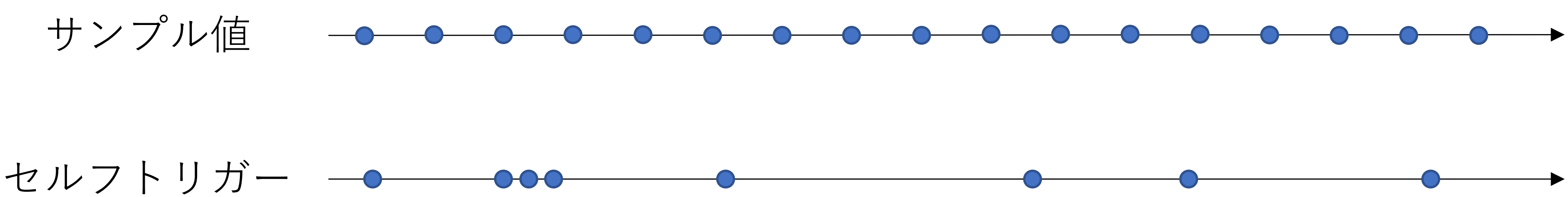
セルフトリガー制御に対する深層強化学習

数理工学専攻 制御システム論分野
竹内 維吹

イントロダクション: 連続時間制御の手法

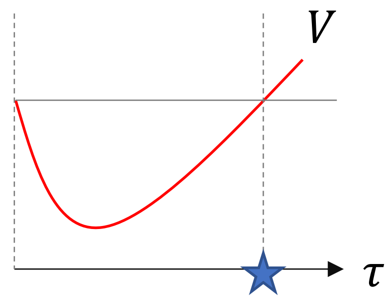
- サンプル値制御 (従来法)
  - 連続時間システムを一定時間間隔で通信し, 制御する手法
  - 各通信の間は, 同じ入力を加え続ける
  - 制御入力の変更が小さい場合は非効率な通信を行うことになる

- 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
  - [1]は連続時間システムに対して



$$\min_{u} u^T u$$
$$\mathrm{s.t.}\ \mathcal{L}_f V(s) + \mathcal{L}_g V(s) u + \varepsilon V(s) \ \le 0$$

の解uを加え続けた際, 次ステップでのリアプノフ関数Vが減少する, 最大の通信間隔をτとする手法を提案

- 長時間制御全体の通信コストの最適性は考慮していない
  - 陽に考慮した最適化問題を考える(本研究)

[1]: G. Yang et al., "Self-triggered Control for Safety Critical Systems Using Control Barrier Functions.", In *Proc. of ACC*, 2019.

最適セルフトリガー制御問題の定式化

- 最適化問題

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

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

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

  - α,β,E,F: ハイパーパラメータ
  - $\pi(s)$: セルフトリガー制御則, 入力uと通信間隔τを出力
  - $\delta_c(t)$: $\int_0^{\infty} e^{-\alpha t}\beta\delta_c(t)dt=\sum_{t\in C} e^{-\alpha t}\beta$ （Cは通信した時刻の集合）

    （↑通信した時刻の$e^{-\alpha t_i}\beta$が積分に足されていくイメージ）

- 本研究では, この問題を強化学習で解けるのかを検討

準備: 強化学習とは

- 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

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

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

i: ステップ数

環境雑音

  - $V^{\pi}(s)$: 全ステップ, 方策πで制御した際の累積コスト
  - $r(s,a)$: 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

$$Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V^{\pi}(s'(s,a))]$$ (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をもつ関数(DNN等)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[2]
  - 方策が状態sから行動aへの関数の場合の勾配

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

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

[2]: D. Silver et al., "Deterministic Policy Gradient Algorithms", In *Proc. of ICML*, 2014.

