セルフトリガー制御に対する深層強化学習

数理工学専攻 制御システム論分野

竹内 維吹

準備: 強化学習の基礎知識

- 強化学習の目的
  - 全ステップの累積コストを最小化する方策π^*を求める

$$\min_{\pi} \ \mathbb{E}_{s\sim d_0}[V^{\pi}(s)]$$

$$s.t.\ V^{\pi}(s) = \mathbb{E}_{w}\left[\sum_{i=0}^{\infty} \gamma^i r(s_i, \pi(s_i)) \,|s_0 = s\right] \quad i\text{: ステップ数}$$

環境雑音

  - $\pi(s)$: 制御則
  - $r(s,a)$: 状態sで行動aをとったときのコスト
  - $\gamma \in [0,1)$: 割引率, 小さいほど先のステップのコストを軽視

- 行動価値関数$Q^{\pi}(s,a)$
  - 状態sでまず自由に行動aを行い, 次ステップから方策πで制御したときの割引付き累積コスト

$$Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V^{\pi}(s'(s,a))] \quad \text{(ベルマン方程式)}$$

準備: 方策勾配型(深層)強化学習

- 方策πをパラメータθをもつニューラルネットワークで表現
- 評価関数$J(\pi_\theta) = \mathbb{E}_{s\sim d_0}[V^\pi(s)]$の$\theta$勾配を用いて方策を更新
- 決定的方策勾配定理[1]
  - 方策が状態sから行動aへの関数の場合の勾配

$$\nabla_\theta J(\pi_\theta) = \int_S \rho^{\pi_\theta}(s)\nabla_\theta \pi_\theta(s)\nabla_a Q^{\pi_\theta}(s,a)|_{a=\pi_\theta(s)} ds$$

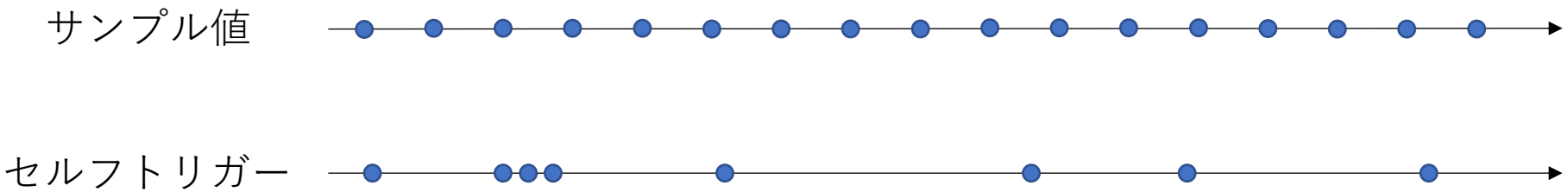
$$\rho^{\pi_\theta}(s) = \int_S \sum_{t=0}^{\infty} \gamma^t d_0(s_0)\mathbb{P}(s_0 \to s, t, \pi) ds_0$$

[1]: Silver et al., “Deterministic Policy Gradient Algorithms”, *ICML*, 2014

イントロダクション

- サンプル値制御
  - 連続時間システムを一定時間間隔で制御する手法
  - 各通信の間は, 同じ入力を加え続ける
  - 制御入力の変更が小さい場合は非効率な通信を行うことになる

- セルフトリガー制御
  - システムの状態などから, 次の通信時刻を制御器が臨機応変に決定

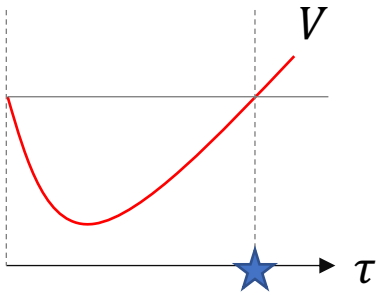


セルフトリガー制御における先行研究

- 1ステップの最適化による手法
  - [2]は連続時間システムに対して

$$\min_{u} u^T u$$

$$\text{s.t. } \mathcal{L}_f V(s) + \mathcal{L}_g V(s) u + \varepsilon V(s) \leq 0$$



