Nonlinear Systems Identification

Machine Learning in Feedback System #10

Yahya Sattar
Postdoc with Prof. Sarah Dean

Recap: System Identification

Linear dynamical system with state observations

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{w}_t$$

Recap: System Identification

Linear dynamical system with state observations

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{w}_t$$

- Identify the dynamics model $\boldsymbol{s}_{0:T} \rightarrow \hat{\boldsymbol{F}}$

Recap: System Identification

Linear dynamical system with state observations

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{w}_t$$

- Identify the dynamics model $\boldsymbol{s}_{0:T} \rightarrow \hat{\boldsymbol{F}}$
- State, noise $\boldsymbol{s}_t, \boldsymbol{w}_t \in \mathbb{R}^n$

Recap: System Identification

Linear dynamical system with state observations

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{w}_t$$

- Identify the dynamics model $\boldsymbol{s}_{0:T} \rightarrow \hat{\boldsymbol{F}}$
- State, noise $\boldsymbol{s}_t, \boldsymbol{w}_t \in \mathbb{R}^n$
- System dynamics $\boldsymbol{F} \in \mathbb{R}^{n \times n}$

Recap: System Identification

Least-squares problem with matrix variable

$$\hat{\boldsymbol{F}} = \arg\min_{\boldsymbol{F}} \sum_{t=0}^{T-1} \|\boldsymbol{s}_{t+1} - \boldsymbol{F}\boldsymbol{s}_t\|_{\ell_2}^2$$

Recap: System Identification

Least-squares problem with matrix variable

$$\hat{\boldsymbol{F}} = \arg\min_{\boldsymbol{F}} \sum_{t=0}^{T-1} \|\boldsymbol{s}_{t+1} - \boldsymbol{F}\boldsymbol{s}_t\|_{\ell_2}^2$$

$$= \arg\min_{\boldsymbol{F}} \sum_{t=0}^{T-1} \|\boldsymbol{w}_t\|_{\ell_2}^2 \text{ s.t. } \underbrace{\begin{bmatrix} \boldsymbol{s}_1 \\ \vdots \\ \boldsymbol{s}_T \end{bmatrix}}_{\boldsymbol{S}_+} = \underbrace{\begin{bmatrix} \boldsymbol{s}_0 \\ \vdots \\ \boldsymbol{s}_{T-1} \end{bmatrix}}_{\boldsymbol{S}_-} \boldsymbol{F}^\mathsf{T} + \underbrace{\begin{bmatrix} \boldsymbol{w}_0 \\ \vdots \\ \boldsymbol{w}_{T-1} \end{bmatrix}}_{\boldsymbol{W}}$$

Recap: System Identification

Least-squares problem with matrix variable

$$\hat{\boldsymbol{F}} = \arg\min_{\boldsymbol{F}} \sum_{t=0}^{T-1} \|\boldsymbol{s}_{t+1} - \boldsymbol{F}\boldsymbol{s}_t\|_{\ell_2}^2$$

$$= \arg\min_{\boldsymbol{F}} \sum_{t=0}^{T-1} \|\boldsymbol{w}_t\|_{\ell_2}^2 \text{ s.t. } \underbrace{\begin{bmatrix} \boldsymbol{s}_1 \\ \vdots \\ \boldsymbol{s}_T \end{bmatrix}}_{\boldsymbol{S}_+} = \underbrace{\begin{bmatrix} \boldsymbol{s}_0 \\ \vdots \\ \boldsymbol{s}_{T-1} \end{bmatrix}}_{\boldsymbol{S}_-} \boldsymbol{F}^\mathsf{T} + \underbrace{\begin{bmatrix} \boldsymbol{w}_0 \\ \vdots \\ \boldsymbol{w}_{T-1} \end{bmatrix}}_{\boldsymbol{W}}$$

$$= \arg\min_{\boldsymbol{F}} \|\boldsymbol{S}_+ - \boldsymbol{S}_-\boldsymbol{F}^\mathsf{T}\|_F^2$$

