Weekly Report

M2 Ibuki Takeuchi

- Report on last week
  - It was possible to apply RL to event-triggered control
  - It will be applicable to self-triggered control as well
  - (At least with same steps,) it was not conducted well
    - Algorithm did not converge
  - Conjecture:

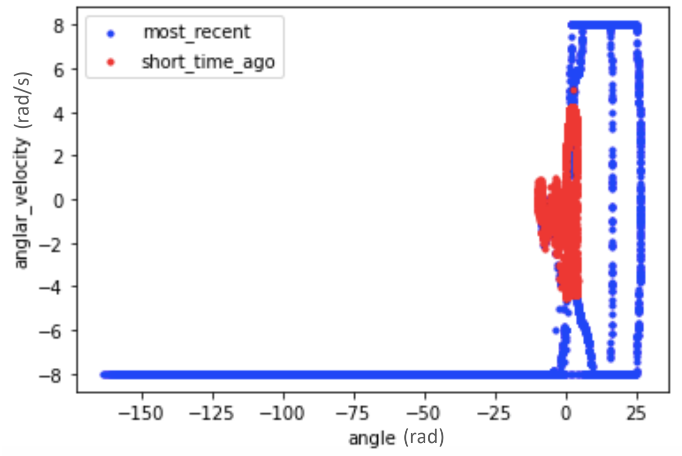
    The control performance changes rapidly with little parameter change
  - If the conjecture is valid, research the condition and consider solution

→ Check the validity (conclusion: NO)

Weekly Report

M2 Ibuki Takeuchi

- Recently experienced data



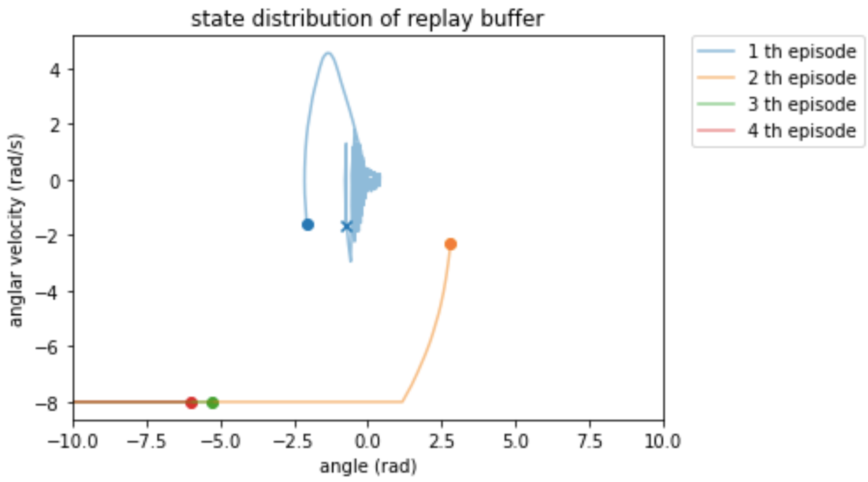
- Although parameter changes slightly, control path changes drastically(?)

- This scatter plot does not show control path

Weekly Report

M2 Ibuki Takeuchi

- Depict control path for each episode(new starting point)



- Assume
  - Policy parameters are updated in all steps
  - Initial policy stabilize all initial states

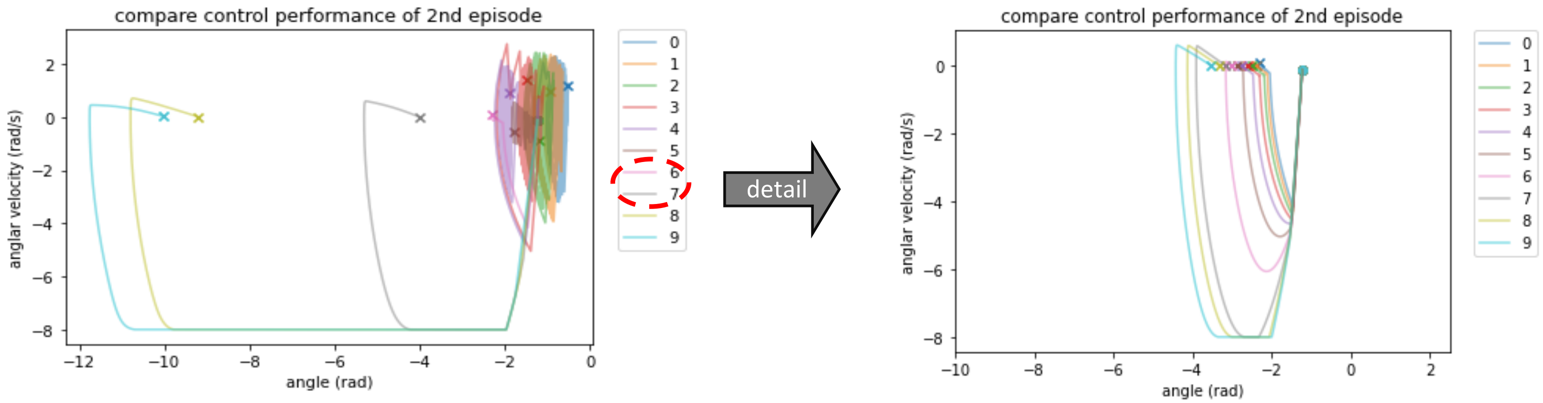
- It is not possible to see why policy became bad with this picture
  - (P1) Control path suddenly changed in second episode (conjecture)
  - (P2) Overfit to blue path
  - (P3) Other reason

