Huan HU, Qing-ling WANG, 2020. Proximal policy optimization with an integral compensator for quadrotor control. *Frontiers of Information Technology & Electronic Engineering,* 21(5):777-795. https://doi.org/10.1631/FITEE.1900641

Proximal policy optimization with an integral compensator for quadrotor control

Key words: Reinforcement learning; Proximal policy optimization (PPO); Quadrotor control; Neural network

Corresponding author: Qing-ling WANG
E-mail: qlwang@seu.edu.cn
ORCID: https://orcid.org/0000-0003-2045-2920

Motivation

- The quadrotor is a highly nonlinear multi-input multi-output under-drive coupling system, which makes the control design very difficult and complicated. The model contains a lot of unmodeled dynamic and nonlinear external interferences. Designing a quadrotor controller with anti-jamming capability has gradually become a hot research topic.
- In view of the influence of many factors such as the complexity of the model, machine learning, especially model-free reinforcement learning (RL), can be used to design the controller, and reinforcement learning has been successfully applied in quadrotor flight control.

Main innovation

- We propose a proximal policy optimization with integral compensator (PPO-IC) algorithm by introducing a state integrator. The algorithm significantly improves the tracking accuracy in motion control.
- We adopt a two-stage learning mode to train the model. In the offline phase, we train a simplified model in the simulation to learn a controller with robustness. Then, we train a real quadrotor in the actual scene, and optimize the control strategies to build a high-performance flight controller in the online phase.

Quadrotor model

The quadrotor has four rotors. Each rotor is distributed at the end of the cruciform frame. The distance between each rotor and the center of mass is *L*.

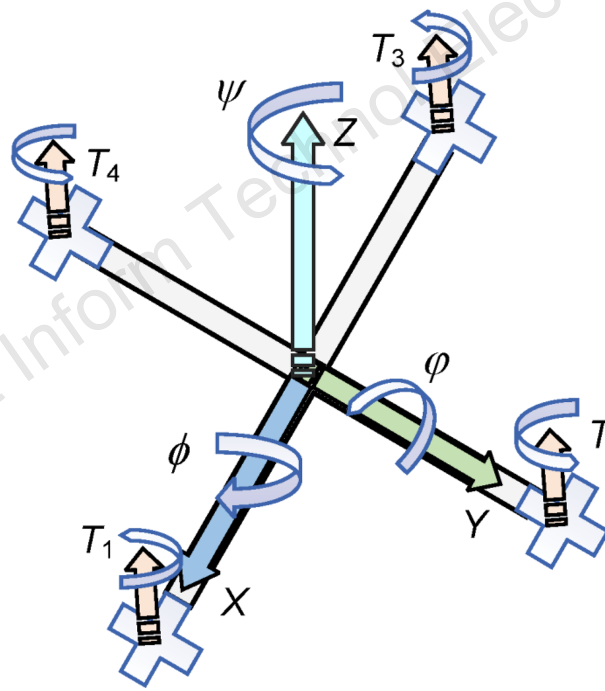

Fig. 1 The quadrotor model and body-fixed frame (Rozi et al., 2017)

Reinforcement learning

Modeling the quadrotor flight control problem as a Markov decision process (MDP), the RL algorithm framework is as follows.