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, $V^{\pi}(s)$を分解

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

$$= \sum_{i=0}^{\infty} e^{-\alpha t_i} \underbrace{\int_0^{\tau_i} e^{-\alpha t} \mathbb{E}_w[s^T(t)Qs(t) + u_i{}^T R u_i + \beta | s(0) = s_i, \pi]\, dt}_{\text{コスト関数} r(s_i, u_i, \tau_i)}$$

- 本問題におけるベルマン方程式

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

  - 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

$$\nabla_\theta J(\pi_\theta)$$
$$= \sum_{i=0}^{\infty} \int_S \cdots \int_S d_0(s_0) \Pr(s_0 \longrightarrow s_1, 1, \pi_\theta) \cdots \Pr(s_i \longrightarrow s_{i+1}, 1, \pi_\theta)$$
$$e^{-\alpha t_i} \left\{ \nabla_\theta \pi_\theta(s_i) \nabla_a Q^{\pi_\theta}(s_i, a)|_{a=\pi_\theta(s_i)} + \nabla_\theta e^{-\alpha \tau(s_i)} V^{\pi_\theta}(s_{i+1}) \right\} ds_{i+1} ds_i \cdots ds_0$$

  - $\Pr(s \longrightarrow s', i, \pi)$: 状態$s$から方策$\pi$で$i$ステップ制御して$s'$にいる確率

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

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

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

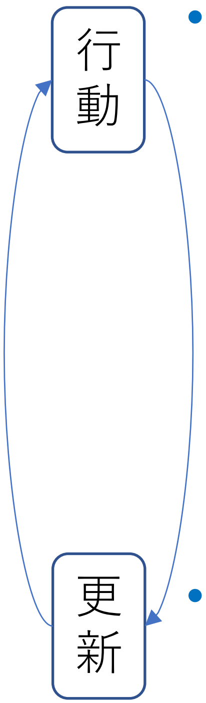
$$e^{-\alpha t_i} \left\{ \nabla_\theta \pi_\theta(s_i) \nabla_a Q^{\pi_\theta}(s_i, a)|_{a=\pi_\theta(s_i)} + \nabla_\theta e^{-\alpha \tau(s_i)} V^{\pi_\theta}(s_{i+1}) \right\} ds_{i+1} ds_i \cdots ds_0$$

  - $\Pr(s \longrightarrow s', i, \pi)$: 状態$s$から方策$\pi$で$i$ステップ制御して$s'$にいる確率

- 右辺は, 制御パス毎に計算される$G = \sum_{i=0}^{\infty} e^{-\alpha t_i} \{\sim\}$の期待値
  - 無限個の初期状態$s_0 \sim d_0$に対し, 各s_0から無限本の制御パスをとり, それぞれのパスで計算したGを平均すればいい
  - 方策更新ごとにデータを取るのは非現実的

方策勾配の近似計算に対する提案手法

- データ収集と方策更新を繰り返す



  - 1. 行動し, 経験データをメモリに貯める

理想:
今の方策π_θでの
無限個の制御パス

  - 2. メモリのデータで方策勾配を近似して方策更新

方策勾配の近似計算に対する提案手法

行動

更新

- データ収集と方策更新を繰り返す
  - 1. 行動し, 経験データをメモリに貯める
    - 様々なs_0からのパスが必要 → T秒ごとに新しいs_0から制御(エピソード)
    - 状態s_iで行動$\pi(s_i)$をとり, コストr_iと次状態s_{i+1}を観測
    - データ組$\{s_i, \pi(s_i), r_i, s_{i+1}, t_i\}$を保存し, 古いデータを捨てる
      - t_i: 各エピソードでの経過時間
    - 方策更新が小さい→ メモリ内には似た方策による複数の制御パス
  - 2. メモリのデータで方策勾配を近似して方策更新

方策勾配の近似計算に対する提案手法

- データ収集と方策更新を繰り返す

行動

- 1. 行動し, 経験データをメモリに貯める
  - 様々なs_0からのパスが必要 → T秒ごとに新しいs_0から制御(エピソード)
  - 状態s_iで行動$\pi(s_i)$をとり, コストr_iと次状態s_{i+1}を観測
  - データ組$\{s_i, \pi(s_i), r_i, s_{i+1}, t_i\}$を保存し, 古いデータを捨てる
    - t_i: 各エピソードでの経過時間
  - 方策更新が小さい→ メモリ内には似た方策による複数の制御パス