の解uを加え続けたときに, 次ステップでのリアプノフ関数Vの値が減少する最大の通信時間τを選択する手法を提案した

- 長時間の制御全体における, 通信コストの最適性は考慮していない

[2]: Silver et al., “Deterministic Policy Gradient Algorithms”, *ICML*, 2014

最適セルフトリガー制御問題の定式化

- 最適化問題

$$\min_{\theta} \ \mathbb{E}_{s\sim d_0}[V^{\pi_\theta}(s)]$$

$$s.t.\ V^{\pi_\theta}(s) = \int_0^{\infty} e^{-\alpha t}\mathbb{E}_w[s^T(t)Es(t) + u^T(t)Fu(t) + \beta\delta(t)C(t)|s(0) = s]\,dt$$

$$\dot{s} = f(s) + g(s)u + \dot{w}$$

  - $e^{-\alpha}$: 割引率, 小さいほど直近のコストを軽視
  - $C(t)$: 時刻tにおいて通信をしたかどうかの0,1変数
  - $u(t), C(t)$は方策$\pi_\theta(s)$によって決定
  - α, β, E, F: ハイパーパラメータ

- 本研究では, この問題を強化学習で解けるのかを検討

強化学習問題としての定式化

- 強化学習はステップ毎のコストの和に対する最適化問題
- 通信することを1つのステップとみなして, 価値関数を分解

$$V^{\pi}(s) = \int_0^{\infty} e^{-\alpha t}\mathbb{E}_w[s^T(t)Qs(t) + u^T(t)Ru(t) + \beta\delta(t)C(t)|s(0) = s]\,dt$$

$$= \sum_{i=0}^{\infty} e^{-\alpha t_i}\int_0^{\tau_i} e^{-\alpha t}\mathbb{E}_w[s^T(t)Qs(t) + {u_i}^T R u_i + \beta|s(0) = s_i]\,dt$$

コスト関数$r(s_i, u_i, \tau_i)$

- Q関数のベルマン方程式

$$Q^{\pi}(s, u, \tau) = r(s, u, \tau) + e^{-\alpha\tau}\mathbb{E}_{s'}[Q^{\pi}(s'(s, u, \tau), \pi(s'(s, u, \tau)))]$$

  - 次ステップ以降の価値にかかる割引率がτによって変動する
  - 決定的方策勾配定理が使えない

本研究での主結果

- 最適セルフトリガー制御問題に対する決定的方策勾配

$$\nabla_\theta J(\pi_\theta)$$

$$= \sum_{i=0}^{\infty} \int_S \cdots \int_S d_0(s_0) \Pr(s_0 \longrightarrow s_1, 1, \pi_\theta) \cdots \Pr(s_i \longrightarrow s_{i+1}, 1, \pi_\theta)$$

$$e^{-\alpha t_i} \left\{ \nabla_\theta \pi_\theta(s_i) \nabla_a Q^{\pi_\theta}(s_i, a)|_{a=\pi_\theta(a_i)} + \nabla_\theta e^{-\alpha \tau(s_i)} V^{\pi_\theta}(s_{i+1}) \right\} ds_{i+1} ds_i \cdots ds_0$$

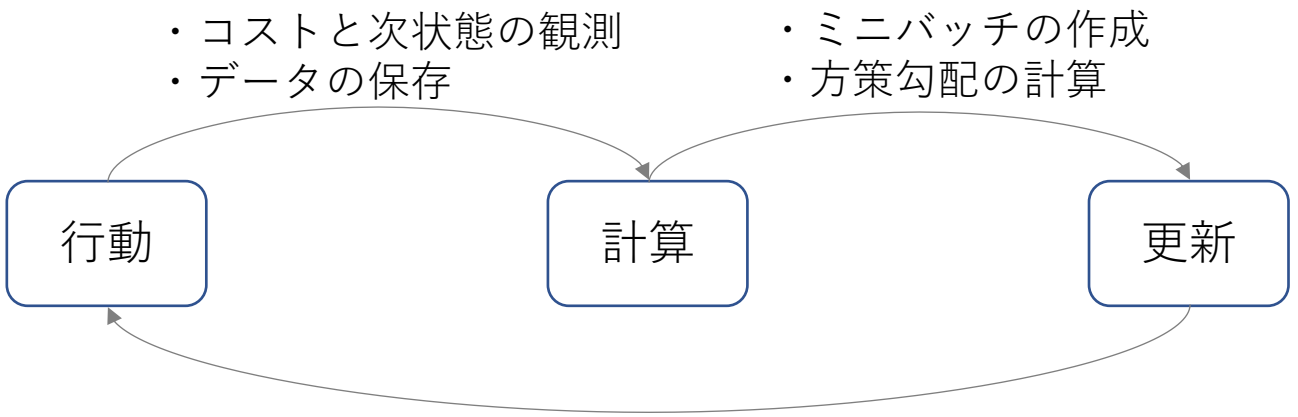
- $\Pr(s \longrightarrow s', i, \pi)$: 状態$s$から方策$\pi$で$i$ステップ制御して$s'$にいる確率
- 方策π_θでの制御パス$\{s_0, s_1, \cdots\}$に対する青文字部の和の期待値
  - 方策更新毎に、何本ものパスをシミュレーションするのは非現実的