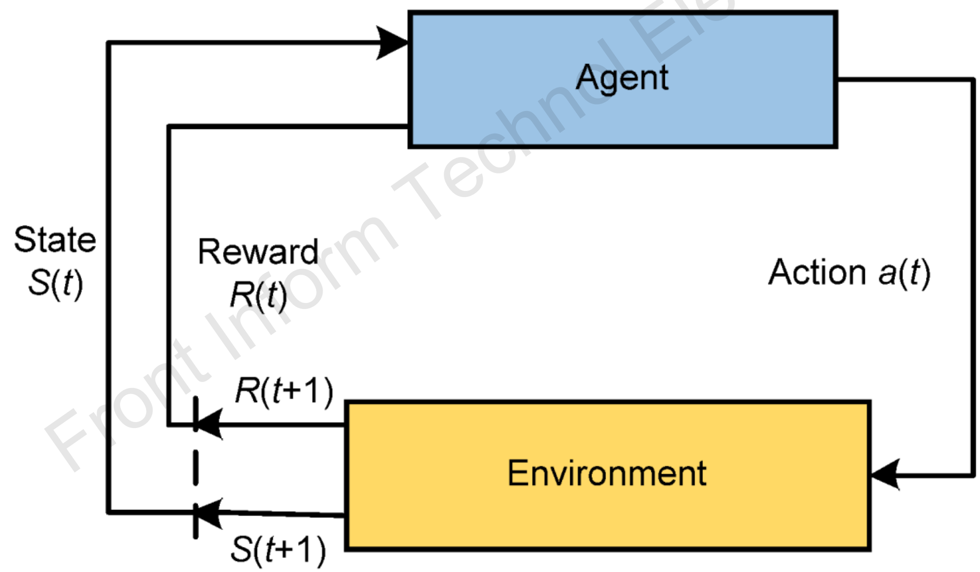

Fig. 2 The agent-environment interaction in a Markov decision process

PPO-IC structure

The actor network is composed of four sub-networks, which are used to select the actions to be performed, and the critic network is used to evaluate the quality of the actions.

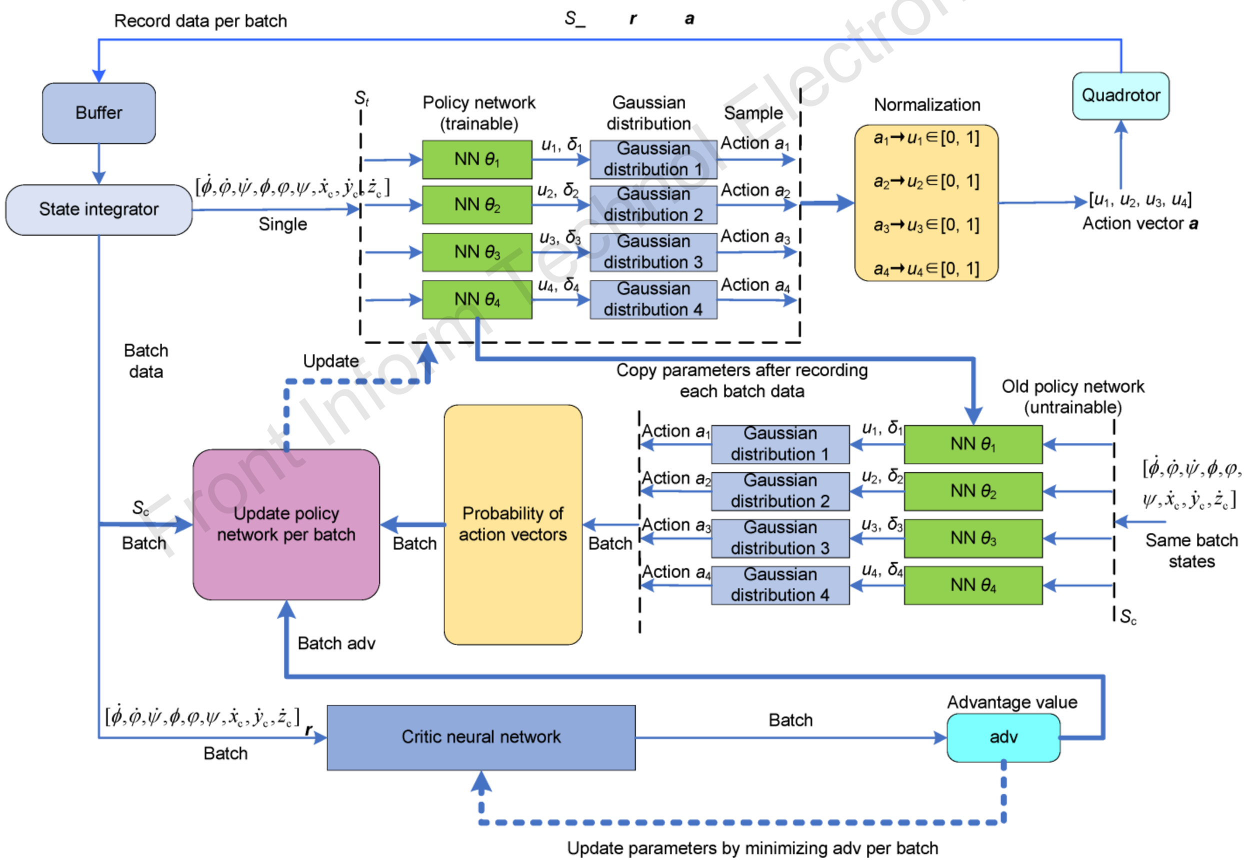

Fig. 3 System network framework structure

Actor-critic network

Policy sub-networks output two items: one is the mean μ of the Gaussian distribution, and the other is the standard deviation δ of the Gaussian distribution. The critic network outputs an advantage value, which is used to evaluate how well a given action is taken in a given state.

