehicles (UAVs) agent on embedded platforms, considering quality-of-flight (QoF) metrics, such as the energy consumed, endurance, and the average trajectory length, and mitigating the hardware gap w.r.t. the discrepancy in the flight time using artificial delays in training.

In the article titled "Dealing with Multiple Experts and Non-Stationarity in Inverse Reinforcement Learning: An Application to Real-Life Problems", the authors Amarildo Likmeta, Alberto Maria Metelli, Giorgia Ramponi, Andrea Tirinzoni, Matteo Giuliani, and Marcello Restelli introduce batch model-free inverse RL approaches to handling data from multiple experts and non-stationarity in reward functions, and evaluate the performance with scenarios of highway driving, user preference inference in social networks, and water management.

In the article titled "Lessons on Off-policy Methods from a Notification Component of a Chatbot", the authors Scott Rome, Tianwen Chen, Michael Kreisel, and Ding Zhou present their experience applying off-policy techniques to train and evaluate a contextual bandit model for troubleshooting notification in a chatbot, considering a null action, a limited number of bandit arms, small data, reward design, logging policy, and model post-training.

The authors Wenjie Shang, Qingyang Li, Zhiwei Qin, Yang Yu, Yiping Meng, and Jieping Ye in the article titled "Partially Observable Environment Estimation with Uplift Inference for Reinforcement Learning based Recommendation" propose to estimate a

partially observable environment from past data by treating hidden variables as a hidden policy, with a deep uplift inference network model, and conduct experiments in both simulated and real-life environments, in particular, for a program recommender system on a large-scale riding-hailing platform.

In the article titled "Automatic Discovery of Interpretable Planning Strategies", the authors Julian Skirzyński, Frederic Becker, and Falk Lieder propose to transform RL policies into interpretable descriptions with imitation learning, program induction, and clustering, and evaluate the performance with behavioral experiments.

The article by authors Sabina Tomkins, Peng Liao, Predrag Klasnja, and Susan Murphy on "IntelligentPooling: Practical Thompson Sampling for mHealth" undertakes a study to address the challenges of learning personalized user policies from limited data and non-stationary responses to treatments, with a high probability regret bound, an empirical evaluation, and a pilot study in a live clinical trial.

Collectively, these 11 articles illustrate the diverse range of issues currently being investigated in the field of reinforcement learning for real life.