方策勾配の近似計算に対する提案手法

- データ収集を行いながら, 少しづつ方策を更新
  - T秒の制御ごとに初期点を変えて制御
  - 各ステップで, メモリにデータ組$\{s, u, \tau, r, s', t\}$を保存し,古いデータを捨てる
  - 方策更新が小さければ, メモリ内は近い方策による複数の制御パス
  - 各データ組を確率$e^{-\alpha t}$の重み付きで選んでミニバッチEを作成
  - ミニバッチEの各データに対する

$$\nabla_\theta \pi_\theta(s) \nabla_a Q^{\pi_\theta}(s, u, \tau)|_{(u,\tau)=\pi_\theta(s)} + \nabla_\theta e^{-\alpha \tau_\theta(s)} Q^{\pi_\theta}(s', \pi_\theta(s))$$

を平均して, 近似方策勾配として用いる



計算に用いるQ関数

- 近似方策勾配の計算には$Q^{\pi_\theta}(s,u,\tau)$を用いる

$$\nabla_\theta \pi_\theta(s) \nabla_a Q^{\pi_\theta}(s,u,\tau)|_{(u,\tau)=\pi_\theta(s)} + \nabla_\theta e^{-\alpha\tau_\theta(s)} Q^{\pi_\theta}(s', \pi_\theta(s))$$

- 真の$Q^{\pi_\theta}(s,u,\tau)$は未知なので, 関数$Q(s,u,\tau|\omega)$を用いて近似
  - ミニバッチEに対するTD誤差のMSEをωに関して最小化する

$$\frac{\partial}{\partial \omega} \frac{1}{N} \sum_{(s,u,\tau)\in E} \left(Q(s,u,\tau|\omega) - \{ r(s,u,\tau) + \gamma Q(s', \pi(s')|\omega) \} \right)^2$$

TD 誤差: 真の Q はこれを0にする

数値実験 (線形システム)

- 制御対象

$$\dot{s} = \begin{bmatrix} -1 & 4 \\ 2 & -3 \end{bmatrix} s + \begin{bmatrix} 2 \\ 4 \end{bmatrix} u + \begin{bmatrix} 0.6 \\ 0.3 \end{bmatrix} \dot{w} \qquad \dot{w}: \text{ウィーナー過程による雑音}$$

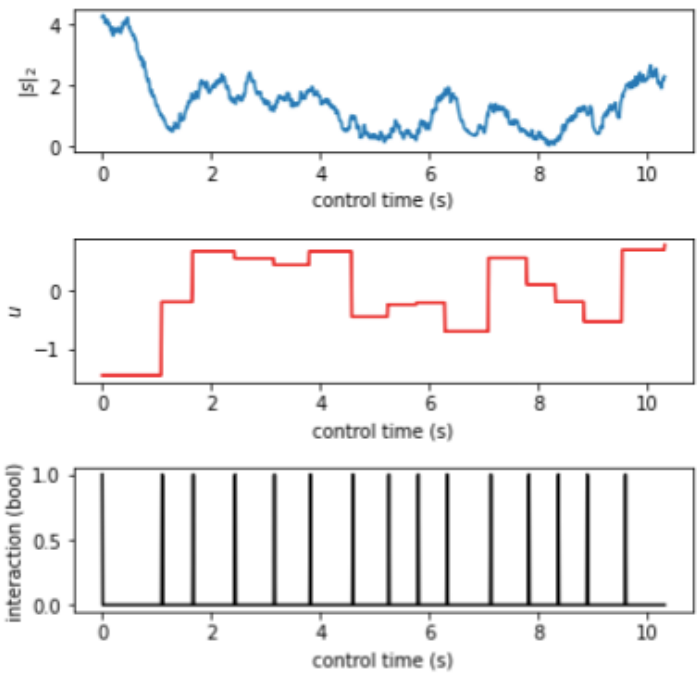
- 初期方策
  - 素朴に設計した方策

$$\pi(s) = \underset{u,\tau}{\operatorname{argmin}} \{u^2 - \lambda\tau + V^*_{cont}(s'_e, \Sigma)\} \qquad \begin{aligned} s'_e &= \mathbb{E}_w[s'(s,u,\tau)] \\ \Sigma &= Var[s'(s,u,\tau)] \end{aligned}$$

