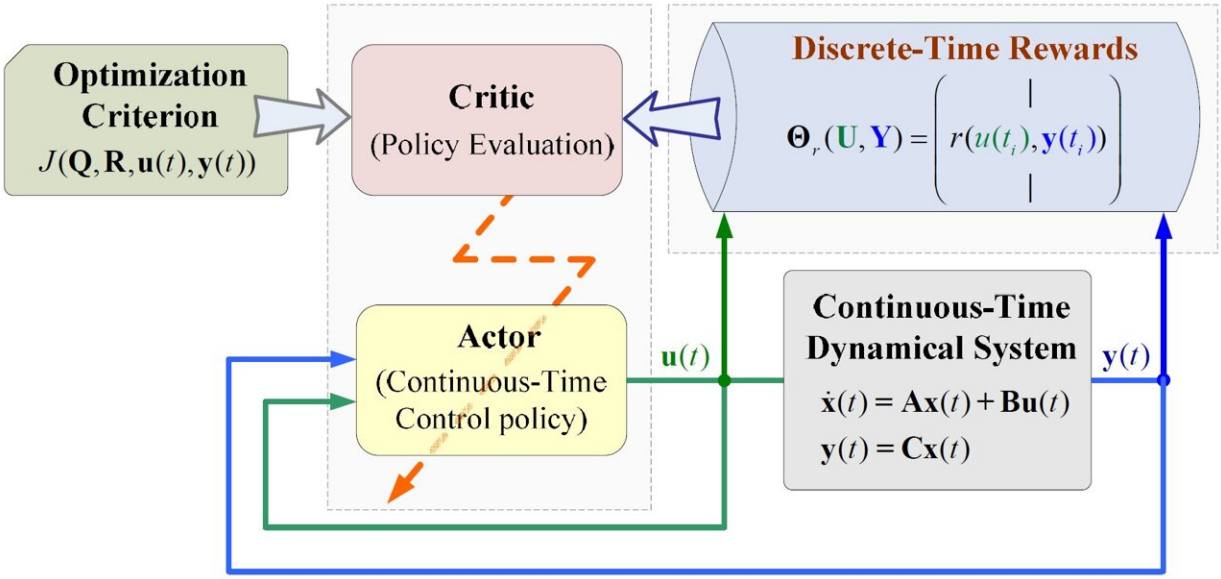
Discrete-time rewards efficiently guide the extraction of continuous-time optimal control policy from system data

June 28 2024



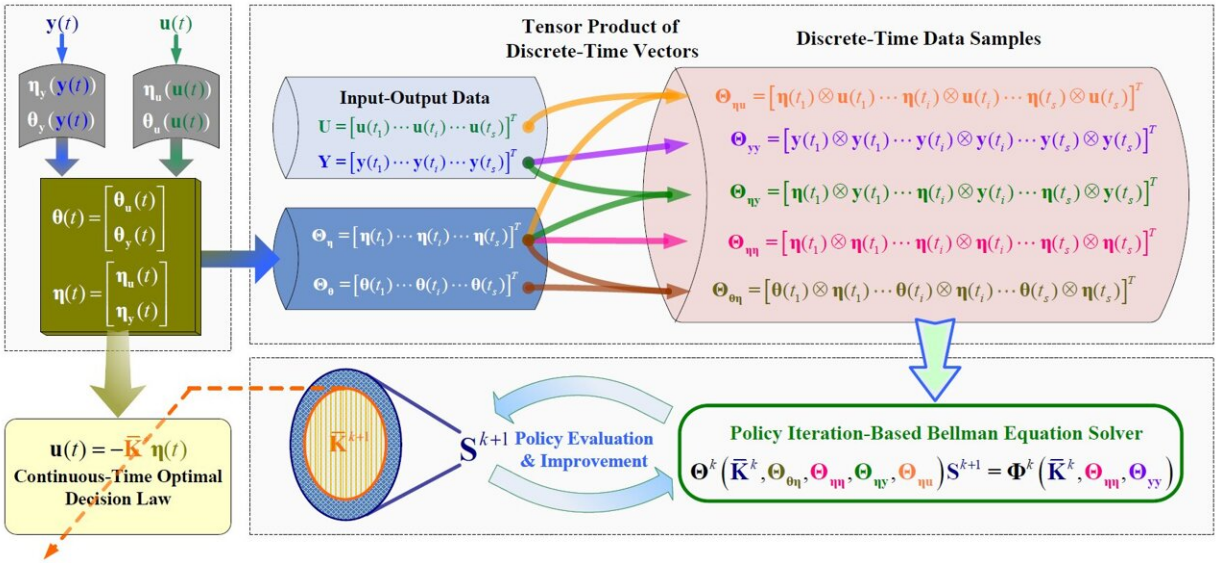
At each sampling time instant, one observes system output and action to form discrete-time rewards. The sampled input-output data are collected along the trajectory of the dynamical system in real-time, and are stacked over the time interval of interest as the discrete-time input-output data. The input-output data, associated with the prescribed optimization criterion, are used for updating the value estimate given in the critic module, based on which the control policy in the actor module is updated. The ultimate goal of this framework is to use the input-output data for learning the optimal decision law that minimizes the user-defined optimization criterion. Credit: Science China Press

