Reinforcement Learning with Fast Stabilization in Linear Dynamical Systems

Sahin Lale

Joint work with: Kamyar Azizzadenesheli, Babak Hassibi and Anima Anandkumar

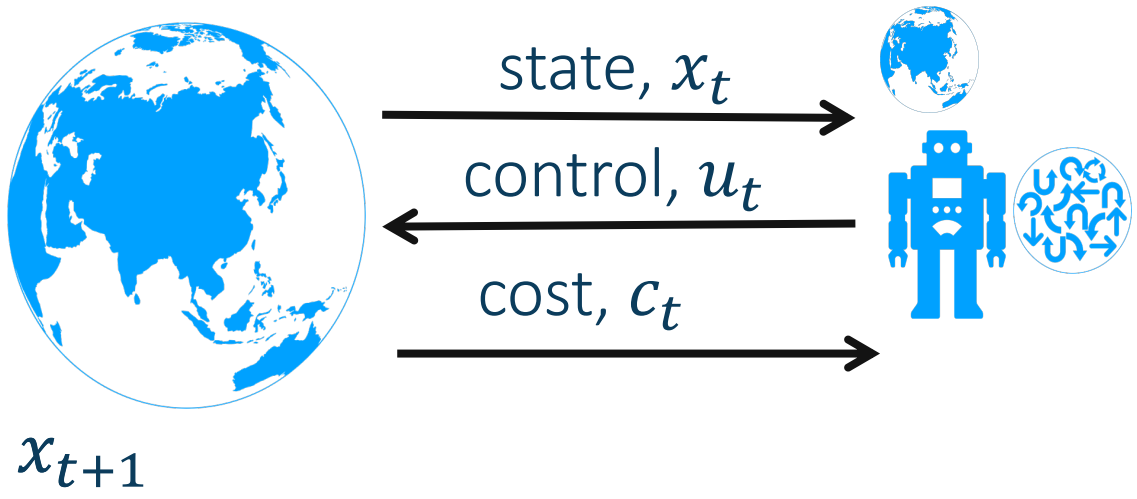
Caltech

Study of Reinforcement Learning & Adaptive Control

- **Central Goal:** To design learning agents that autonomously adapt to unknown environments with minimal information and enjoy finite-time stability and performance guarantees.

- In this work, we study this goal in fully observable linear time-invariant dynamical systems

Linear Quadratic Regulator (LQR)



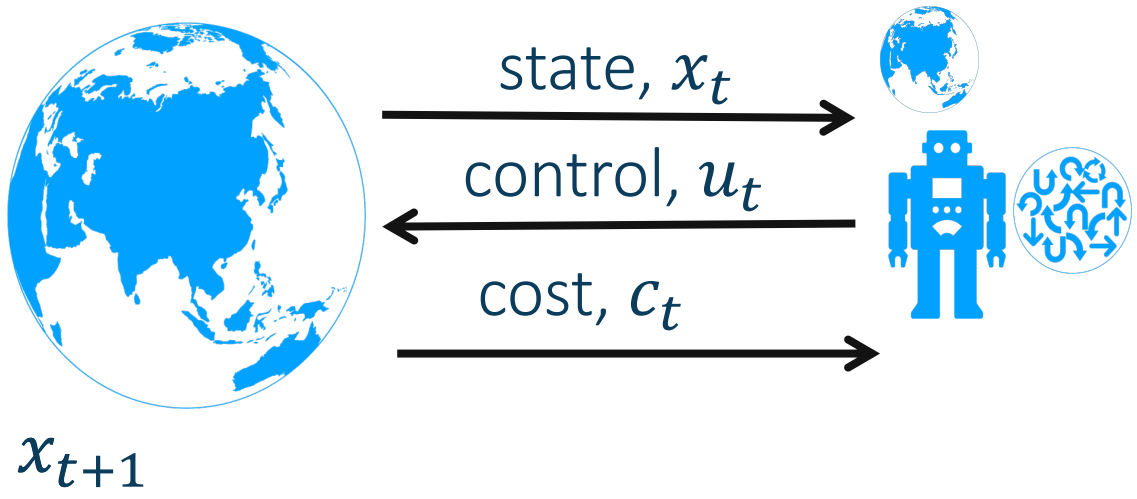
- Linear dynamics
  $$x_{t+1} = A\,x_t + B\,u_t + w_t$$
- Gaussian noise $w_t \sim N(0, \sigma_w^2 I)$
- Stabilizable, i.e. there exists a policy which makes the system asymptotically stable
- Quadratic costs $c_t = x_t^T Q x_t + u_t^T R u_t$
- State feedback (MDP)

Optimal Control (Known model):

$$J_* = \lim_{T\to\infty} \min_{u=[u_1,\ldots,u_T]} \frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} x_t^\top Q x_t + u_t^\top R u_t\right]$$

$$\text{s.t. } x_{t+1} = Ax_t + Bu_t + w_t$$

Linear Quadratic Regulator (LQR)