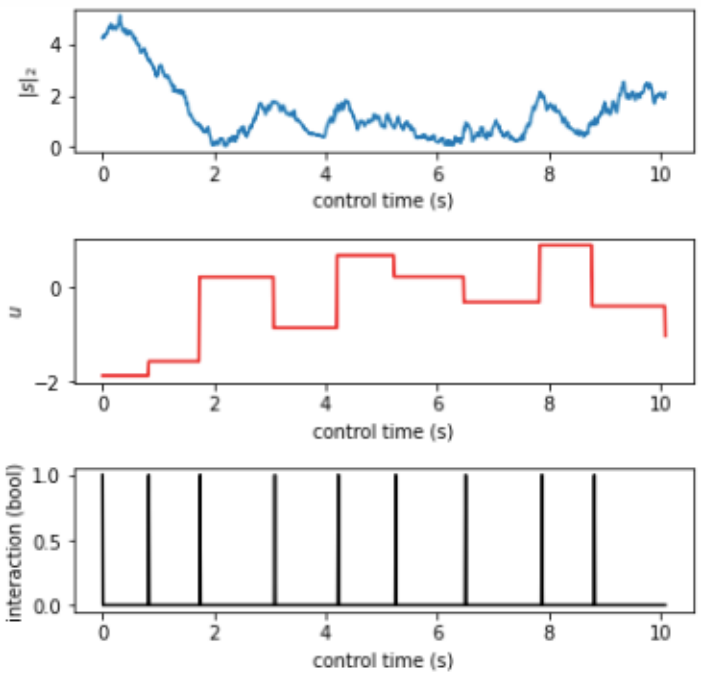
  - $V^*_{cont}(s'_e, \Sigma)$は, 連続的に最適制御した際の制御コスト(×通信コスト)
  - 次ステップで高い制御コストを必要とする状態に行かないようにしたい

数値実験の結果 (線形システム)

- 初期方策(左)と, 学習で得た方策(右)の制御性能比較
  - 初期値$s_0 = [3., 3.]$からの制御
  - 上から, 状態変数の2ノルム, 各時刻の入力u, 通信の有無を表す真偽値



$J = 11.3$



$J = 6.9$

状態変化を抑えながら, 通信回数の減少

数値実験 (非線形システム)

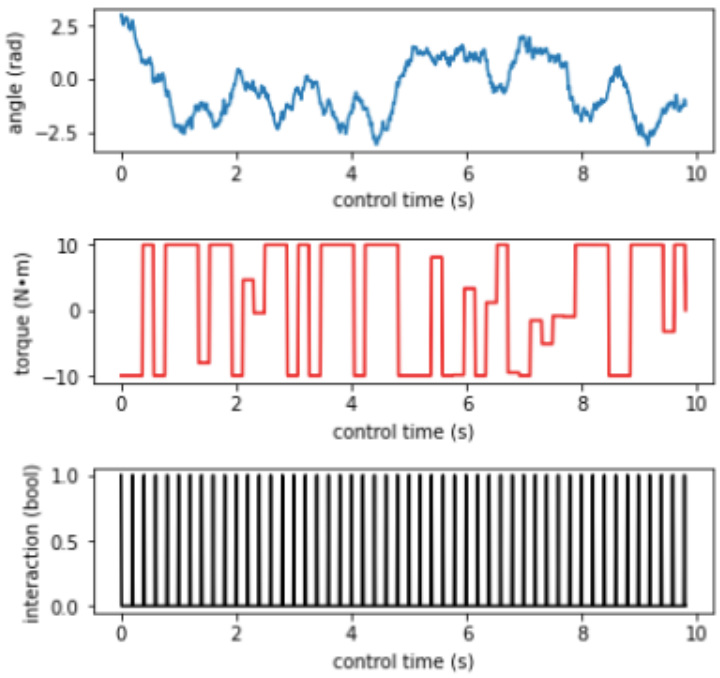
- 制御対象: 倒立振子

$$\frac{d}{dt}\begin{bmatrix}\varphi \\ \dot{\varphi}\end{bmatrix} = \begin{bmatrix}\dot{\varphi} \\ \frac{3g}{2g}sin\varphi + \frac{3}{ml^2}u\end{bmatrix} + \begin{bmatrix}1 \\ 1\end{bmatrix}\dot{w} \qquad \dot{w}:\text{ウィーナー過程による雑音}$$

