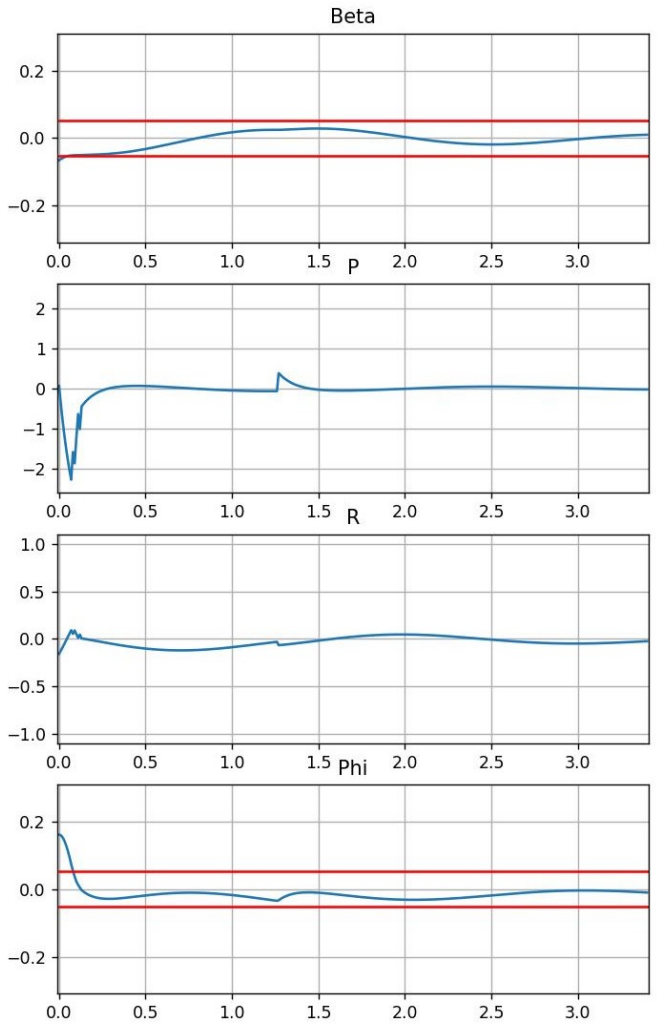
# Meeting

Po Hsun Wu

September, 2022

# Progress report

- Writing conference paper
- Study PPO and TRPO algorithm



# TRPO algorithm

- TRPO optimal problem

$$\begin{aligned} &\max_{\theta} \hat{\mathbb{E}}_t \left[ \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t \right] \\ &\text{subject to } \hat{\mathbb{E}}_t \left[ \text{KL} \left[ \pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t) \right] \right] \leq \delta \end{aligned} \tag{1}$$

- Trust Region method is to solving nonlinear problem

$$\begin{aligned} &\min m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p \\ &\text{subject to } \|p\| \leq \Delta_k \end{aligned} \tag{2}$$

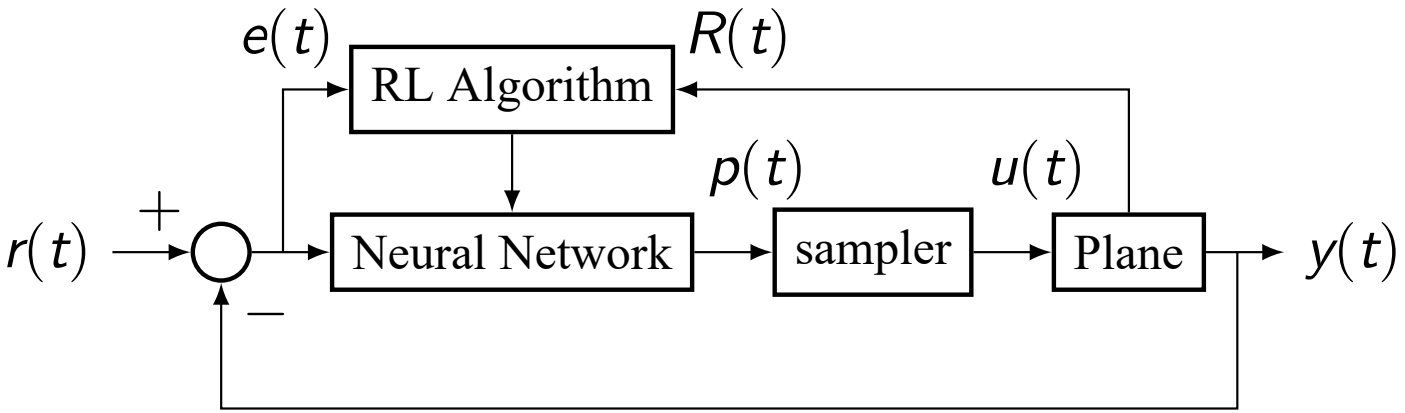
# PPO algorithm

- PPO(Proximal Policy Optimization) is upgrade from TRPO(Trust Region Policy Optimization)
- PPO optimal problem

$$\min_{\theta} \hat{\mathbb{E}}_t \left[ \min \left( \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t,\ \text{clip} \left( \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)},\ 1-\epsilon,\ 1+\epsilon \right) \hat{A}_t \right) \right] \tag{3}$$

clip function is to set the policy change inside $[1-\epsilon, 1+\epsilon]$

# Architecture




$r(t)$: reference
$y(t)$: state
$e(t)$: error
$R(t)$: reward
$p(t)$: probability of actions
$u(t)$: control