更新

- 2. メモリのデータで方策勾配を近似して方策更新
  - 各データ組を確率$e^{-\alpha t}$の重み付きでN個選んでデータセットEを作成
  - データセットEの各データに対して, 以下を平均

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

$G = \sum_{i=0}^{\infty} e^{-\alpha t_i}\{\sim\}$を表現

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

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

パラメータ

- 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
    - データセットEに対するTD誤差のMSEをωに関して最小化する

$$\frac{\partial}{\partial\omega}\frac{1}{N}\sum_{(s,u,\tau)\in E}\Big(\underbrace{Q(s,u,\tau|\omega) - \{r(s,u,\tau) + e^{-\alpha\tau}Q(s',\pi(s')|\omega)\}}_{\text{TD 誤差: 真の } Q^{\pi_\theta} \text{ はこれを0にする}}\Big)^2$$

数値実験

- 制御対象: 倒立振子(非線形システム)

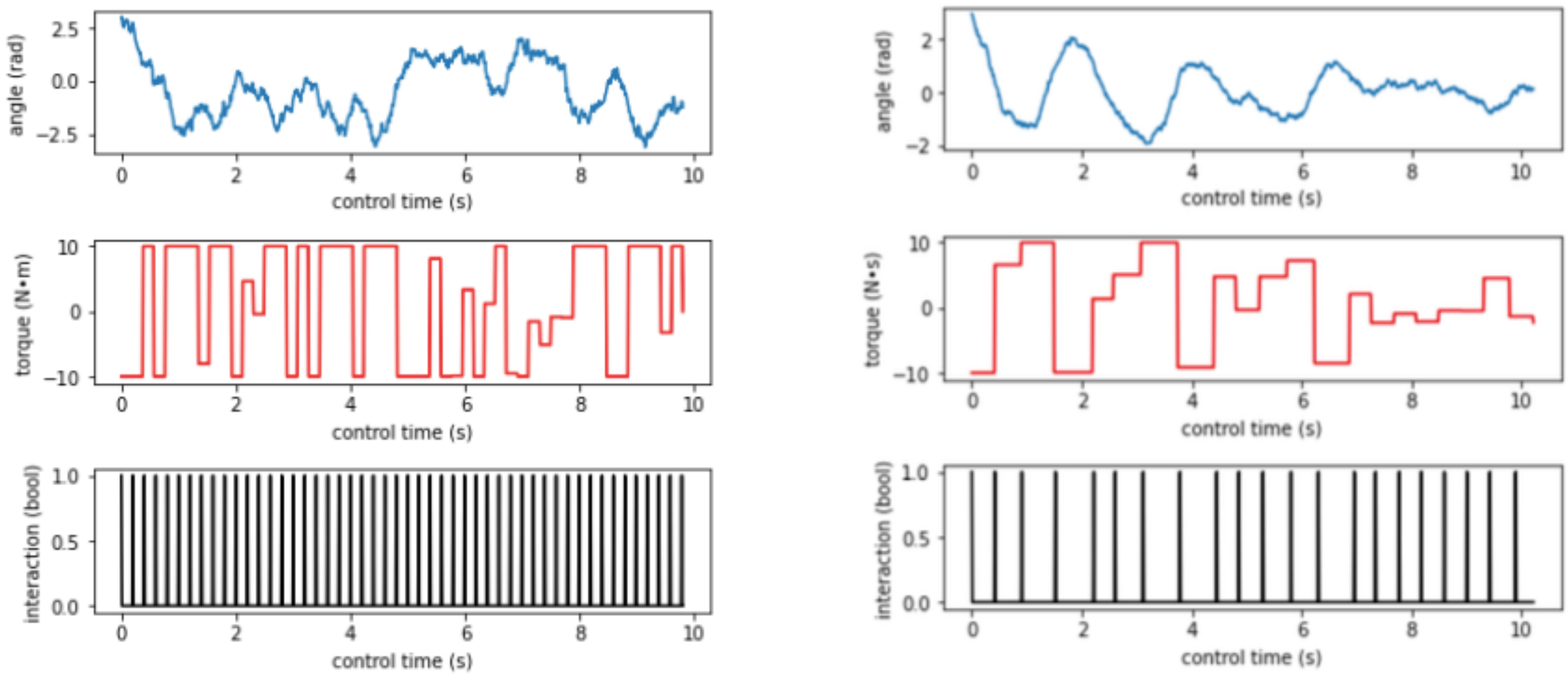
$$\frac{d}{dt}\begin{bmatrix}\varphi\\ \dot{\varphi}\end{bmatrix} = \begin{bmatrix}\dot{\varphi}\\ \frac{3g}{2l}sin\varphi + \frac{3}{ml^2}u\end{bmatrix} + \begin{bmatrix}1\\1\end{bmatrix}\dot{w}$$

