制御システム論分野研究会資料

2020 年 8 月 15 日

M2 竹内 維吹

**進捗報告 8.15**

# 1　問題設定 (作成中)

以下のような入力アフィン系システムを考える.

$$\dot{s} = f(s) + g(s)a \tag{1}$$

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.

さらに, 図 1 のようなシステム (1) に対するイベント駆動型制御を考える.

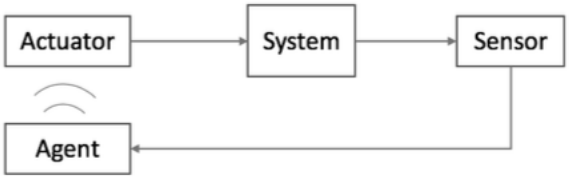


図 1　イベント駆動型制御

ここではエージェントの制御則 $\pi(s)$ を設計するとし, $\pi(s)$ はシステムに加えるべき入力信号 $a$ と, それを送信するかどうかを決定する 2 値変数 $c$ ($c = 1$ ならば送信) で構成されるとする. つまり

$$\pi(\cdot) = [\ a, c\ ]^\top \tag{2}$$

であるとする. このようなイベント駆動型制御に対して, 以下の条件を満たす最適イベント駆動型制御則 $\pi^*(\cdot)$ を導出する問題を考える.

$$\pi^*(\cdot) = \underset{\pi}{\mathrm{argmax}}\, \mathbb{E}_{s_0 \sim d_0}\left[V^\pi(s_0)\right] \tag{3}$$

$$V^\pi(s_0) = \sum_{t=0}^{\infty} (-s_t^\top Q s_t - \pi^\top(s_t) R \pi(s_t) - \lambda \gamma_t) \tag{4}$$

ここで, $\gamma_t$ は時刻 $t$ においてエージェントがアクチュエータと通信を行ったかを表す 2 値変数である. また, $Q, R, \gamma$ はそれぞれ正定値, 半正定値, 正のハイパーパラメータであり, $d_0$ は初期状態を与える確率分布である. 式 (4) より, 「最小限の入力エネルギーで」かつ「最小限の通信回数で」「状態 $s$ を 0 に素早く漸近させる」と $V$ が最大化される.

さて, システム (1) が未知であるという設定のもとで, 実環境とのインタラクションによってデータ組 $(s_t, a_t, r_t, s_{t+1})$ を収集し, それらを活用することで上記の問題を解いていたのは [1] である.

本研究ではさらに一歩踏み込んで, データ収集の実環境とのインタラクションの際に状態制約

$$s \in C, C \subset S \tag{5}$$

を全時刻において満たしながら $\pi^*(\cdot)$ を求める問題を考える. ただし, 式 (3) における $d_0$ は $\mathrm{support}(d_0) \subset C$ を満たすとする. また, $S$ は実環境において考えられうる全状態の集合であるとする.

(現状, 上記の問題を解くことが研究の目標であると考えています. ただ, この先どのような仮定を置くかはまだ考えられていません. )

## 2 DDPG と CBF を用いた解法

システム (1) に対して, 関数 $h(s)$ が以下の式を満たすならば, $h(s)$ は制御バリア関数 (CBF) と呼ばれる.

$$\sup_{a \in A} \left\{ \frac{\partial h}{\partial s}(f(s) + g(s)a) + K(h(s)) \right\} \geq 0 \tag{6}$$

ただし, $K(s)$ はクラス K 関数である.

さて, 式 (5) における状態制約 $C$ が

$$C = \{s \in S \mid h(s) \geq 0\} \tag{7}$$

として与えられているとする. このとき $h(s)$ が制御バリア関数であるならば, 状態 $s \in C$ に対して次の時刻における状態 $s'$ が $s' \in C$ を満たすようにする入力が存在することを保証する. そのような入力集合は現時刻での状態 $s$ に依存し,

$$U(s) = \left\{ a \in A \mid \frac{\partial h}{\partial s}(f(s) + g(s)a) + K(h(s)) \geq 0 \right\}, \forall s \in C \tag{8}$$

としてその集合を与える.

さて, DDPG などの方策 on 型の強化学習では, 制御則 $\pi(\cdot)$ を用いて実環境とインタラクションを行い, データを収集・利活用することで, より $V^\pi$ を大きくできるような制御則 $\pi'(\cdot)$ を模索する. その過程で制御入力 $\pi(s)$ が $\pi(s) \notin U(s)$ となるならば, 次時刻において状態制約 $C$ を満たさない状態に遷移してしまう. それを避けるため, $\pi(s) \notin U(s)$ となった各時刻では, $\pi(s)$ ではなく, $\pi(s)$ に最も近い $U(s)$ の元を用いてインタラクションを行う手法を考えてみる.