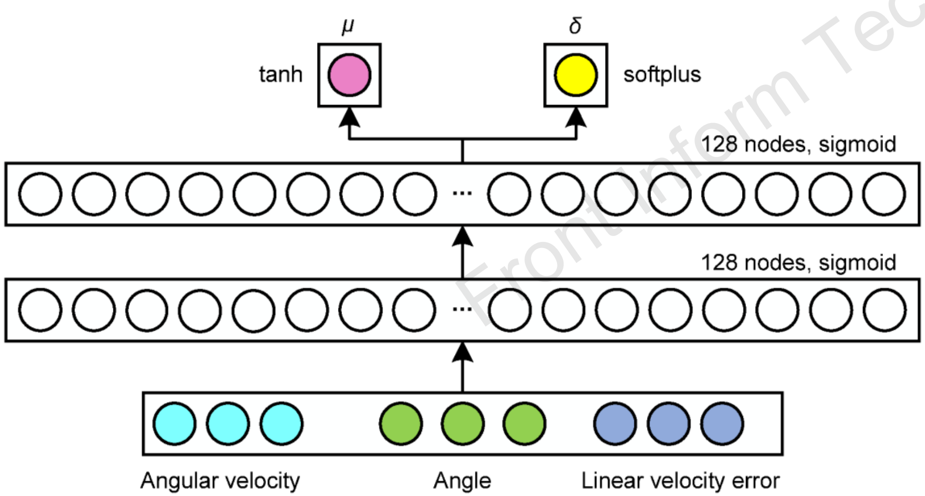

Fig. 4 The structure of four policy sub-networks used in this work

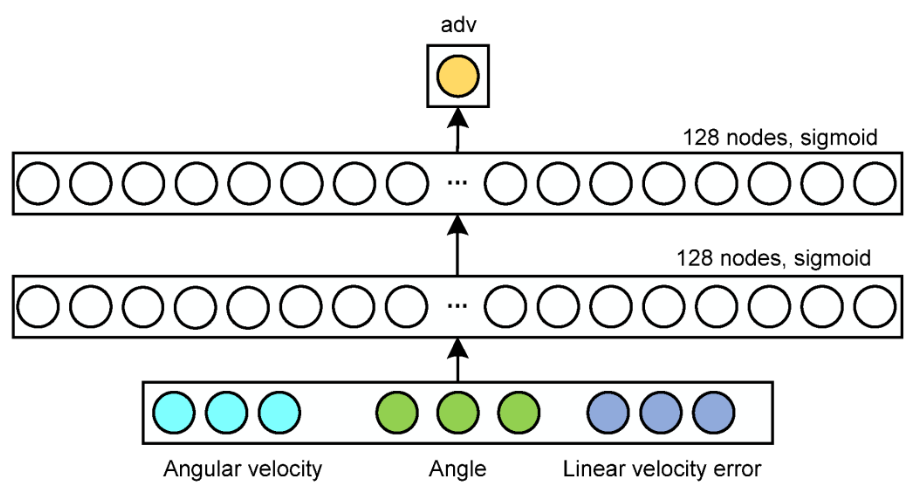

Fig. 5 The structure of the critic neural network used in this work

Major results

Comparison of the reward and velocity tracking curves of PPO and PPO-IC

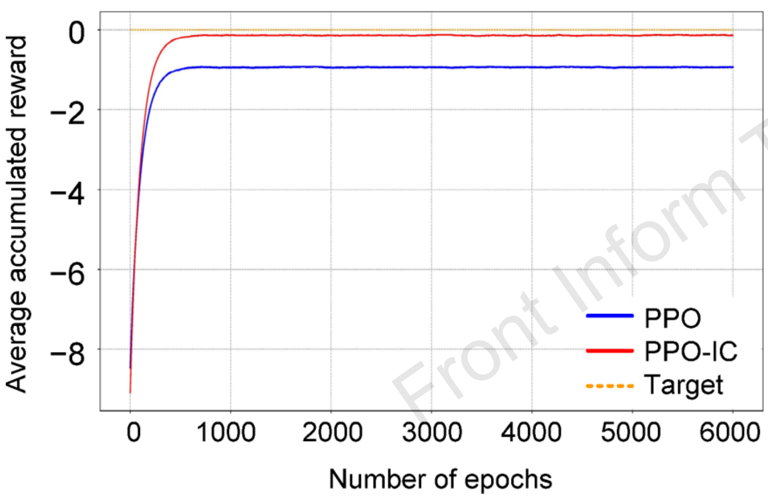

Fig. 7 Averaged accumulated reward in the evaluation of polices learned by PPO-IC and PPO

