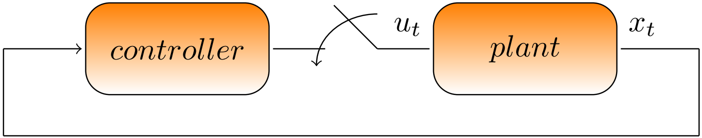
Thompson-sampling based reinforcement learning for networked control of unknown linear systems

Borna Sayedana[1]
Joint work with: Mohammad Afshari[2], Peter E. Caines[1] , Aditya Mahajan[1]

McGill University[1], CIM[1], GERAD[1],Georgia Institute of Technology[2]

IEEE Conference on Decision and Control
December 2022

Network Control Systems



- 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?

Literature review

Planning in NCS

[Sinopoli et al., 2005, Antsaklis and Baillieul, 2007, Sinopoli et al., 2004]

Reinforcement learning for NCS

[Jiang et al., 2017, Fan et al., 2019, Li et al., 2020]

Related Models : Switched Linear Systems

[Sarkar et al., 2019, Shi et al., 2023]
[Sattar et al., 2021, Sayedana et al., 2021]

- In all these works, switching signal is known or controlled.

Notation

- $J(\theta)$: performance of the optimal policy for paramter θ.
- 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:

$$\|a_n\| \leq Kb_n$$

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

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

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:

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

- 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

$$c(x_t, u_t, \nu_t) = x_t^\intercal Q x_t + \nu_t u_t^\intercal R u_t, \quad Q \succ 0, R \succ 0$$

Optimization Setup

- $\theta^\mathsf{T} = [A, B]$: parameters of the system.
- q : probability of successful transmission
- Performance of policy π:

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

Planning Solution

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

$$S(\theta) = Q + A^\intercal S(\theta)A - qA^\intercal S(\theta)B(R + B^\intercal S(\theta)B)^{-1}B^\intercal S(\theta)A$$

Optimal policy

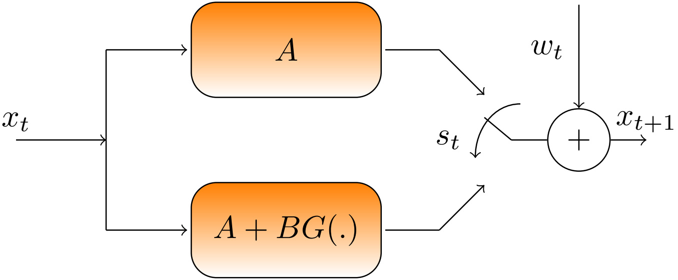
- Optimal control action:

$$u_t = G(\theta)x_t,$$

- Gain:

$$G(\theta) = -(R + B^\intercal S(\theta)B)^{-1}B^\intercal S(\theta)A$$

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_θ, B_θ) is controllable.
- **Sufficient condition for planning** [Sinopoli et al., 2005] :

$$1 - q \leq \frac{1}{|\lambda_{\max}(A_\theta)|^2}, \quad \forall \theta \in \Theta$$

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

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

Assumption on the stability

- **Stability**: $\delta^q \sigma^{1-q} < 1$.
- Average contractivity of dynamical system

Assumption on the prior

Assumption on prior

- Distribution of prior:

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

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

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

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

- $\xi_t^i(\theta^i) = \mathcal{N}(\mu_t^i, \Sigma_t)$

Posterior distribution

Posterior distribution

- Update rule for $\{\mu_t^i\}_{i=1}^n$ and Σ_t:

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

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

- where $z_t = \mathrm{vec}(x_t, \nu_t u_t)$, and $x_t = [x_t^1, \ldots, x_t^n]$

TSDE

- We present a variation of TSDE for NCS.
  t_k :start of episode, T_k:length of episode

Episodes restarts

$$t_{k+1} = \min\left\{ t > t_k \,\middle|\, \begin{array}{l} t - t_k > T_{k-1} \text{ or} \\ \det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k} \end{array} \right\}.$$

TSDE

At the episode K:

1. θ_k is sampled from posterior p_{t_k}.
2. Control inputs are generated using θ_k:

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

TSDE

Thompson Sampling with Dynamic Episodes

input: Θ, $\hat{\theta}$, Σ_1
initialization: $t \leftarrow 1,\ t_0 \leftarrow -T_{\min},\ T_{-1} \leftarrow T_{\min},\ k \leftarrow 0.$
for $t = 1, 2, \ldots$ **do**
    observe x_t
    update p_t.
    if $\big((t - t_k > T_{k-1})$ or $(\det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k})\big)$ **then**
        $T_k \leftarrow t - t_k,\ k \leftarrow k + 1,\ t_k \leftarrow t$
        sample $\theta_k \sim \mu_t$
    end if
    Apply control $u_t = G(\theta_k) x_t$
end for