- **Unknown dynamics (A & B unknown)**

$$x_{t+1} = A\,x_t + B\,u_t + w_t$$

- Gaussian noise $w_t \sim N(0, \sigma_w^2 I)$
- Stabilizable, i.e. there exists a policy which makes the system asymptotically stable
- Quadratic costs $c_t = x_t^T Q x_t + u_t^T R u_t$
- State feedback (MDP)

Optimal Control (Known model):

$$J_* = \lim_{T\to\infty} \min_{u=[u_1,\ldots,u_T]} \frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} x_t^\top Q x_t + u_t^\top R u_t\right]$$

$$\text{s.t. } x_{t+1} = Ax_t + Bu_t + w_t$$

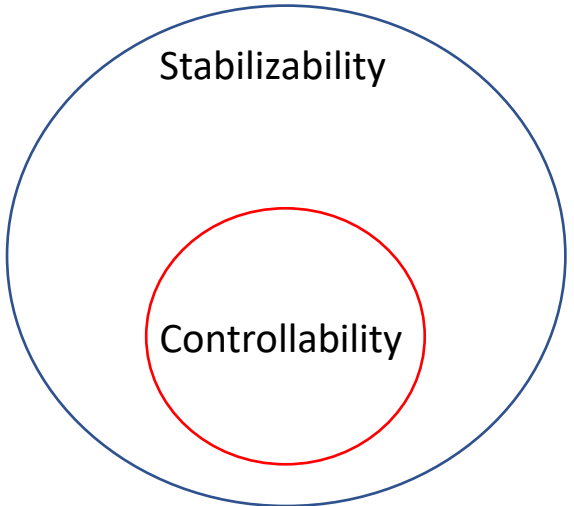
Regret:

$$\sum_{t=1}^{T} c_t - T J_*$$

Prior Works on LQR

Authors	Setting	Regret	Stability
[Abbasi-Yadkori & Szepesvari, 2011]	Controllable	$e^n \sqrt{T}$	No Initial Stabilizing Controller
[Dean et al., 2018]	Controllable	$\mathrm{poly}(n)\, T^{2/3}$	Initial Stabilizing Controller
[Mania et al., 2019]	Controllable	$\mathrm{poly}(n)\sqrt{T}$	Initial Stabilizing Controller
[Simchowitz & Foster, 2020]	Stabilizable	$\mathrm{poly}(n)\, \sqrt{T}$	Initial Stabilizing Controller

- Controllable system can be brought to $x = 0$ in finite steps
- **Restrictive than stabilizability**



Prior Works on LQR

Authors	Setting	Regret	Stability
[Abbasi-Yadkori & Szepesvari, 2011]	Controllable	$e^n \sqrt{T}$	No Initial Stabilizing Controller
[Dean et al., 2018]	Controllable	$\mathrm{poly}(n)\, T^{2/3}$	Initial Stabilizing Controller
[Mania et al., 2019]	Controllable	$\mathrm{poly}(n)\sqrt{T}$	Initial Stabilizing Controller
[Simchowitz & Foster, 2020]	Stabilizable	$\mathrm{poly}(n)\, \sqrt{T}$	Initial Stabilizing Controller
This Work	**Stabilizable**	$\boldsymbol{\mathrm{poly}(n)\, \sqrt{T}}$	**No Initial Stabilizing Controller**

Main Results

- We propose the first RL algorithm, **Stab**ilizing **L**earning (**StabL**), that achieves order-optimal regret $\tilde{\mathcal{O}}(\mathrm{poly}(\mathrm{n})\sqrt{\mathrm{T}})$ in all stabilizable LQRs without a given initial stabilizing policy.

- We show that **StabL** achieves fast stabilization of the system dynamics and attains state-of-the-art regret performance in various adaptive control tasks.

Algorithmic Intuitions

- Sophisticated RL strategies such as optimism fails to explore state-space effectively to design stabilizing policies due to lack of reliable model estimates

- This lack of knowledge also results in state blowups without stability.

- The goal of RL algorithms in dynamical systems should be to **certify fast stabilization by an effective exploration** while still **aiming to control the system**.