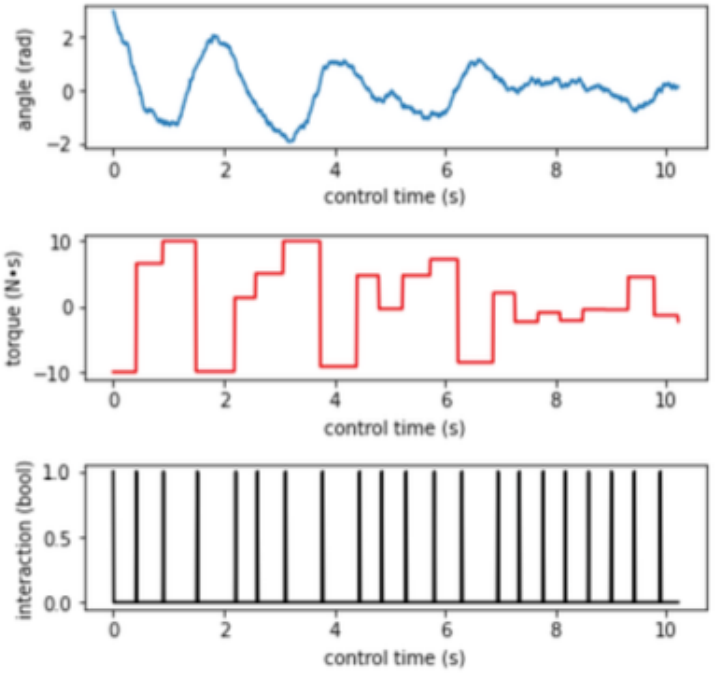
- 初期方策
  - サンプル値制御$(\tau = 0.2)$
  - 各時刻の制御入力は原点付近で線形化したシステムの連続時間LQRによって設計

数値実験の結果 (線形システム)

- 初期方策(左)と, 学習で得た方策(右)の制御性能比較
  - 初期値$s_0 = [3., 3.]$からの制御
  - 上から, 角度φ rad, 各時刻のトルクu N·m, 通信の有無を表す真偽値



$J = 62.5$



$J = 30.6$

状態変化を抑えながら, 通信回数が減少

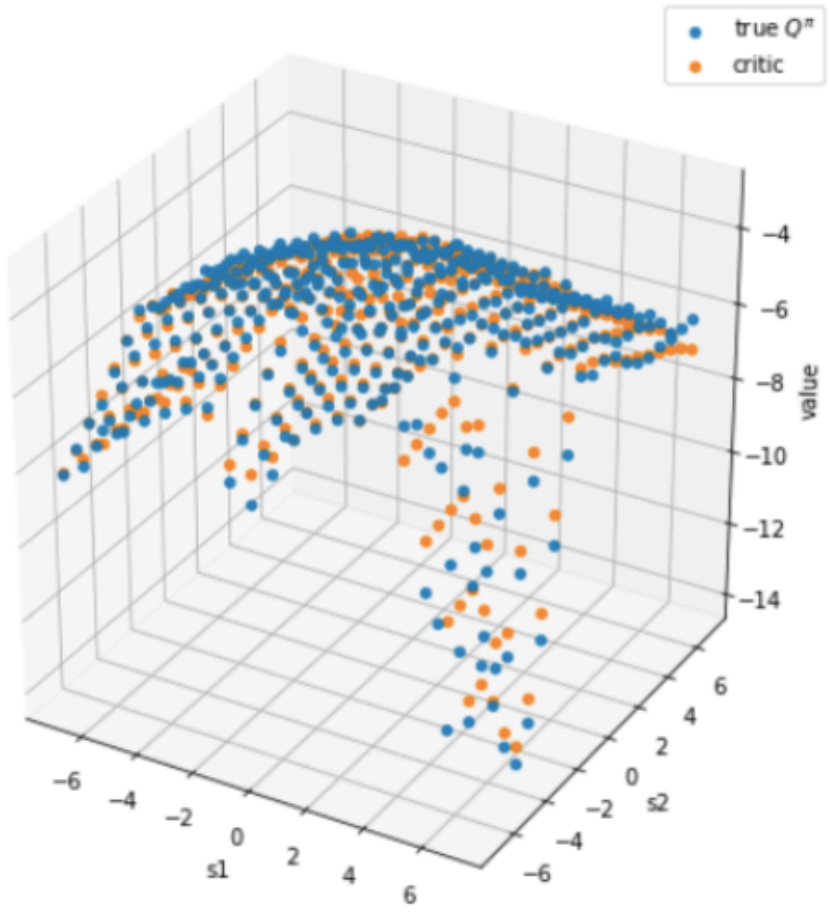
結論

- 先行研究で考慮されていなかった, 長時間制御全体での通信コストを陽に組み込んだ最適セルフトリガー制御問題の定式化

- 定式化した問題の方策勾配型強化学習を用いた解き方の考案

付録A: Q関数の近似

- 格子状に状態変数をとってTD学習



付録B: 用いたニューラルネットワーク

付録C: ハイパーパラメータ