Recap: System Identification

Least-squares problem with matrix variable

$$\hat{\boldsymbol{F}} = \arg\min_{\boldsymbol{F}} \sum_{t=0}^{T-1} \|\boldsymbol{s}_{t+1} - \boldsymbol{F}\boldsymbol{s}_t\|_{\ell_2}^2$$

$$= \arg\min_{\boldsymbol{F}} \sum_{t=0}^{T-1} \|\boldsymbol{w}_t\|_{\ell_2}^2 \text{ s.t. } \underbrace{\begin{bmatrix} \boldsymbol{s}_1 \\ \vdots \\ \boldsymbol{s}_T \end{bmatrix}}_{\boldsymbol{S}_+} = \underbrace{\begin{bmatrix} \boldsymbol{s}_0 \\ \vdots \\ \boldsymbol{s}_{T-1} \end{bmatrix}}_{\boldsymbol{S}_-} \boldsymbol{F}^\mathsf{T} + \underbrace{\begin{bmatrix} \boldsymbol{w}_0 \\ \vdots \\ \boldsymbol{w}_{T-1} \end{bmatrix}}_{\boldsymbol{W}}$$

$$= \arg\min_{\boldsymbol{F}} \|\boldsymbol{S}_+ - \boldsymbol{S}_-\boldsymbol{F}^\mathsf{T}\|_F^2$$

$$\hat{\boldsymbol{F}} = \boldsymbol{S}_+^\mathsf{T}\boldsymbol{S}_-(\boldsymbol{S}_-^\mathsf{T}\boldsymbol{S}_-)^{-1}$$

Recap: System Identification

Least-squares problem with matrix variable

$$\hat{\boldsymbol{F}} = \boldsymbol{S}_+^\mathsf{T} \boldsymbol{S}_- (\boldsymbol{S}_-^\mathsf{T} \boldsymbol{S}_-)^{-1}$$

- Unlike standard linear regression, the "features" $(\boldsymbol{s}_0, \ldots, \boldsymbol{s}_T)$ are not fixed or drawn independently
  - dependence: $\boldsymbol{s}_t = \boldsymbol{F}^t \boldsymbol{s}_0 + \sum_{\tau=0}^{t-1} \boldsymbol{F}^{t-\tau-1} \boldsymbol{w}_\tau$

Recap: System Identification

Least-squares problem with matrix variable

$$\hat{\boldsymbol{F}} = \boldsymbol{S}_+^\mathsf{T} \boldsymbol{S}_- (\boldsymbol{S}_-^\mathsf{T} \boldsymbol{S}_-)^{-1}$$

- Unlike standard linear regression, the "features" $(\boldsymbol{s}_0, \ldots, \boldsymbol{s}_T)$ are not fixed or drawn independently
  - dependence: $\boldsymbol{s}_t = \boldsymbol{F}^t \boldsymbol{s}_0 + \sum_{\tau=0}^{t-1} \boldsymbol{F}^{t-\tau-1} \boldsymbol{w}_\tau$
- Similar to standard linear regression, the estimation error (when $\boldsymbol{S}_-$ is full rank) can be bounded as

$$\|\hat{\boldsymbol{F}} - \boldsymbol{F}\| = \|\boldsymbol{W}^\mathsf{T} \boldsymbol{S}_- (\boldsymbol{S}_-^\mathsf{T} \boldsymbol{S}_-)^{-1}\|$$

Recap: System Identification

Least-squares problem with matrix variable

$$\hat{\boldsymbol{F}} = \boldsymbol{S}_+^\mathsf{T} \boldsymbol{S}_- (\boldsymbol{S}_-^\mathsf{T} \boldsymbol{S}_-)^{-1}$$

- Unlike standard linear regression, the "features" $(\boldsymbol{s}_0, \ldots, \boldsymbol{s}_T)$ are not fixed or drawn independently
  - dependence: $\boldsymbol{s}_t = \boldsymbol{F}^t \boldsymbol{s}_0 + \sum_{\tau=0}^{t-1} \boldsymbol{F}^{t-\tau-1} \boldsymbol{w}_\tau$
- Similar to standard linear regression, the estimation error (when $\boldsymbol{S}_-$ is full rank) can be bounded as

$$\begin{aligned} \|\hat{\boldsymbol{F}} - \boldsymbol{F}\| &= \|\boldsymbol{W}^\mathsf{T} \boldsymbol{S}_- (\boldsymbol{S}_-^\mathsf{T} \boldsymbol{S}_-)^{-1}\| \\ &\leq \frac{\|\boldsymbol{W}^\mathsf{T} \boldsymbol{S}_-\|}{\lambda_{\min}(\boldsymbol{S}_-^\mathsf{T} \boldsymbol{S}_-)} \end{aligned}$$

Today: Control Inputs

Linear dynamical system with control Inputs

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t$$

Today: Control Inputs

Linear dynamical system with control Inputs

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t$$

- System dynamics $\boldsymbol{F} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{G} \in \mathbb{R}^{n \times n}$

Today: Control Inputs

Linear dynamical system with control Inputs

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t$$

- System dynamics $\boldsymbol{F} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{G} \in \mathbb{R}^{n \times n}$
- Two types of inputs:

Today: Control Inputs

Linear dynamical system with control Inputs

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t$$

- System dynamics $\boldsymbol{F} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{G} \in \mathbb{R}^{n \times n}$
- Two types of inputs:
  - the disturbance $\boldsymbol{w}_t \in \mathbb{R}^n$

Today: Control Inputs

Linear dynamical system with control Inputs

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t$$

- System dynamics $\boldsymbol{F} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{G} \in \mathbb{R}^{n \times n}$
- Two types of inputs:
  - the disturbance $\boldsymbol{w}_t \in \mathbb{R}^n$
  - the control input $\boldsymbol{u}_t \in \mathbb{R}^m$

Today: Control Inputs

Linear dynamical system with control Inputs

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t$$

- System dynamics $\boldsymbol{F} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{G} \in \mathbb{R}^{n \times n}$
- Two types of inputs:
  - the disturbance $\boldsymbol{w}_t \in \mathbb{R}^n$
  - the control input $\boldsymbol{u}_t \in \mathbb{R}^m$
- Identify the dynamics model $\boldsymbol{s}_{0:T}, \boldsymbol{u}_{0:T} \rightarrow \hat{\boldsymbol{F}}, \hat{\boldsymbol{G}}$

Today: Control Inputs

Linear dynamical system with control Inputs

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t$$

Today: Control Inputs

Linear dynamical system with control Inputs

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t$$

- Suppose $\rho(\boldsymbol{F}) > 1$, can we choose $\boldsymbol{u}_t$ to stabilize the system?

Today: Control Inputs

Linear dynamical system with control Inputs

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$ $\quad \boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t}$
- Suppose $\rho(\boldsymbol{F}) > 1$, can we choose $\boldsymbol{u}_t$ to stabilize the system?
- Choose $\boldsymbol{u}_t = -\boldsymbol{K}\boldsymbol{s}_t$: $\quad \boxed{\boldsymbol{s}_{t+1} = (\boldsymbol{F} - \boldsymbol{G}\boldsymbol{K})\boldsymbol{s}_t + \boldsymbol{w}_t}$

Today: Control Inputs

Linear dynamical system with control Inputs

- Observe states $\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_T$ $\quad \boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t}$
- Suppose $\rho(\boldsymbol{F}) > 1$, can we choose $\boldsymbol{u}_t$ to stabilize the system?
- Choose $\boldsymbol{u}_t = -\boldsymbol{K}\boldsymbol{s}_t$: $\quad \boxed{\boldsymbol{s}_{t+1} = (\boldsymbol{F} - \boldsymbol{G}\boldsymbol{K})\boldsymbol{s}_t + \boldsymbol{w}_t}$
- closed-loop dynamics $\tilde{\boldsymbol{F}} = (\boldsymbol{F} - \boldsymbol{G}\boldsymbol{K}) \in \mathbb{R}^{n \times n}$

Today: Nonlinear Dynamical Systems

Nonlinear dynamical systems with state observations

- state $\boldsymbol{s}_t \in \mathbb{R}^n$
- input $\boldsymbol{u}_t \in \mathbb{R}^m$
- noise $\boldsymbol{w}_t \in \mathbb{R}^n$
- system dynamics $\boldsymbol{\theta}_\star \in \mathbb{R}^d$

$$\boldsymbol{s}_{t+1} = \phi(\boldsymbol{s}_t, \boldsymbol{u}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t$$

Today: Nonlinear Dynamical Systems

Nonlinear dynamical systems with state observations

$$\boxed{\boldsymbol{s}_{t+1} = \phi(\boldsymbol{s}_t, \boldsymbol{u}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t}$$

Example:

- Standard linear dynamical system
  - State equation: $\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t$
  - System dynamics: $\theta_\star = [\boldsymbol{F}\ \boldsymbol{G}]$

Goal: Learning the dynamics $\theta_\star$ from data

Today: Nonlinear Dynamical Systems

Nonlinear dynamical systems with state observations

$$\boxed{\boldsymbol{s}_{t+1} = \phi(\boldsymbol{s}_t, \boldsymbol{u}_t; \boldsymbol{\theta}_\star) + \boldsymbol{w}_t}$$

Example:

- RNN type nonlinear dynamical system
  - State equation: $\boldsymbol{s}_{t+1} = \phi(\boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t) + \boldsymbol{w}_t$
  - System dynamics: $\boldsymbol{\theta}_\star = [\boldsymbol{F}\ \ \boldsymbol{G}]$

Goal: Learning the dynamics $\boldsymbol{\theta}_\star$ from data

Learning from Finite Data

Run the system until time T, collect $(\boldsymbol{s}_t, \boldsymbol{u}_t)_{t=0}^{T}$

Learning from Finite Data

Run the system until time T, collect $(\boldsymbol{s}_t, \boldsymbol{u}_t)_{t=0}^{T}$

1. Set loss function $\hat{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{2(T-L)} \sum_{t=L}^{T-1} \|\boldsymbol{s}_{t+1} - \phi(\boldsymbol{s}_t, \boldsymbol{u}_t; \boldsymbol{\theta})\|_{\ell_2}^2$.

