Weekly Report

M2 Ibuki Takeuchi

- Summary
  - Stance: Consider the next step for master thesis

  - Modify approximation of value function $V^{\pi}(s)$
  - Extract issues from comparison between $Q^{\pi}(s,a)$ and critic $Q(s,a|\omega)$

approximates

- Review
  - $V^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t r\big(s_t, \pi(s_t)\big), s_0 = s, \gamma \in (0,1]$
  - $Q^{\pi}(s,a) = r(s,a) + \gamma V^{\pi}(s')$, s': next state
  - Policy gradient:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s\sim\rho^{\pi_\theta}}\left[\nabla_\theta \pi_\theta(s) \nabla_a Q^{\pi_\theta}(s,a)|_{a=\pi_\theta(s)}\right]$$

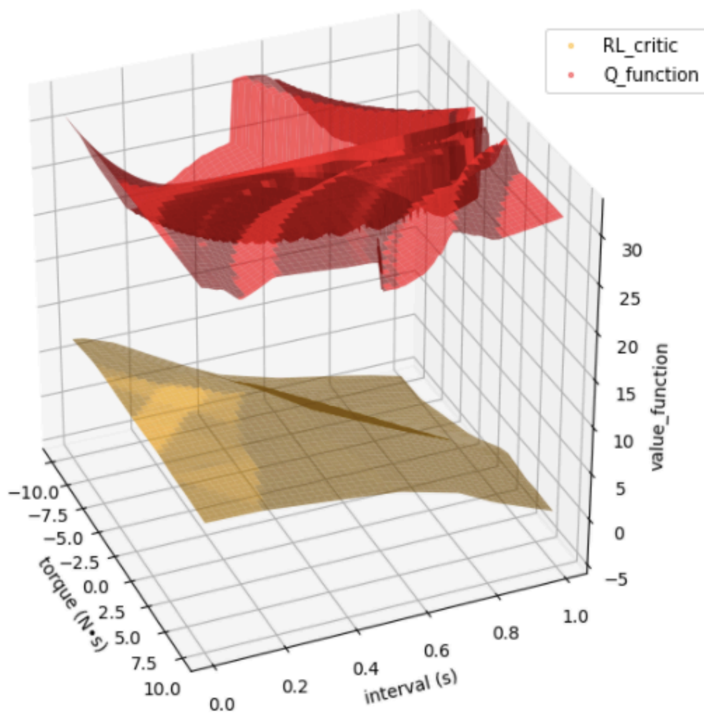
$$\rho^{\pi_\theta}(s) = \int_S \sum_{t=0}^{\infty} \gamma^t d_0(s_0)\mathbb{P}(s_0 \to s, t, \pi_\theta) ds_0$$

Weekly Report

M2 Ibuki Takeuchi

- Comparison between $Q^{\pi}(s,a)$ and critic $Q(s,a|\omega)$
  - DDPG assumes that $\nabla_a Q^{\pi}(s,a) = \nabla_a Q(s,a|\omega)$
  - (at least) the shape of Q function for one s should be similar

  - Shape of 2 functions at $s = [0,0]$ (a is 2 dimension)



- Critic could not learn Q function during reinforcement learning

Weekly Report

M2 Ibuki Takeuchi

- The reason for poor approximation performance
  - critic $Q(s, a|\omega)$ is fitted with supervised learning
  - The variance of (s, a) should be large
    (algorithm requires performance only for high-frequency states in a distribution ρ^{π_θ})
  - To meet request above, enough action exploration is needed

Exploration-Exploitation Dilemma

  - Distribution of experienced states should not be dissociated from ρ^{π_θ}

$$g = \frac{1}{N}\sum_{s \in E} \left[\nabla_\theta \pi(s|\theta) \nabla_a Q(s, a|\omega)|_{a=\pi(s|\theta)}\right] \quad \text{actor's gradient}$$

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

experienced data

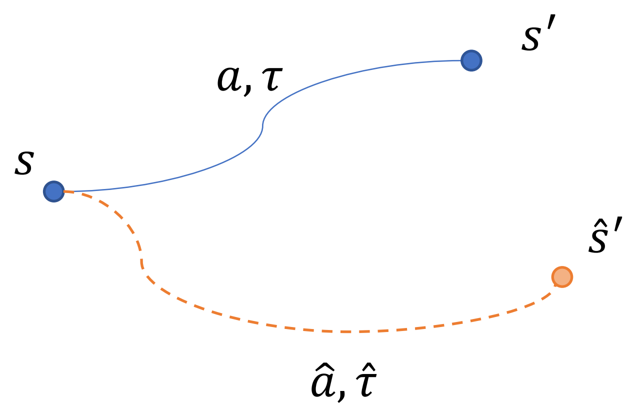
Weekly Report

M2 Ibuki Takeuchi

- Idea of thesis
  - To propose a method of good exploration noise ($\mathrm{u} = \pi(s) + e$)

1. Similarity of the empirical state distribution and ρ^{π_θ}
2. Various inputs for each state

  - Adaptive noise scaling($\simeq$ variance) w.r.t. control path



- s' is a function of (s, a, τ)

  If $\dfrac{\partial s'}{\partial a}, \dfrac{\partial s'}{\partial \tau}$ is large
  $\rightarrow$ small noise
  else:
  $\rightarrow$ large noise

  - s' needs f, g of $\dot{s} = f(s) + g(s)a$

Weekly Report

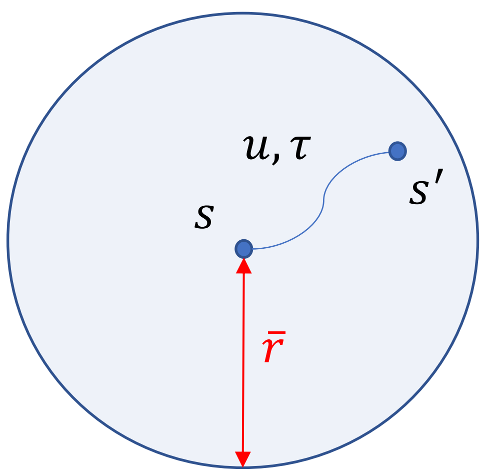
M2 Ibuki Takeuchi

- Lost generality when we use system dynamics i.e. f, g

- [1] shows the upper bound of state change on self-trigger control

$$||s' - s|| \leq \frac{1}{L} ||f(s) + g(s)a||(e^{L\tau} - 1)(= \bar{r}(s, a, \tau))$$

f, g: Lipcshitz continuous



Draft of noise scaling

- derivative of radius
- size of this circle

[1]: G. Yang, C. Belta and R. Tron, "Self-triggered Control for Safety Critical Systems Using Control Barrier Functions," *2019 American Control Conference (ACC)*, Philadelphia, PA, USA, 2019, pp. 4454-4459.

- criticがQ関数を近似できていない
- その原因は経験データの偏りにある
- 経験データの分散を上げる為の探索ノイズを大きくしたい
- 単純にノイズを大きくすればいいってものでもない
- ノイズの大きさの工夫について考えたい