$\dot{w}$: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

- 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

- 長時間制御全体での通信コストを考慮した最適セルフトリガー制御問題の定式化

- 定式化した問題に対する方策勾配の式の導出
  - 及び, それを用いた強化学習法の提案

- 非線形システムに対する有効性の確認
  - 線形システムに対しても同様の結果

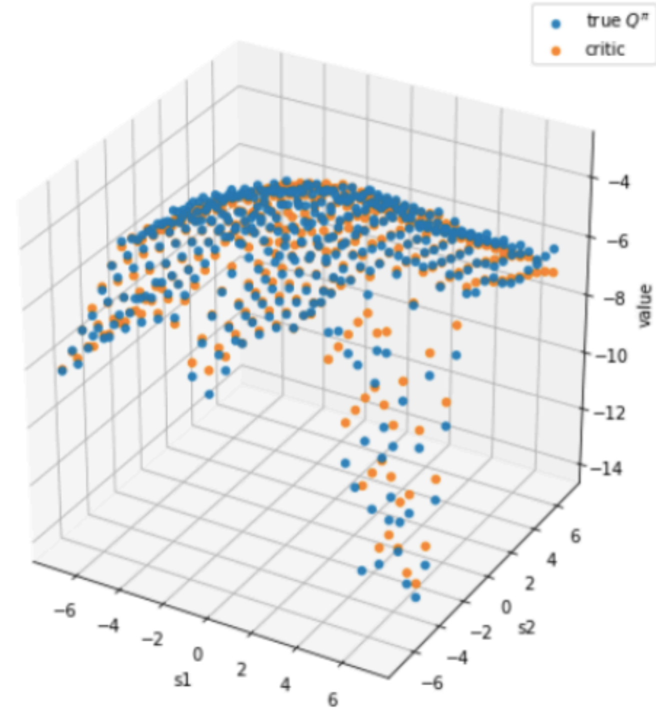
付録A: Q関数の近似

- 格子状に状態変数をとってTD学習

- 相対誤差の平均
  - $V^{\pi_\theta}(s)$と$Q(s,\pi_\theta(s)|\omega)$を比較

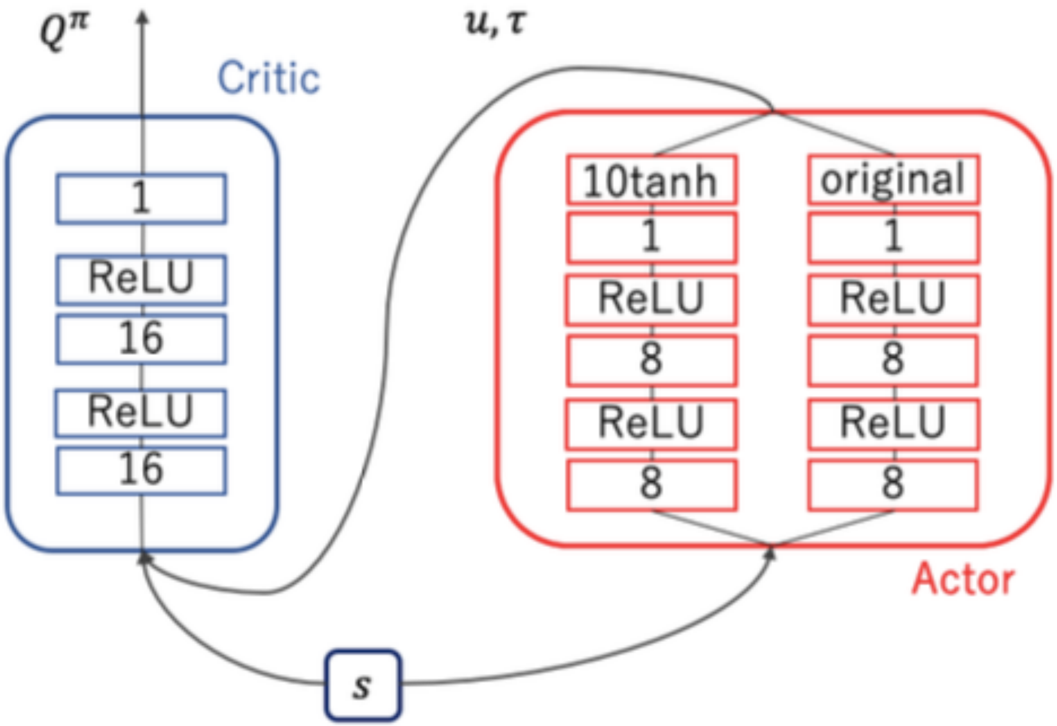
