Deep Reinforcement Learning of PID Control for Rotational Wing Aerial Vehicles

Angelina Kim (angelina.kim.25@bishops.com)
Advisor: Dr. Marcus Jaiclin (marcus.jaiclin@bishops.com)
The Bishop's School, La Jolla, CA

Problem and Hypothesis

- **Problem**: RWAV PID controller setup is static, and it might not be able to maximize its capability to achieve flight goals under varying environments. Can the emerging DRL network manage PID controllers and improve flight performance?
- **Hypothesis**: RWAV can hold position better when PID controller coefficients are adjusted dynamically by DRL network after simulated training based on the RWAV flight telemetry data.

Background

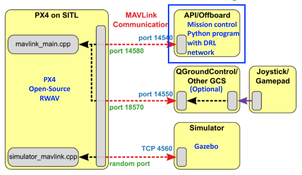
The automatic control of RWAV is essential for safety. The automation in flight is managed by nested Proportional, Integral, and Derivative (PID) controllers. RWAV has multiple levels of position, velocity, acceleration, attitude, and rate PID controllers. They are fixed to a conservative set of values with margin for stable flight. Under extreme conditions, such as strong wind, RWAV might not be able to maintain its flight and mission. If PID controllers are tuned against such conditions, it still might have the capability to overcome the conditions.

Deep Reinforced Learning (DRL) network has been recently introduced as an attractive machine-learning algorithm for complex problem solving (REINFORCE agent, 2023). It operates from observations to generate actions, then learn from rewards. Through iterative trainings, it has demonstrated performance beyond human intelligence.

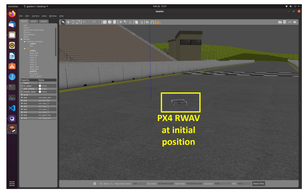
Using a Software-In-The-Loop (SITL) physics simulation of RWAV and DRL network, flight and mission controls can be integrated together. There is a potential for dynamic performance and safety improvement in flight control.

Setup and Configuration

PX4 RWAV simulation block diagram



Gazebo physics simulator initial status



PX4 autopilot is adopted as an open-source platform for ground, marine, and aerial vehicle development (PX4 Drone Autopilot, 2023). PX4 supports both SITL and Hardware In The Loop (HITL) simulations (PX4 Autopilot user guide, 2023).

Gazebo is an open-source 3D robotics simulator with a high-performance physics engine adopted by NASA and DARPA robotics challenges (Gazebo, 2023). Gazebo's wind plugin introduces random wind with average and variation values for velocity and direction (x, y, z). Only horizontal wind at random direction and velocity with varying velocities is used for the DRL network training and evaluation.

The Micro Aerial Vehicle Link (MAVLink) connects PX4, Gazebo simulator, QGroundControl, and mission control program through data networks. All components are implemented in a single computer and the MAVLink is established through a UDP loop back (MAVLink Developer Guide, 2023).

The mission control program is a custom Python code designed for this project to communicate with PX4 RWAV, bring the vehicle through the test procedure, and integrate the DRL network agent for training and evaluation.

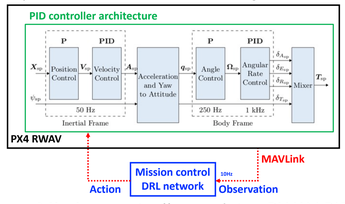
DRL Network for Dynamic PID Update

Mission controller Python program communicates with RWAV flight controller through MAVLink. Telemetry data is updated at 10Hz rate, and 50 data points are fed to the DRL network as an observation vector. The DRL network generates PID controller coefficients as an action vector. There are two cases: 1) Rate PID controllers only with 8 coefficients, and 2) all PID controllers with 19 coefficients updated by the DRL network.

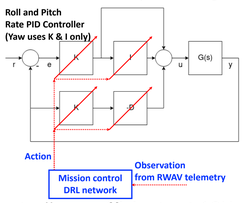
PID Controller Coefficients

#	Controller	PID parameter name
1	Rate 1000Hz	MC_PITCHRATE_K
2		MC_PITCHRATE_D
3		MC_PITCHRATE_I
4		MC_ROLLRATE_K
5		MC_ROLLRATE_D
6		MC_ROLLRATE_I
7		MC_YAWRATE_K
8		MC_YAWRATE_I
9	Attitude 250Hz	MC_ROLL_P
10		MC_PITCH_P
11		MC_YAW_P
12	Velocity 50Hz	MPC_XY_VEL_P_ACC
13		MPC_XY_VEL_I_ACC
14		MPC_XY_VEL_D_ACC
15		MPC_Z_VEL_P_ACC
16		MPC_Z_VEL_I_ACC
17		MPC_Z_VEL_D_ACC
18	Position 50Hz	MPC_XY_P
19		MPC_Z_P