## 3 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.

$$\theta_{t+1} = \theta_t + \dot{\theta}_t \delta_t + \frac{3g}{2l} \sin\theta_t \delta_t^2 + \frac{3}{ml^2} a \delta_t^2 \tag{9}$$

$$\dot{\theta}_{t+1} = \dot{\theta}_t + \frac{3g}{2l} \sin\theta_t \delta_t + \frac{3}{ml^2} a \delta_t \tag{10}$$

これは式 (1) 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 (5) を満たしながら $\pi^*$ を求めることができるのかを検証する. ただし, $\delta_t$ は離散化定数であり $\delta_t = 0.05$ とする.

ただし, これは離散化された状態方程式であるため, CBF による状態制約の前進不変性をより厳密に議論するために対応する連続時間システムを書き下すと

$$\frac{\mathrm{d}}{\mathrm{dt}} \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} \dot{\theta} \\ \frac{3g}{2l} \sin\theta + \frac{3}{ml^2} a \end{pmatrix} \tag{11}$$

となる. 与えられた関数 $h(s)$ が制御バリア関数か否かを調べるには, 式 (11) と式 (6) を用いる.

さて, $s = [\ \theta, \dot{\theta}\ ]^\top$ とすると $S = \{s \mid \theta \in [-\pi, \pi], \dot{\theta} \in \mathbb{R}\}$ である. ここで, 状態制約集合 $C \in S$ を

$$C = \{s \in S \mid h(s) \geq 0\} \tag{12}$$

とし, $h(s) = (1 - \theta^2 - \alpha\dot{\theta}^2), \alpha > 0$ とおく. すると, 状態制約集合 $C$ は以下のように分布する.

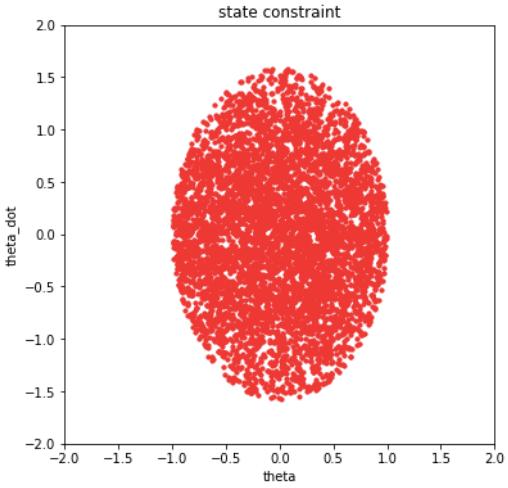


図 2　状態集合 $C$

式 (6) における $K$ を $K(x) = \gamma x$ とおくと式 (6) は次のようになる.

$$\sup_{a \in A} \left\{ -2\theta\dot{\theta} - \frac{3g\alpha}{l}\dot{\theta}\sin\theta - \frac{6\alpha}{ml^2}\dot{\theta}a + \gamma(1 - \theta^2 - \alpha\dot{\theta}^2) \right\} \geq 0 \tag{13}$$

括弧の中は $a$ に関する 1 次式となっているため $\dot{\theta}$ の正負に合わせて $U(s)$ を定義できる ($\frac{6\alpha}{ml^2}$ は必ず正のため).

さて, 括弧の中身を

$$p(a) = -2\theta\dot{\theta} - \frac{3g\alpha}{l}\dot{\theta}\sin\theta - \frac{6\alpha}{ml^2}\dot{\theta}a + \gamma(1 - \theta^2 - \alpha\dot{\theta}^2) \tag{14}$$

としたとき, $p(a) = 0$ の解を $a^*$ とすれば, $a$ は 1 次元なので $\theta < 0$ のとき $U(s)$ は以下のようになる.

$$U(s) = \begin{cases} [\ a^*, 10\ ] & \text{if} \quad a^* > -10 \\ [\ -10, 10\ ] & \text{if} \quad a^* \leq -10 \end{cases} \tag{15}$$

また, $\theta > 0$ のときも同様にすると

$$U(s) = \begin{cases} [\ -10, a^*\ ] & \text{if} \quad a^* < 10 \\ [\ -10, 10\ ] & \text{if} \quad a^* \geq 10 \end{cases} \tag{16}$$

として与えることができる.

