Thompson-sampling based reinforcement learning for networked control of unknown linear systems

Borna Sayedana\textsuperscript{1}
Joint work with: Mohammad Afshari\textsuperscript{2}, Peter E. Caines\textsuperscript{1}, Aditya Mahajan\textsuperscript{1}

McGill University\textsuperscript{1}, CIM\textsuperscript{1}, GERAD\textsuperscript{1}, Georgia Institute of Technology\textsuperscript{2}

IEEE Conference on Decision and Control
December 2022
The control loops are closed through *wireless channel*.

These channels can be between plant and controller / sensors and controller.

Applications: Platooning of self-driving trucks, Smart grid, Robotics, Wireless sensor networks

**Question**: How can we control *unknown* NCS?
In all these works, switching signal is known or controlled.
Notation

- $J(\theta)$: performance of the optimal policy for parameter $\theta$.
- Given the prior over $\theta \in \Theta$, the Bayesian regret:

$$
\mathcal{R}(T; \pi) = \mathbb{E}^\pi \left[ \sum_{t=1}^{T} c(x_t, u_t, \nu_t) - TJ(\theta) \right]
$$

- $a_n = \mathcal{O}(b_n)$, if there exists a positive constant $K$, such that:

$$
\|a_n\| \leq K b_n
$$

- $\tilde{\mathcal{O}}(c_n)$ means:

$$
\tilde{\mathcal{O}}(c_n) = \mathcal{O}(c_n \log^k(n))
$$
Bayesian Reinforcement learning for linear systems

Regret bounds for Linear systems

- [Abbasi-Yadkori and Szepesvari, 2014, Faradonbeh et al., 2020b, Faradonbeh et al., 2020a, Simchowitz and Foster, 2020]

Thompson sampling

- [Gagrani et al., 2021, Ouyang et al., 2020] for LQR problem:

\[ \mathcal{R}(T; \text{TSDE}) \leq \tilde{O}(\sigma_w^2 (n + m) \sqrt{nT}). \]
Contribution

- Bayesian reinforcement learning for networked control system
- Variation of TSDE algorithm [Ouyang et al., 2020]
- Connection with Markov jump linear systems.
- Achieve Bayesian regret bound of:
  \[
  R(T; \text{TSDE}) \leq \tilde{O}(\sigma_w^2(n + m)\sqrt{nT}).
  \]
- Show same regret bound is true for the NCS model.
Networked Control Systems

System’s dynamics

\[ x_{t+1} = Ax_t + \nu_t Bu_t + w_t, \quad t \geq 1, \]

- \( \{w_t\}_{t \geq 1} \): is an i.i.d. Gaussian process with \( w_t \sim \mathcal{N}(0, \sigma_w^2 I) \).
- \( \{\nu_t\}_{t \geq 1} \): is an i.i.d. Bernoulli process with \( \mathbb{P}(\nu_t = 1) = q \).

Switching per step cost

- Per-step cost given by

\[ c(x_t, u_t, \nu_t) = x_t^\top Q x_t + \nu_t u_t^\top R u_t, \quad Q \succ 0, \ R \succ 0 \]
Optimization Setup

- $\theta^\top = [A, B]$ : parameters of the system.
- $q$ : probability of successful transmission
- Performance of policy $\pi$:

$$
J(\pi; \theta) = \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi} \left[ \sum_{t=1}^{T} c(x_t, u_t, \nu_t) \right]
$$
Planning Solution

- Planning problem: [Sinopoli et al., 2005]: \( J(\theta) = \sigma_w^2 \text{tr}(S_\theta) \)
- \( s(\theta) \succ 0 \) solution to modified Riccati:

\[
S(\theta) = Q + A^T S(\theta) A - q A^T S(\theta) B (R + B^T S(\theta) B)^{-1} B^T S(\theta) A
\]

Optimal policy

- Optimal control action:

\[
u_t = G(\theta) x_t,
\]

- Gain:

\[
G(\theta) = -(R + B^T S(\theta) B)^{-1} B^T S(\theta) A
\]
NCS as a switching system

Switching between open/closed loop dynamics

- If $\nu_t = 1$ : closed loop dynamics
- If $\nu_t = 0$ : open loop dynamics
Problem Formulation

Our setup

- $\theta = (A, B)$ are unknown.
- $(q, Q, R)$ are known.
- We have a prior on $\theta \in \Theta$.

Definition of regret:

$$\mathcal{R}(T; \pi) = \mathbb{E}^\pi \left[ \sum_{t=1}^{T} c(x_t, u_t, \nu_t) - TJ(\theta) \right]$$
Assumptions on the Model

- **Controllability**: \( \forall \theta \in \Theta, \) pair \((A_\theta, B_\theta)\) is controllable.

