Weekly Report

M2 Ibuki Takeuchi

- Read papers
  - [1]: T. Yu, G. Thomas, L. Yu, S. Ermon, J. Zou, S. Levine, C. Finn and T. Ma. “MOPO: Model-based Offline Policy Optimization.” *arXiv preprint arXiv: 2005.13239,* 2020.
  - [2]: M. Janner, J. Fu, M. Zhang and S. Levine. “When to Trust Your Model: Model-Based Policy Optimization.” *In Advances in Neural Information Processing Systems*, pp. 12498-12509, 2019. (in progress)

- MOPO[1] is an application of MBPO[2] to offline RL

Weekly Report

M2 Ibuki Takeuchi

- MOPO[1] is (a)model-based (b)offline RL

- (a) model-based approach
  - has better efficiency of data sampling
  - has (not small) issue of evaluation error ← MBPO[2]

- (b) offline approach
  - avoid interaction between environment (possibly dangerous)
  - does not have method to avoid the distributional shift ← MOPO[1]

- It is known that model-based RL is superior to model-free RL in offline fashion.

Weekly Report

M2 Ibuki Takeuchi

- In [1]
  - MDP $M = (S, A, T, r, \mu_0, \gamma)$
    - $T(s'|\, s, a)$ is a transition probability distribution (True system)
  - MDP $\widehat{M} = \left(S, A, \hat{T}, r, \mu_0, \gamma\right)$
    - $\hat{T}(s'|\, s, a)$ is the estimation of $T(s'|\, s, a)$, learned by MBPO[2]
  - $\rho_{\hat{T}}^{\pi}(s, a)$: probability to have (s, a) with $\hat{T}$ and π

  - $d_F(T(s, a), \hat{T}(s, a))$: a kind of distance between T and $\hat{T}$
    - We assume $d_F\left(T(s, a), \hat{T}(s, a)\right) \leq u(s, a)$
    - Penalized MDP : $\widetilde{M} = \left(S, A, \hat{T}, \tilde{r}, \mu_0, \gamma\right)$, $\tilde{r}(s, a) = r(s, a) - \lambda u(s, a)$
    - $\widetilde{M}$ gives lower bound: $\eta_M(\pi) \geq \eta_{\widetilde{M}}(\pi)$

      where $\eta_M(\pi) = \mathbb{E}_{\pi, T, \mu_0}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$

Weekly Report

M2 Ibuki Takeuchi

- Apply policy optimization (like SAC) over $\widetilde{M}$

$$\hat{\pi} = argmax_{\pi}\, \eta_{\widetilde{M}}(\pi)$$

- If we have $u(s,a)$
  - Learned policy $\hat{\pi}$ is evaluated as follows:

$$\eta_M(\hat{\pi}) \geq sup_{\pi}\{\eta_M(\pi) - 2\lambda\varepsilon_u(\pi)\}$$

where $\varepsilon_u(\pi) = \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}} u(s,a)$

- Because we cannot have exact $u(s,a)$, [1] uses heuristic surrogate
- [1,2] uses neural network to approximate dynamics T