## 4　現状

図 2 にプロットされている状態 $s$ に対して, $U(x)$ の元を入力しているにもかかわらず, 「シミュレーション上の」次時刻の $h(s_{t+1})$ が 0 よりも小さくなる時があります.

その原因を探るべく, 水曜日の zoom での MTG で, $C$ 内での $U(x)$ をプロットしてみるということも試してみようと話していた (?) ので, 見てみました.

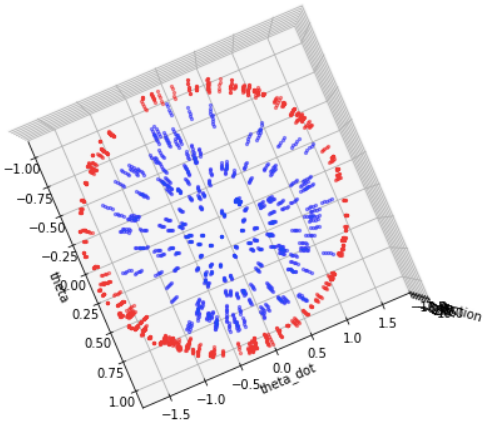


図 3　$\theta, \dot{\theta}$ 平面への投影

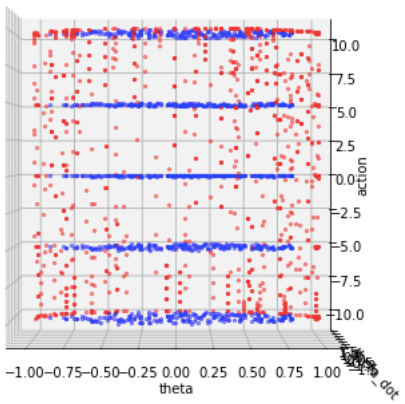


図 4　$\theta, a$ 平面への投影

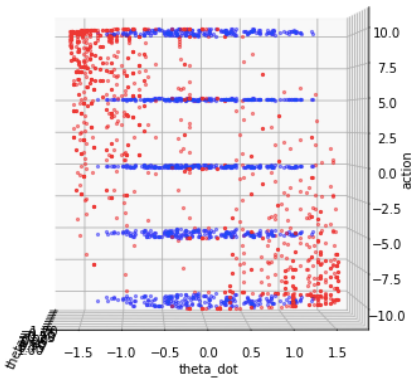


図 5　$\dot{\theta}, a$ 平面への投影

$C$ の要素の $\theta, \dot{\theta}$ に対して, $U(s)$ の要素を 5 個づつ等間隔に選び, 3 次元空間上にプロットしました. 赤色の点は $C$ のフチで, 青色の点は $C$ の内部の $\theta, \dot{\theta}$ に対する $U(s)$ を示しています. 図 4 を見ると, $\theta$ が $-1$ に近く

ても, 負の方向への入力を許していることがわかります. さらに図 5 を見ると, その場合には $\dot{\theta}$ が正方向にあるとわかります. これは直感通り (フチにいても, 逆方向の速度があれば OK, という直感) なので, 大まかには $U(s)$ を正しく設定できていると思います.

なぜそれでも制約が壊れるかを調べるため, $C$ のフチの状態 $s = [-0.99006819, -0.16353134]$ を 1 つ取ってきて, 離散化定数 $\delta_t$ を変えてシミュレーションしてみました.

この時 $U(s)$ は $[-4.192729709679552, 10.0]$ と計算されるため, 入力 $a = -4.192729709679552$ を加えてみました. 離散化定数 $\delta_t = 0.05$ のときは

$$\begin{aligned} s' &= [-1.06104261 - 1.44708933] \\ h(x) &= -0.9634384239222334 \end{aligned}$$

となるのに対して, 離散化定数 $\delta_t = 0.005$ のときは

$$\begin{aligned} s' &= [-0.99015624 - 0.17609127] \\ h(x) &= 0.0071873749226185266 \end{aligned}$$

となり, 状態制約を満たします. ここから, 先生のおっしゃっていた通り, 離散化エラーが問題であったことがわかりました. さらに, 離散化定数を限りなく 0 に近づければ $h(s) = 1 - \theta^2 - \alpha\dot{\theta}^2$ によって安全強化学習できそうだと予測できると思います.

## 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.