Main results

Theorem

The regret of TSDE is upper bounded by

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

- n is dimension of state
- m is dimension of control input

Discussion on the Assumptions

- Feasible region for planning : $\mathcal{Q}_p(\Theta) = [q_p, 1]$

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

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

Relation between $\mathcal{Q}_p(\Theta)$ and $\mathcal{Q}_l(\Theta)$

- Relation between $\mathcal{Q}_p(\Theta)$ and $\mathcal{Q}_l(\Theta)$ and is in general a function of Θ.
- Both $\mathcal{Q}_p(\Theta) \subset \mathcal{Q}_l(\Theta)$ and $\mathcal{Q}_l(\Theta) \subset \mathcal{Q}_p(\Theta)$ might hold.

Conclusion

- Bayesian reinforcement learning for Networked control systems
- Use variation of TSDE algorithm and show Bayesian regret of:

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

- No partial ordering between $\mathcal{Q}_p(\Theta)$ and $\mathcal{Q}_l(\Theta)$ in general.
- TSDE has the same regret bound as the the case of linear systems.

Thank you!

Bibliography I

Abbasi-Yadkori, Y. and Szepesvari, C. (2014).
Bayesian optimal control of smoothly parameterized systems: The lazy posterior sampling algorithm.
arXiv preprint arXiv:1406.3926.

Antsaklis, P. J. and Baillieul, J., editors (2007).
: *Specical issue on Technology of Networked Control Systems*, volume 95.

Fan, J., Wu, Q., Jiang, Y., Chai, T., and Lewis, F. L. (2019).
Model-free optimal output regulation for linear discrete-time lossy networked control systems.
IEEE Transactions on Systems, Man, and Cybernetics: Systems, 50(11):4033–4042.

Bibliography II

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. (2020a).
Input perturbations for adaptive control and learning.
Automatica, 117:108950.

Faradonbeh, M. K. S., Tewari, A., and Michailidis, G. (2020b).
On adaptive Linear–Quadratic regulators.
Automatica, 117:108982.

Gagrani, M., Sudhakara, S., Mahajan, A., Nayyar, A., and Ouyang, Y. (2021).
A relaxed technical assumption for posterior sampling-based reinforcement learning for control of unknown linear systems.
arXiv preprint arXiv:2108.08502.

Bibliography III

Jiang, Y., Fan, J., Chai, T., Lewis, F. L., and Li, J. (2017).
Tracking control for linear discrete-time networked control systems with unknown dynamics and dropout.
IEEE transactions on neural networks and learning systems, 29(10):4607–4620.

Li, J., Xiao, Z., Li, P., and Ding, Z. (2020).
Networked controller and observer design of discrete-time systems with inaccurate model parameters.
ISA transactions, 98:75–86.

Ouyang, Y., Gagrani, M., and Jain, R. (2020).
Posterior sampling-based reinforcement learning for control of unknown linear systems.
65(6):3600–3607.

Bibliography IV

Sarkar, T., Rakhlin, A., and Dahleh, M. (2019).
Nonparametric system identification of stochastic switched linear systems.
In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3623–3628. IEEE.

Sattar, Y., Du, Z., Tarzanagh, D. A., Ozay, N., Balzano, L., and Oymak, S. (2021).
Identification and adaptive control of markov jump systems: Sample complexity and regret bounds.
In *ICML Workshop on Reinforcement Learning Theory*.

Sayedana, B., Afshari, M., Caines, P. E., and Mahajan, A. (2021).
Consistency and rate of convergence of switched least squares system identification for autonomous switched linear systems.
arXiv preprint arXiv:2112.10753.

Bibliography V

Shi, S., Mazhar, O., and De Schutter, B. (2023).
Finite-sample analysis of identification of switched linear systems with arbitrary or restricted switching.
IEEE Control Systems Letters, 7:121–126.

Simchowitz, M. and Foster, D. (2020).
Naive exploration is optimal for online lqr.
In *International Conference on Machine Learning*, pages 8937–8948. PMLR.

Sinopoli, B., Schenato, L., Franceschetti, M., Poolla, K., Jordan, M. I., and Sastry, S. S. (2004).
Kalman filtering with intermittent observations.
IEEE transactions on Automatic Control, 49(9):1453–1464.

Bibliography VI

Sinopoli, B., Schenato, L., Franceschetti, M., Poolla, K., and Sastry, S. (2005).
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.

Sternby, J. (1977).
On consistency for the method of least squares using martingale theory.
IEEE Transactions on Automatic Control, 22(3):346–352.