- **Sufficient condition for planning** [Sinopoli et al., 2005]:
  \[
  1 - q \leq \frac{1}{|\lambda_{\text{max}}(A_\theta)|^2}, \quad \forall \theta \in \Theta
  \]

  \( \lambda_{\text{max}}(A_\theta) \): Maximum eigen-value of \( A_\theta \)

  \[
  \delta := \sup_{\theta, \phi \in \Theta} \| A_\theta + B_\theta G(\phi) \|, \quad \sigma := \sup_{\theta \in \Theta} \| A_\theta \|.
  \]

Assumption on the stability

- **Stability**: \( \delta^q \sigma^{1-q} < 1. \)

- Average contractivity of dynamical system
Assumption on the prior

Distribution of prior:

\[ p_1(\theta) = \left[ \prod_{i=1}^{n} \xi_1^i(\theta^i) \right]_{\Theta} \]

\[ \xi_1^i = \mathcal{N}(\mu_1^i, \Sigma_1), \mu_1^i \in \mathbb{R}^d. \]

Given the assumption we get posterior, [Sternby, 1977]:

\[ p_t(\theta) = \left[ \prod_{i=1}^{n} \xi_t^i(\theta^i) \right]_{\Theta} \]

\[ \xi_t^i(\theta^i) = \mathcal{N}(\mu_t^i, \Sigma_t) \]
Update rule for $\{\mu_i^t\}_{i=1}^n$ and $\Sigma_t$:

$$\mu_{t+1}^i = \mu_t^i + \frac{\Sigma_t z_t (x_{t+1}^i - (\mu_t^i)^T z_t)}{\sigma_w^2 + z_t^T \Sigma_t z_t},$$

$$\Sigma_{t+1}^{-1} = \Sigma_t^{-1} + \frac{1}{\sigma_w^2} z_t z_t^T,$$

where $z_t = \text{vec}(x_t, \nu_t u_t)$, and $x_t = [x_1^t, \ldots, x_n^t]$. 
We present a variation of TSDE for NCS.

\( t_k \): start of episode, \( T_k \): length of episode

### Episodes restarts

\[
 t_{k+1} = \min \left\{ t > t_k \left| \begin{array}{l}
 t - t_k > T_{k-1} \text{ or } \\
 \det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k}
\end{array} \right. \right\}.
\]
At the episode $K$:

1. $\theta_k$ is sampled from posterior $p_{tk}$.
2. Control inputs are generated using $\theta_k$:

$$u_t = G(\theta_k)x_t, \quad t_k \leq t \leq t_{k+1} - 1$$
Thompson Sampling with Dynamic Episodes

1: **input:** $\Theta, \hat{\theta}, \Sigma_1$
2: **initialization:** $t \leftarrow 1, t_0 \leftarrow -T_{\text{min}}, T_{-1} \leftarrow T_{\text{min}}, k \leftarrow 0.$
3: **for** $t = 1, 2, \ldots$ **do**
4:   observe $x_t$
5:   update $p_t$.
6:   **if** $( (t - t_k > T_{k-1}) \text{ or } (\det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k}) )$ **then**
7:     $T_k \leftarrow t - t_k, k \leftarrow k + 1, t_k \leftarrow t$
8:     sample $\theta_k \sim \mu_t$
9:   **end if**
10:  Apply control $u_t = G(\theta_k)x_t$
11: **end for**
Main results

**Theorem**

The regret of TSDE is upper bounded by

$$\mathcal{R}(T; \text{TSDE}) \leq \tilde{O}(\sigma_w^2(n + m)\sqrt{nT}).$$

- $n$ is dimension of state
- $m$ is dimension of control input
Feasible region for planning: \( Q_p(\Theta) = [q_p, 1] \)

\[
q_p = \sup_{\theta \in \Theta} \left[ 1 - \frac{1}{|\lambda_{\text{max}}(A_{\theta})|^2} \right]^+ ,
\]

Feasible region for learning: \( Q_\ell(\Theta) = \{ q \in [0, 1] : \delta q \sigma^{1-q} < 1 \} \)

**Relation between \( Q_p(\Theta) \) and \( Q_\ell(\Theta) \)**

- Relation between \( Q_p(\Theta) \) and \( Q_\ell(\Theta) \) and is in general a function of \( \Theta \).
- Both \( Q_p(\Theta) \subset Q_\ell(\Theta) \) and \( Q_\ell(\Theta) \subset Q_p(\Theta) \) might hold.
Conclusion

- Bayesian reinforcement learning for Networked control systems
- Use variation of TSDE algorithm and show Bayesian regret of:
  \[ R(T; \text{TSDE}) \leq \tilde{O}(\sigma_w^2 (n + m) \sqrt{nT}) \].
- No partial ordering between \( Q_p(\Theta) \) and \( Q_l(\Theta) \) in general.
- TSDE has the same regret bound as the the case of linear systems.
Thank you!


An LQG optimal linear controller for control systems with packet losses.
In Proceedings of the 44th IEEE Conference on Decision and Control, pages 458–463. IEEE.

On consistency for the method of least squares using martingale theory.