Weekly Report

M2 Ibuki Takeuchi

- How the policy changes in 2nd episode
  - Pick up policy in equally step interval

    ex.) step 100, 200, 300, ...



(Attention: policy are not updated in each path)

- Control path does not change suddenly
  - The conjecture is not valid

Weekly Report

M2 Ibuki Takeuchi

- Next step: Reconsider the cause for policy deterioration
  - policy gradient approximation

$$g = \frac{1}{N}\sum_{s\in E}\left[\nabla_\theta \pi(s|\theta)\nabla_a Q(s,a|\omega)|_{a=\pi(s|\theta)}\right]$$
$$\approx \mathbb{E}_{s\sim\rho^{\pi_\theta}}\left[\nabla_\theta \pi_\theta(s)\nabla_a Q^\theta(s,a)|_{a=\pi(s|\theta)}\right]$$

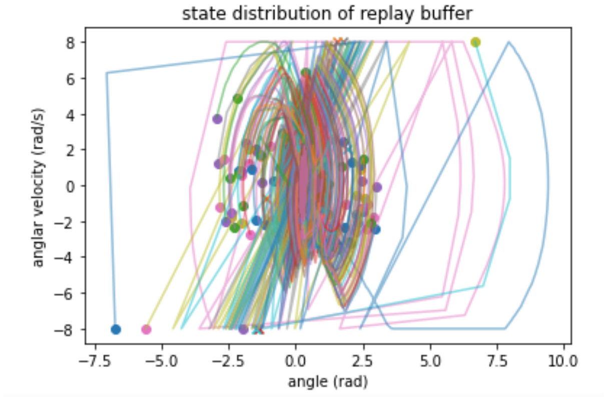
    - This assumes experienced states is well approximates ρ^{π_θ}
    - In other words, improve policy by prioritizing only experienced states
    - Lack of the number of episodes makes policy overfitting

  - For data efficiency, by enlarging minimum interval time, increase the number of control paths in replay buffer (N steps experienced data)

Weekly Report

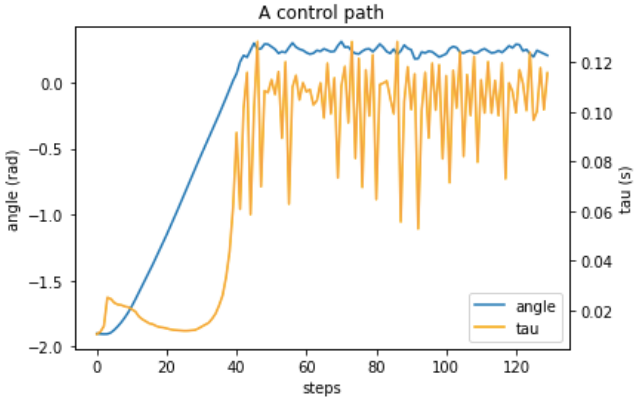
M2 Ibuki Takeuchi

- By changing learning configuration as last slide, policy may improved
  - Distribution of replay buffer



- There is no divergent paths

  - Learned policy



There is no guarantee that this policy is the best policy …

- Wide interval around origin and frequent otherwise
- Stabilize the system

Weekly Report

M2 Ibuki Takeuchi

- This week
  - Reconsider and discuss the theme

  - Ideas for stable learning
    - Configuration of optimizer (learning late hyperparameter etc..)
    - Safe learning