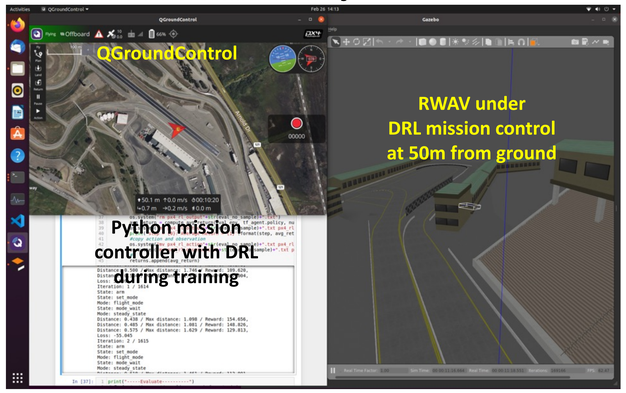
Dynamic PID Controller Block Diagram



Rate Controller PID Update Diagram



Mission controller with DRL network in training



Mission controller updates RWAV PID controller coefficients at 10Hz rate. During flight, RWAV's performance is rated by calculating distance from the target position, and it is provided to the DRL network as rewards during the DRL training and Evaluation. Using collected observation, action, and reward, the DRL is trained to improve PID controller coefficient set-up action strategy.

RWAV and Simulator Set-Up

Dynamic PID Controller Block Diagram

Telemetry data set requested for DRL

MAVLink message #	MAVLink message name	Description	Number of data
26	SCALED_IMU	X, Y, and Z angular acceleration and speed	6
30	ATTITUDE	Roll, pitch, and yaw angle and angular speed	6
31	ATTITUDE_QUATERNION	Quaternion component 1 -4, roll, pitch, and yaw angular speed	7
32	LOCAL_POSITION_NED	Local position and speed in X, Y, and Z with respect to North, East, and Down (NED)	6
36	SERVO_OUTPUT_RAW	Servo 1-4 output values	4
74	VFR_HUD	Air and ground speed, heading, throttle, altitude, and climb rate	4
83	ATTITUDE_TARGET	Current vehicle attitude target in attitude quaternion, body roll, pitch, and yaw rate, thrust	8
230	ESTIMATOR_STATUS	Output of EKF estimator. Velocity, horizontal, vertical innovation test ratio	4
340	UTM_GLOBAL_POSITION	X, Y, and Z velocity	3
12901	OPEN_DRONE_ID_LOCATION	Horizontal and vertical speed	2
		Total parameters	50

The mission controller Python program integrates PX4 RWAV SITL and Gazebo physics simulation to run experiments. The mission control initializes MAVLink connection and DRL network instance. RWAV is put at full command off-board mode for mission control through MAVLink. Actuators are armed to prepare take off. A total of 50 telemetry parameters are requested through MAVLink message at 10Hz update rate. Telemetry data set includes gyroscope, attitude, attitude target, estimator status, actuator output, position, and ground velocity

Target position command is sent to RWAV at 10Hz to meet 2Hz refresh rate requirement. Otherwise, RWAV changes mode to failsafe. The mission control waits until the vehicle is at target position at 50m above coordinate origin within 0.3m. 50m allows RWAV's altitude loss and recovery to avoid crash during simulation. Virtual test site at Sonoma Raceway with wind plugin emulates real flight environment.

DRL Network Set-Up and Procedure

Reinforcement learning agent is formed and initialized. The network has five inner layers with 50 – 400 nodes at each layer. DRL network observation is 50 telemetry data set obtained from PX4 at 10Hz. The DRL network uses REINFORCE TF agent and has continuous range mapped to PID coefficients (REINFORCE agent, 2023). DRL output actions are PID controller coefficients and are updated at 10Hz to PID controllers. RWAV is positioned at 50m above origin, and training starts when it is within 0.3m range from target.

- Reward function = 1 / [(1 + current distance)
  * (1 + accumulated average distance)
  * (1 + accumulated maximum distance)]
- Training samples: 600, collected for 1 minutes
- Evaluation samples: 1200, evaluated for 2 minutes

Learning rate is 0.0001, and total training iteration is 1500. 3 episodes are collected per iteration. Checkpoints are saved at every training, and a copy store at every 10 trainings. At every 50-training iterations, the DRL performance is assessed by collecting 30 times of 2-minute sessions to get average. About 1 in 500 trainings, RWAV crashes and gets flipped. Then simulation can resume from a stable checkpoint.

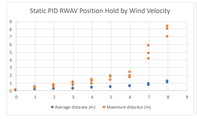
Static PID Position Hold with Random Wind

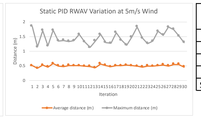
The effect of wind is analyzed from a RWAV with a static PID controller as a baseline. It reveals the RWAV's operation limit in simulation environment, and the wind environment setup is derived, where the performance of the DRL-engaged PID controllers should be tested.

Wind velocity sweep shows wind plugin and PID RWAV's behavior and limits. Standard deviation was set at 0.5m/s from mean velocity sweep. Max velocity is limited to 2m/s from mean velocity. Average distance is obtained from target over 1200 samples with 10Hz data for 2 minutes. 3 episodes per wind set up to observe repeatability. RWAV is put within 0.3m from the target position before collect data.

Maximum distance after 6m/s increases significantly. It suggests that static PID RWAV's capability is limited to 6m/s wind. RWAV crashed multiple times at 9m/s and failed to recover.