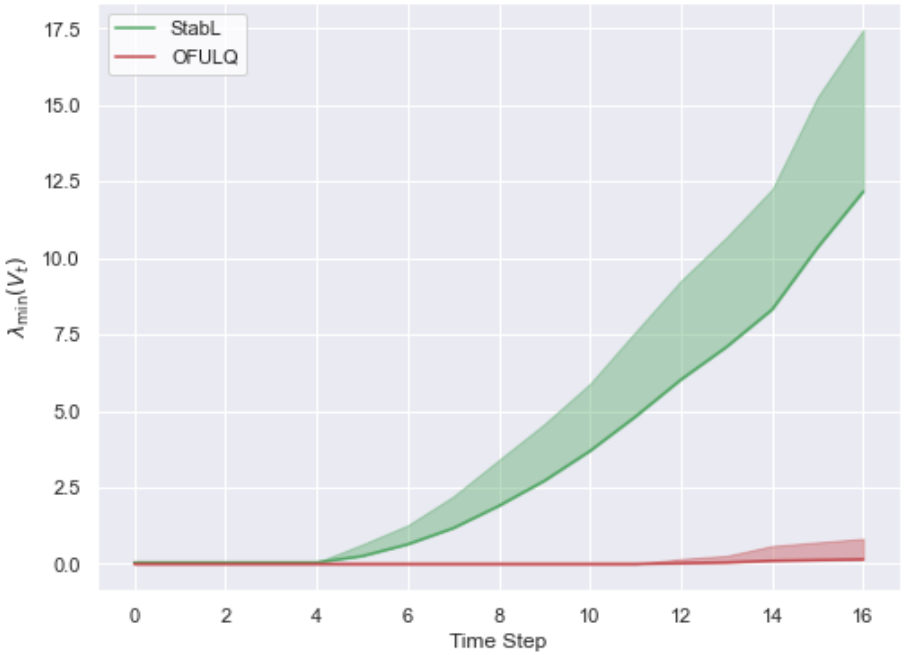
- Frequent policy changes can also cause unstable dynamics in stabilizable systems even if the policies are stabilizing.

Stabilizing Learning Algorithm: StabL

- StabL has two phases:
  1. Adaptive Control with Improved Exploration
  2. Stabilizing Adaptive Control
- In the first phase, StabL uses an improved exploration strategy along with optimism
  - To certify fast stabilization of the system,
  - To minimize the regret.
- Uses isotropic perturbations with optimistic controller uniformly excites the system as well as the dimensions that have more promising impact on the control performance.
- For long enough improved exploration phase, StabL is guaranteed to design of stabilizing controllers henceforth in stabilizing adaptive control phase.
- Avoids frequent controller updates and uses the same controller at least for a fixed time period

Longitudinal flight control of Boeing 747 with linearized dynamics

Minimum Eigenvalue of Design Matrix



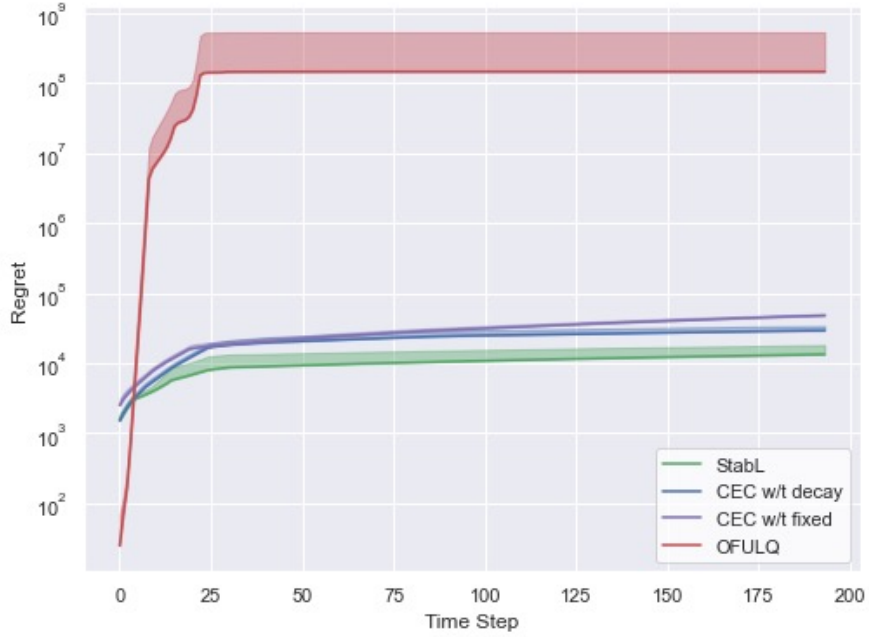
Linearly scaling due to improved exploration which allows finding stabilizing neighborhood!

Maximum State Norm

Algorithm	**Average** $\max \|x\|_2$
StabL	$\mathbf{3.38 \times 10^1}$
OFULQ	1.62×10^3
CEC w/t Fixed	4.97×10^1
CEC w/t Decay	4.60×10^1

Bounded state due to fast stabilization!

Regret



The best regret performance via fast stabilization!

Main Takeaway

- The benefit of early improved exploration to achieve improved control performance in dynamical systems

- At the expense of a slight increase in regret in the early stages with improved exploration, RL algorithms can attain much smaller regret in the long run