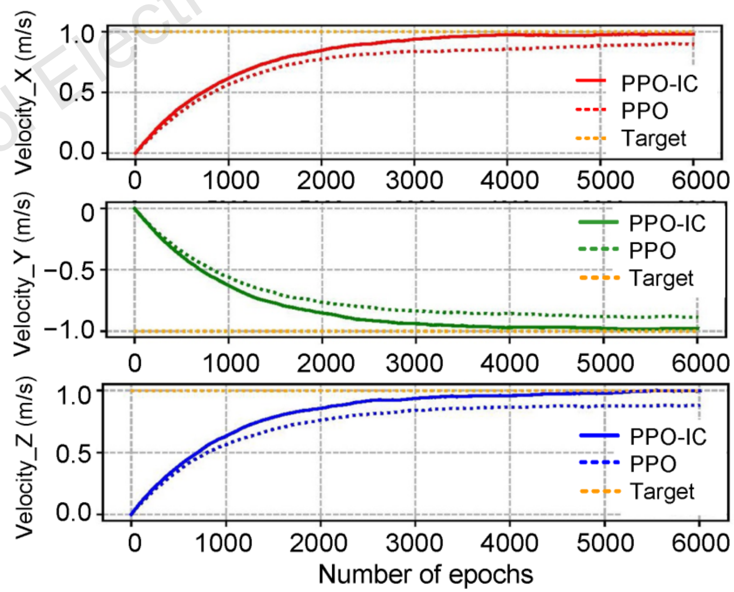

Fig. 8 Velocity tracking curves during the learning and training of PPO-IC and PPO

Major results (Cont'd)