Wind velocity at 5m/s is used to compare static PID RWAV and RDL-engaged RWAV. Average distance is observed from total 30 episodes for 60 minutes. 30 episodes' wind statistics show there is up to 17% average distance changes and 7.4% standard deviation.





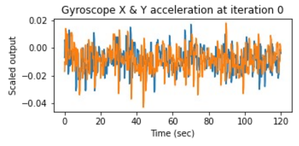
	Average distance	Maximum distance
Mean (m)	0.50	1.46
Standard deviation (m)	0.04	0.21
Minimum distance (m)	0.41	1.13
Maximum distance (m)	0.58	1.85
Standard deviation/mean (%)	7.4%	14.7%

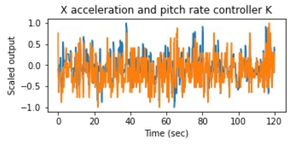
Results with DRL-Engaged Rate PID Controllers

The DRL action is set to engage with roll, pitch, and yaw rate controllers, as rate controller is the most fundamental and fastest (1kHz) in RWAV. Rate controller's roll and pitch rate KDI, and yaw rate KI coefficients are updated by DRL agents. There are total 8 action PID coefficient outputs from the DRL network.

Plots show gyroscope X- and Y-axis acceleration as observations to the DRL at iteration 0, and the DRL's action to rate controller coefficient K at iteration 1300 over X acceleration. The coefficient K tracks X acceleration to certain degree.

After 1500 iterations, the DRL could hold the target position as good as static rate PID controllers regarding average distances to target position. Normalized standard deviation was improved slightly to 6.2% from 7.4%.



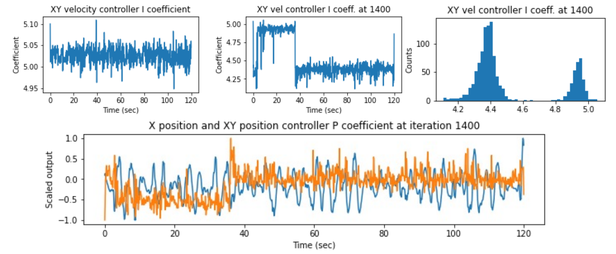


DRL Network with All PID Controllers

In addition to rate controllers, attitude, velocity, and position controllers are directed by the DRL to assess if there is further improvements from additional DRL involvement. Total 19 action PID coefficients are generated with the DRL network

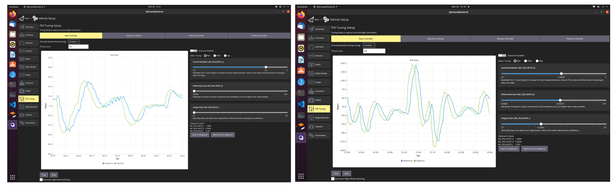
- Roll, pitch, and yaw attitude P
- Horizontal and vertical velocity PID
- Horizontal and vertical position P

Plots show XY velocity controller's I coefficient change between initial status and after 1400 iterations. The I coefficient's histogram shows bimodal distribution at two levels and variations around the levels. XY position controller coefficient P has similar behavior. DRL with all PID controller coefficient update improved position holding by 45% at 0.27m.



RWAV Behavior under DRL-Updated PID

Rate controller set point and tracking case 1 and case 2



QGroundControl visualizes RWAV PID controller's setpoint and tracking behavior to assist static PID tuning or even with DRL's dynamic PID coefficient update.

During training, RWAV could be in unstable mode to crash or to be flipped. This is likely from DRL's training induced random parameters to explore effectiveness in training. Also, DRL-updated coefficients seem to set unrealistic and vibrating responses time to time.

On the other hand, there was no crash reported during evaluation, where PID controllers were still dynamically updated by the DRL. PID controller loops were able to maintain flight within the tuning.

Analysis

The DRL with dynamic rate controller coefficient update showed a similar mean distance from target as statis PIC controller. Its normalized standard deviation was reduced to 6.2% from 7.4%. Rate controller runs at 1000Hz, while DRL update is 10Hz. The dynamic rate that DRL can update rate controller coefficients is a long-term update relatively, and it might not be able to operate effectively

The DRL with all controller coefficient update improved mean distance by 45% to 0.27m. The mechanism how the tracking behavior improves need further analysis. The qualitative understanding of telemetry data, PID controller coefficients, and DRL parameters would help refine the DRL mission controller

Summary of statistical data from experiments

Average distance	Static PID	DRL with Rate controllers	DRL with all PID controllers
Mean (m)	0.50	0.50	0.27
Standard deviation (m)	0.04	0.03	0.02
Standard deviation/mean (%)	7.4%	6.2%	7.9%

Conclusion

The DRL network tuned RWAV PID controllers dynamically, and its position hold was compared with static PID controllers under random 5m/s mean-velocity wind. The DRL-engaged rate PID controllers with 8 coefficients showed a comparable consistency at 0.50m and slightly reduced standard deviation. When the DRL managed all PID controllers with 19 coefficients, position consistency was improved by 45% at 0.27m.