制御システム論分野研究会資料

2020 年 7 月 27 日

M2 竹内 維吹

修論に向けてのテーマ草案

1 仮定

実環境システムが以下のような入力アフィン系であると仮定する.

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

ここで, s_t, a_t はそれぞれ状態ベクトル, 入力ベクトルである. また, f_{true}, g はリプシッツ連続であるとする.

次に, 強化学習エージェントが上記のダイナミクスの公称知モデルが既知であると仮定する. つまりエージェントが関数 f, g を与えられ, 式 (1) が以下のように変形される.

$$s_{t+1} = f(s_t) + g(s_t)a_t + d(s_t) \tag{2}$$

ここで, $d(s_t)$ は未知関数である (これをガウス過程回帰を用いて推定する).

2 準備

2.1 安全領域と制御バリア関数

前節の仮定のもと, マルコフ決定仮定 $M = \{S, A, f, g, d, \gamma, r, d_0\}$ における安全強化学習法を考えたい. ここで S は状態集合, A は入力集合, r は報酬関数, d_0 は初期状態分布である. 安全な強化学習とは, 安全制約 (後述) を与える集合 $S_{safe} \subset S$ に対して, $s \in S \backslash S_{safe}$ を学習過程において取らないようにするということである. これを S_{safe} を前進不変 (forward invariant) にするという.

ここで, 安全状態集合を以下のようにある関数 $h(s)$ に対する優位集合として定義する.

$$S_{safe} = \{s \in S \mid h(s) \geq 0\} \tag{3}$$

次に上記の安全状態集合を前進不変にする安全入力制約を定式化するため, 制御バリア関数を導入する. これは以下を満たす関数 $h(s)$ の集合を表すものである. ($\kappa(\cdot)$ は $\kappa(0) = 0$ を満たす狭義単調増加関数)

$$\sup_{a \in A} \left\{ \frac{\partial h}{\partial s}(f(s) + g(s)a + d(s)) + \kappa(h(s)) \right\} \geq 0, \forall s \in S_{safe} \tag{4}$$

もし, 安全状態集合 S_{safe} を定義する関数 $h(s)$ が制御バリア関数であるならば, $\frac{\partial h}{\partial s}(f(s) + g(s)a + d(s)) + \kappa(h(s)) \geq 0$ を満たす入力 $a \in A$ が, 全ての $s \in S_{safe}$ に対して少なくとも 1 つ存在することを保証する. よって, 状態 s に対する安全入力集合は以下のように与えることができる.

$$U(s) = \left\{ a \in A \mid \frac{\partial h}{\partial s}(f(s) + g(s)a + d(s)) + \kappa(h(s)) \geq 0 \right\}, \forall s \in S_{safe} \tag{5}$$

$U(s)$ を s に対する安全入力集合と呼ぶ理由は, $a \in U(s)$ を入力することにより, S_{safe} を前進不変にできるか

らである. 上記の記法を用いると, 安全方策 $\pi_{safe}(a|s)$ は以下のクラス Π_{safe} に限定できる.

$$\Pi_{safe} = \left\{ \pi \mid \sum_{a \in U(s)} \pi(a|s) = 1 \right\} \tag{6}$$

この Π_{safe} は自明に $\Pi = \left\{\pi \mid \sum_{a \in A} \pi(a|s) = 1\right\}$ のサブクラスになる.

2.2 安全領域の拡張

[2] では, 式 (2) における $d(s)$ をガウス過程回帰により,

$$m(s) - k_\delta \sigma(s) \leq d(s) \leq m(s) + k_\delta \sigma(s) \tag{7}$$

という $\pm k_\delta \sigma$ までの信頼区間を推定し, 以下の最適化問題を解くことで制御バリア関数を決定していた.

$$\begin{aligned} &\max_{\mu} \ \mathrm{vol}(S_{safe}) \\ &\text{s.t.} \ \sup_{a \in A} \left\{ \frac{\partial h_\mu}{\partial s}(f(s) + g(s)a + m(s)) - k_\delta \left| \frac{\partial h_\mu}{\partial s} \right| \sigma(s) + \kappa(h_\mu(s)) \right\} \geq 0 \end{aligned}$$

ここで, 単項式からなるベクトル値関数 $Z(s)$ を用いて, 制御バリア関数の候補を $h_\mu(s) = 1 - Z(s)^\top \mu Z(s)$ としている. この h_μ は凹関数なので, その優位集合 S_{safe} は凸集合となる.

$\mathrm{vol}(S_{safe})$ が大きくなれば, 状態空間の探索可能範囲が広がり $|\sigma(s)|$ が小さくすることができるので, $\mathrm{vol}(S_{safe})$ をより大きくするスキームを作ることができる [2]. $|\sigma(s)|$ は, その大きさが大きい s から探索することで効率よく小さくしていくことが可能である. しかし [2] ではその s に移動させるための入力 u^{explore} が既知であるものとしてした. したがって今後の研究にむけて, 状態空間探索方策をエージェントが自律的に学習できるようにしたい. そのヒントとして次節の [3] がある.

2.3 モデル推定誤差を考慮した強化学習

3 研究の目的

今後研究を進めていくにあたり, 以下の 2 つを目標に据えたい.

- [2] で考慮されていなかった探索方策をエージェントが自律的に決定する手法の開発
- $D_{KL}(\underset{\pi \in \Pi}{\mathrm{argmax}}\, J(\pi) | \underset{\pi \in \Pi_{safe}}{\mathrm{argmax}}\, J(\pi)) = 0$ とできる条件についての調査

まず 1 つ目については, [2] で紹介されていた制御バリア関数と未知関数 $d(s)$ の学習に, [3] を組み合わせることで開発していきたい.

2 つ目については, S_{safe} の凸性を活かしたい. コスト $r(s,a)$ が $s \in S_{safe}$ ならどんな a に対しても $s \in S \backslash S_{safe}$ の $r(s,a)$ よりも大きくあればいいのじゃないか？

未知関数 $d(s)$ の学習が完了すれば, それを用いたオフライン model-based 強化学習 MOPO を用いる

quad の内容 modelbased にした時の評価関数の差状態制約による評価関数の差

参考文献

[1] S. Levine, A. Kumar, G. Tucker and J. Fu. “Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems, ” *arXiv preprint arXiv: 2005.01643*, 2020.

[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] 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.