Learning from Finite Data

Run the system until time T, collect $(\boldsymbol{s}_t, \boldsymbol{u}_t)_{t=0}^{T}$

1. Set loss function $\hat{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{2(T-L)} \sum_{t=L}^{T-1} \|\boldsymbol{s}_{t+1} - \phi(\boldsymbol{s}_t, \boldsymbol{u}_t; \boldsymbol{\theta})\|_{\ell_2}^2$.
2. Find $\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \hat{\mathcal{L}}(\boldsymbol{\theta})$ (e.g. via gradient descent)

Learning from Finite Data

Run the system until time T, collect $(\boldsymbol{s}_t, \boldsymbol{u}_t)_{t=0}^{T}$

1. Set loss function $\hat{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{2(T-L)} \sum_{t=L}^{T-1} \|\boldsymbol{s}_{t+1} - \phi(\boldsymbol{s}_t, \boldsymbol{u}_t; \boldsymbol{\theta})\|_{\ell_2}^2$.
2. Find $\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \hat{\mathcal{L}}(\boldsymbol{\theta})$ (e.g. via gradient descent)
3. **Hope that $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}_\star$**

Learning from Finite Data

Run the system until time T, collect $(\boldsymbol{s}_t, \boldsymbol{u}_t)_{t=0}^{T}$

1. Set loss function $\hat{\mathcal{L}}(\boldsymbol{\theta}) = \frac{1}{2(T-L)} \sum_{t=L}^{T-1} \|\boldsymbol{s}_{t+1} - \phi(\boldsymbol{s}_t, \boldsymbol{u}_t; \boldsymbol{\theta})\|_{\ell_2}^2$.
2. Find $\hat{\boldsymbol{\theta}} := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \hat{\mathcal{L}}(\boldsymbol{\theta})$ (e.g. via gradient descent)
3. **Hope that $\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}_\star$**

Challenges:

- temporal dependence
- nonlinearity
- finite samples

Nonlinear Stability

Nonlinear dynamical systems: Use ρ-stability

Definition (ρ-stabilized system)

(1) Pick inputs $\boldsymbol{u}_t = \boldsymbol{\pi}(\boldsymbol{s}_t) + \boldsymbol{z}_t$. Fix $(\boldsymbol{z}_\tau)_{\tau=0}^{t-1}$ and $(\boldsymbol{w}_\tau)_{\tau=0}^{t-1}$.
(2) Denote the state sequence resulting from initial state $\boldsymbol{s}_0 = \boldsymbol{\alpha}$ by $\boldsymbol{s}_t(\boldsymbol{\alpha})$.
(3) There exists $C_\rho \geq 1$ and $\rho \in (0, 1)$ such that for all $\boldsymbol{\alpha}$, $(\boldsymbol{z}_t)_{t\geq 0}$ and $(\boldsymbol{w}_t)_{t\geq 0}$, we have

$$\|\boldsymbol{s}_t(\boldsymbol{\alpha}) - \boldsymbol{s}_t(0)\|_{\ell_2} \leq C_\rho \rho^t \|\boldsymbol{\alpha}\|_{\ell_2},$$

ρ corresponds to *nonlinear spectral radius* (not easy to calculate).

Nonlinear Stability

Nonlinear dynamical systems: Use ρ-stability

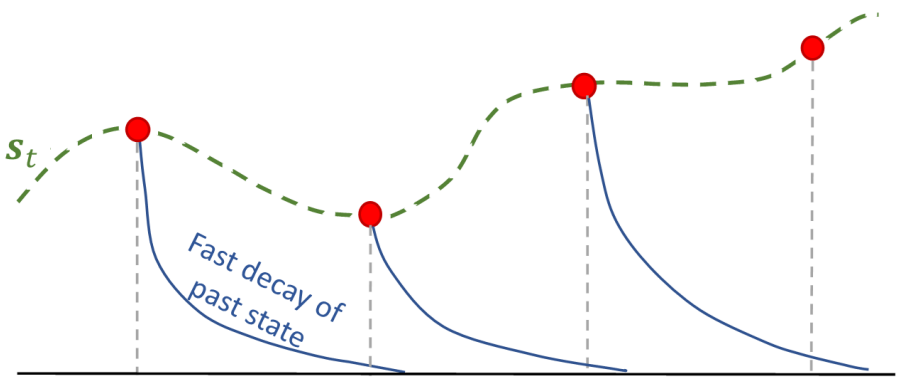
Definition (ρ-stabilized system)