Comparison of PID and PPO-IC

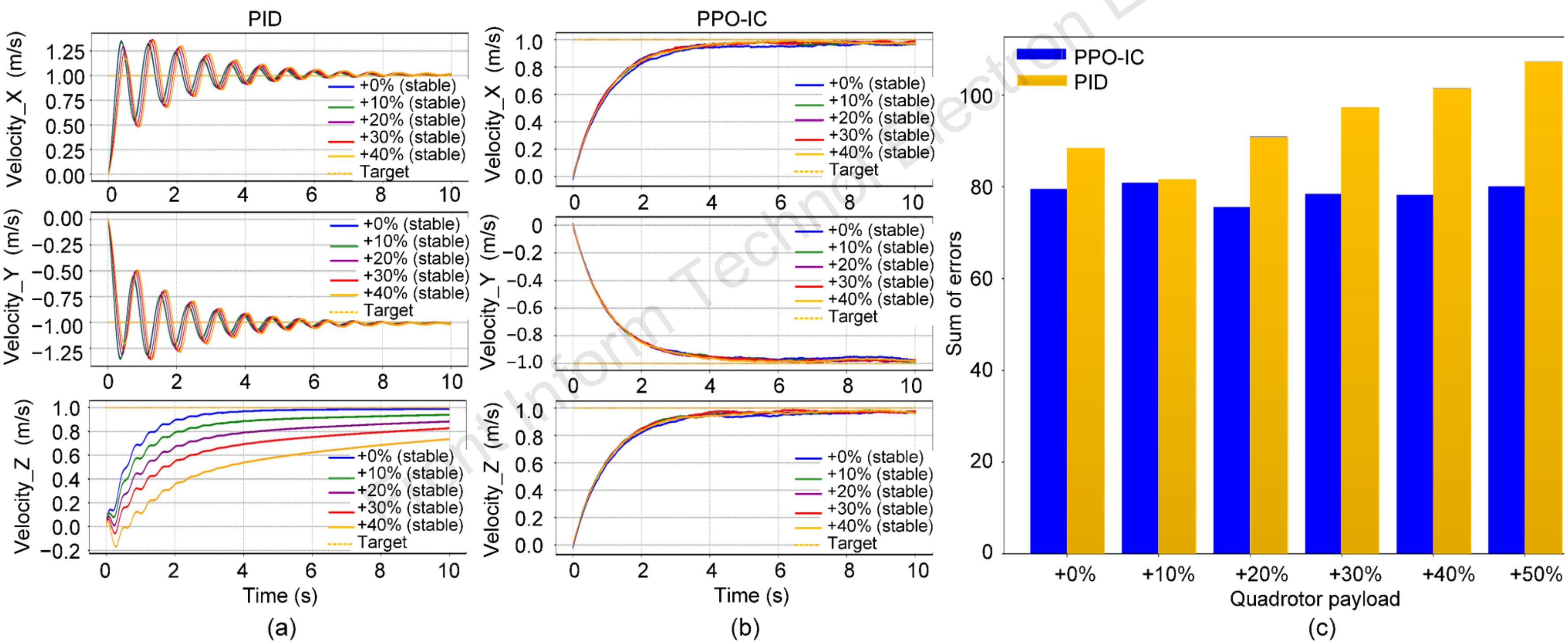

Fig. 14 Comparison of the control performance of the quadrotor model with the PPO-IC algorithm and the PID controller with different payloads: (a) three-dimensional velocity tracking response curve with the PID controller; (b) three-dimensional velocity tracking response curve with the PPO-IC algorithm; (c) sum of the errors under different payloads

Major results (Cont’d)

Comparison of PID and PPO-IC

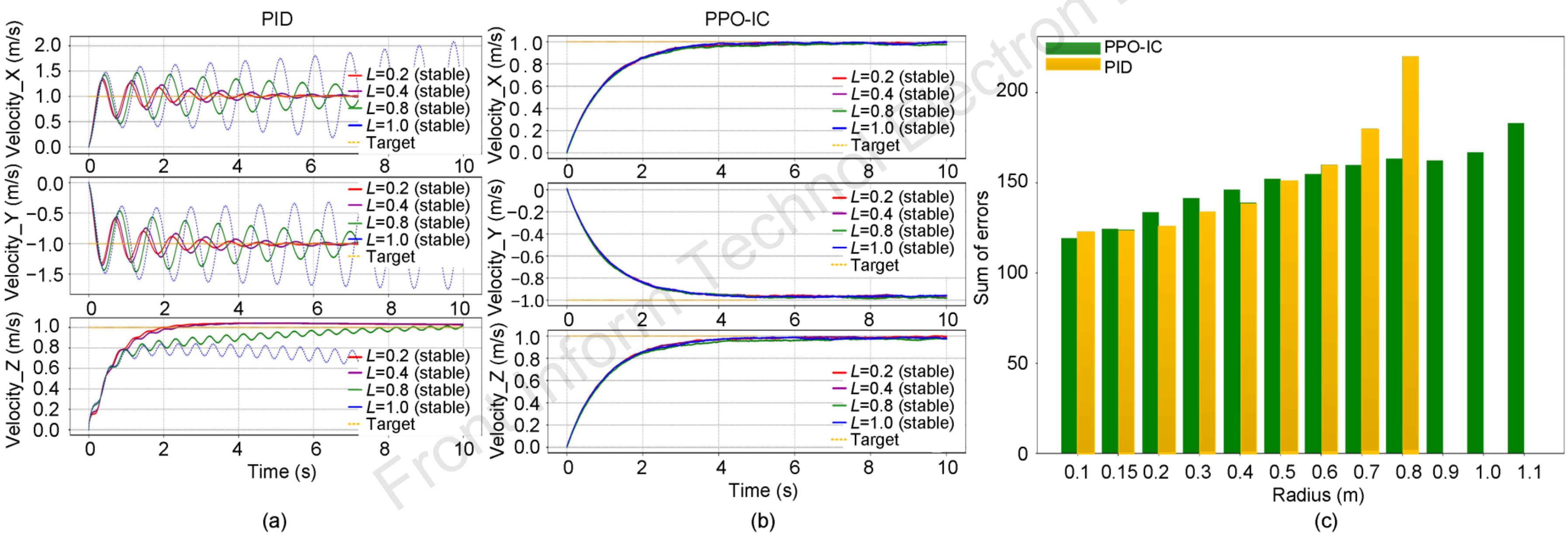

Fig. 15 Comparison of the control performance of the quadrotor model with the PPO-IC algorithm and the PID controller of different sizes: (a) three-dimensional velocity tracking response curve with the PID controller; (b) three-dimensional velocity tracking response curve with the PPO-IC algorithm; (c) sum of the errors of different sizes
If a model of a certain size fails to reach a steady state, the sum of errors is not shown in (c)

Conclusions

- We have proposed a PPO-IC algorithm with state integral for the development of the UAV intelligent controller, which successfully reduces the steady-state error in speed tracking and significantly improves the tracking accuracy.
- This method, together with the proposed reward signal, provides good sample efficiency and reduces the convergence time.
- A two-stage learning program has been proposed to develop a high-performance flight controller.