制御システム論分野研究会資料

2020 年 8 月 28 日

M2 竹内 維吹

進捗報告 8.28

1　サンプル値系システムの強化学習

対象のシステムが

$$s_{t+1} = f(s_t) + g(s_t)a_t \tag{1}$$

と書かれていて, τ ステップ毎に状態を観測し, 状態フィードバック制御則を変えて次の観測までは同じ入力を加え続ける, サンプル値系における最適制御問題の強化学習を考える. この動機は τ ステップ毎に観測・制御則更新を行うのと, それを毎ステップ行うのとで学習に必要なステップ数が変わるのかを検証するためである.

結論から記すと, おそらくはサンプル値系にしたからといって学習時間に大きな変化はないと考える.

2　倒立振子による実験

倒立時の振子の角度を $\theta = 0$ とし, 加えられる入力が $A = [-10\mathrm{N}\cdot\mathrm{m}, 10\mathrm{N}\cdot\mathrm{m}]$ と制限されるような倒立振子を考える. この倒立振子のダイナミクスは, 以下のように与えられる.

$$\theta_{t+1} = \theta_t + \dot{\theta}_t\delta_t + \frac{3g}{2l}\sin\theta_t\delta_t^2 + \frac{3}{ml^2}a\delta_t^2 \tag{2}$$

$$\dot{\theta}_{t+1} = \dot{\theta}_t + \frac{3g}{2l}\sin\theta_t\delta_t + \frac{3}{ml^2}a\delta_t \tag{3}$$

これは式 (**??**) に対応する. 本実験ではこのダイナミクスが既知であるとして $U(s)$ を構築し, 状態制約 (**??**) を満たしながら π^* を求めることができるのかを検証する. ただし, δ_t は離散化定数であり $\delta_t = 0.005$ とする.

これに対応する連続時間システムを書き下すと

$$\frac{\mathrm{d}}{\mathrm{dt}}\begin{pmatrix}\theta\\ \dot{\theta}\end{pmatrix} = \begin{pmatrix}\dot{\theta}\\ \frac{3g}{2l}\sin\theta + \frac{3}{ml^2}a\end{pmatrix} \tag{4}$$

となる.

3　セルフトリガー制御にむけて

毎ステップ観測の最適なエージェントを初期値として, サンプル間隔 τ を変えた時に制御性能を満たしながら”サボる”ことを学習できるのかどうか検証したい.

参考文献

[1] D. Baumann, J. J. Zhu, G. Martius, and S. Trimpe. “Deep Reinforcement Learning for Event-Triggered Control.” *In Proc. of the 57th IEEE International Conference on Decision and Control*, 2018.

[2] Li Wang, Evangelos A Theodorou, and Magnus Egerstedt. "Safe learning of quadrotor dynamics using barrier certificates," *In 2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2460-2465, 2018

[3] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick. "End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks," *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.