$$\frac{|V^{\pi_\theta}(s) - Q(s,\pi_\theta(s)|\omega)|}{|V^{\pi_\theta}(s)|}$$

  - 平均相対誤差 $= 0.03$



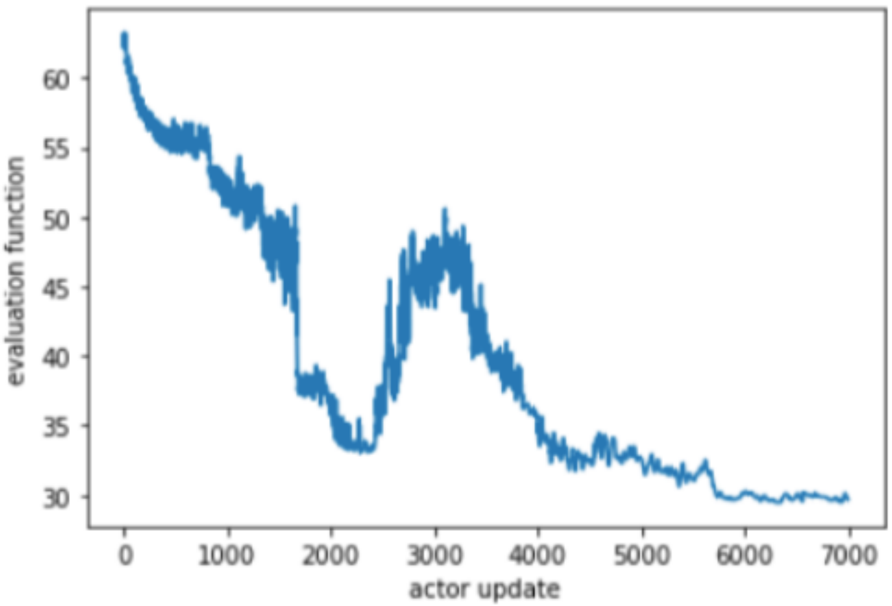
付録B: モデルの設定

- ニューラルネットワーク



付録C: 方策更新に伴う評価関数の履歴

- うまく学習ができた例



- TD学習に求められる精度は非常に高く、その近似精度が低いと、方策が悪化することもよくあった

付録D: 数値実験 (線形システム)

- 制御対象

$$\dot{s} = \begin{bmatrix} -1 & 4 \\ 2 & -3 \end{bmatrix} s + \begin{bmatrix} 2 \\ 4 \end{bmatrix} u + \begin{bmatrix} 0.6 \\ 0.3 \end{bmatrix} \dot{w}$$