The concept of reward is central in reinforcement learning and is also widely used in the natural sciences, engineering, and social sciences. Organisms learn behavior by interacting with their environment and observing the resulting rewarding stimuli. The expression of rewards largely represents the perception of the system and defines the behavioral state of the dynamic system. In reinforcement learning, finding rewards that explain behavioral decisions of dynamic systems has been an open challenge.

The work aims to propose reinforcement learning algorithms using discrete-time rewards in both continuous time and action space, where the continuous space corresponds to the phenomena or behaviors of a system described by the laws of physics. The approach of feeding state derivatives back into the learning process has led to the development of an analytical framework for reinforcement learning based on discrete-time rewards, which is essentially different from existing integral reinforcement learning frameworks.

"When the idea of feedbacking the derivative into the learning process struck, it felt like lightning. And guess what? It mathematically ties into the discrete-time reward-based policy learning," says Dr. Ci Chen, recalling his Eureka moment.

Under the guidance of discrete-time reward, the search process of behavioral decision law is divided into two stages: feed-forward signal learning and feedback gain learning. In their study, it was found that the optimal decision law for continuous-time dynamic systems can be searched from real-time data of dynamic systems using the discrete-time reward-based technique.



The central is to leverage the sampled data to extract laws from data. To this end, pre-process the actions and outputs of the dynamical system and construct the feedforward signals that will be used for the feedback gain learning and the design of an online real-time control loop. Then, measure the input-output data, as well as the feedforward signals, over discrete-time series, based on which the discrete-time data samples are assembled using the tensor product. Calculate the Bellman equation for optimality via policy iterations. Through policy evaluation and improvement, the optimal feedback gain is obtained from the discrete-time data samples with rigorous mathematical operations and convergence deduction. Finally, both the feedforward signal and the feedback gain contribute to the optimal decision law. Credit: Science China Press

The above method has been applied to power system state regulation to achieve optimal design of output feedback. This process eliminates the intermediate stage of identifying dynamic models and significantly improves the computational efficiency by removing the reward integrator operator from the existing integral reinforcement learning framework.

This research uses discrete-time reward guidance to discover optimization strategies for continuous-time dynamical systems, and constructs a computational tool for understanding and improving dynamical systems. This result can play an important role in natural science, engineering, and social science.

The work is published in the journal *National Science Open*.

This study is led by an international team of scientists including Dr. Chen (School of Automation, Guangdong University of Technology, China), Dr. Lihua Xie (School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore), and Dr. Shengli Xie (Guangdong-HongKong-Macao Joint Laboratory for Smart Discrete Manufacturing, Guangdong Key Laboratory of IoT Information Technology, China), co-contributed by Dr. Yilu Liu (Department of Electrical Engineering and Computer Science, University of Tennessee, U.S.) and Dr. Frank L. Lewis (UTA Research Institute, The University of Texas at Arlington, U.S.).

More information: Ci Chen et al, Learning the Continuous-Time Optimal Decision Law from Discrete-Time Rewards, *National Science Open* (2024). DOI: 10.1360/nso/20230054

Provided by Science China Press