$\dot{w}$: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

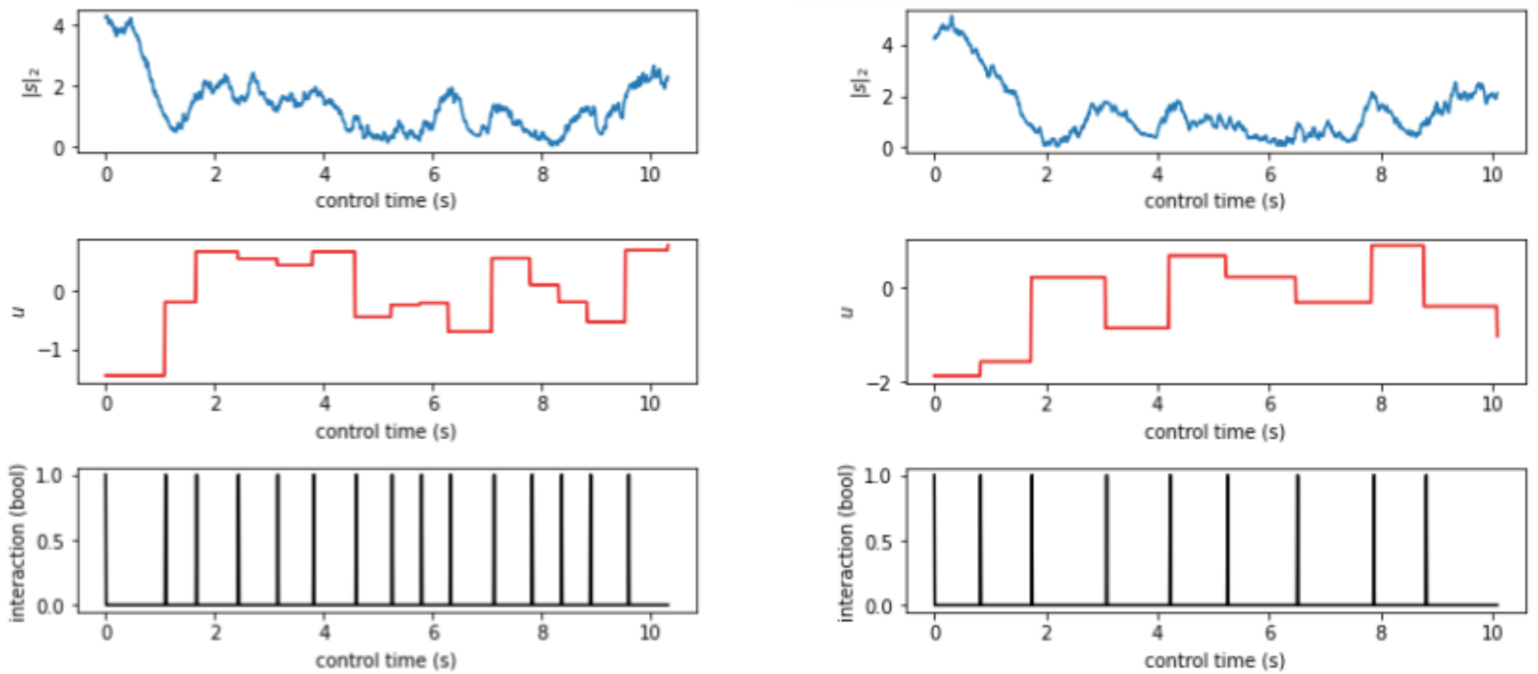
$$\pi(s) = \underset{u,\tau}{\mathrm{argmin}} \{u^2 - \lambda\tau + V_{cont}^*(s_e', \Sigma)\}$$

$$s_e' = \mathbb{E}_w[s'(s,u,\tau)]$$
$$\Sigma = Var[s'(s,u,\tau)]$$

  - $V_{cont}^*(s_e', \Sigma)$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

付録D: 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