(1) Pick inputs $\boldsymbol{u}_t = \boldsymbol{\pi}(\boldsymbol{s}_t) + \boldsymbol{z}_t$. Fix $(\boldsymbol{z}_\tau)_{\tau=0}^{t-1}$ and $(\boldsymbol{w}_\tau)_{\tau=0}^{t-1}$.
(2) Denote the state sequence resulting from initial state $\boldsymbol{s}_0 = \boldsymbol{\alpha}$ by $\boldsymbol{s}_t(\boldsymbol{\alpha})$.
(3) There exists $C_\rho \geq 1$ and $\rho \in (0, 1)$ such that for all $\boldsymbol{\alpha}$, $(\boldsymbol{z}_t)_{t\geq 0}$ and $(\boldsymbol{w}_t)_{t\geq 0}$, we have

$$\|\boldsymbol{s}_t(\boldsymbol{\alpha}) - \boldsymbol{s}_t(0)\|_{\ell_2} \leq C_\rho \rho^t \|\boldsymbol{\alpha}\|_{\ell_2},$$

ρ corresponds to *nonlinear spectral radius* (not easy to calculate).

Key observation: System forgets the past quickly



Nonlinear Stability

Linear dynamical systems: Gelfand's formula

$$\rho(\boldsymbol{F}) = \lim_{k \to \infty} \|\boldsymbol{F}^k\|^{1/k}$$

Nonlinear Stability

Linear dynamical systems: Gelfand's formula

$$\boxed{\rho(\boldsymbol{F}) = \lim_{k\to\infty} \|\boldsymbol{F}^k\|^{1/k}}$$

$$\|\boldsymbol{s}_t(\boldsymbol{\alpha}) - \boldsymbol{s}_t(0)\|_{\ell_2} = \|\boldsymbol{F}^t\boldsymbol{\alpha} - \boldsymbol{F}^t 0\|_{\ell_2}$$

Nonlinear Stability

Linear dynamical systems: Gelfand's formula

$$\boxed{\rho(\boldsymbol{F}) = \lim_{k\to\infty} \|\boldsymbol{F}^k\|^{1/k}}$$

$$\begin{aligned}\|\boldsymbol{s}_t(\boldsymbol{\alpha}) - \boldsymbol{s}_t(0)\|_{\ell_2} &= \|\boldsymbol{F}^t\boldsymbol{\alpha} - \boldsymbol{F}^t 0\|_{\ell_2} \\ &\leq \|\boldsymbol{F}^t\|\|\boldsymbol{\alpha}\|_{\ell_2}\end{aligned}$$

Nonlinear Stability

Linear dynamical systems: Gelfand's formula

$$\rho(\boldsymbol{F}) = \lim_{k\to\infty} \|\boldsymbol{F}^k\|^{1/k}$$

$$\begin{aligned}
\|\boldsymbol{s}_t(\boldsymbol{\alpha}) - \boldsymbol{s}_t(0)\|_{\ell_2} &= \|\boldsymbol{F}^t\boldsymbol{\alpha} - \boldsymbol{F}^t 0\|_{\ell_2} \\
&\leq \|\boldsymbol{F}^t\|\|\boldsymbol{\alpha}\|_{\ell_2} \\
&\leq C_\rho \rho(\boldsymbol{F})^t \|\boldsymbol{\alpha}\|_{\ell_2}
\end{aligned}$$

Assumptions on the System and Inputs

Assumption (Stability)

The closed loop system $\tilde{\phi}$ is ρ-stable.

Assumptions on the System and Inputs

Assumption (Stability)

The closed loop system $\tilde{\phi}$ is ρ-stable.

Assumption (Boundedness)

There exist scalars $B, c_w, \sigma > 0$, such that $(\boldsymbol{z}_t)_{t\geq 0} \overset{i.i.d.}{\sim} \mathcal{D}_z$ and $(\boldsymbol{w}_t)_{t\geq 0} \overset{i.i.d.}{\sim} \mathcal{D}_w$ obey $\|\tilde{\phi}(0, \boldsymbol{z}_t; \boldsymbol{\theta}_\star)\|_{\ell_2} \leq B\sqrt{n}$ and $\|\boldsymbol{w}_t\|_{\ell_\infty} \leq c_w\sigma$ for $0 \leq t \leq T-1$ with probability at least $1 - p_0$ over the generation of data.

Mixing-time Approach

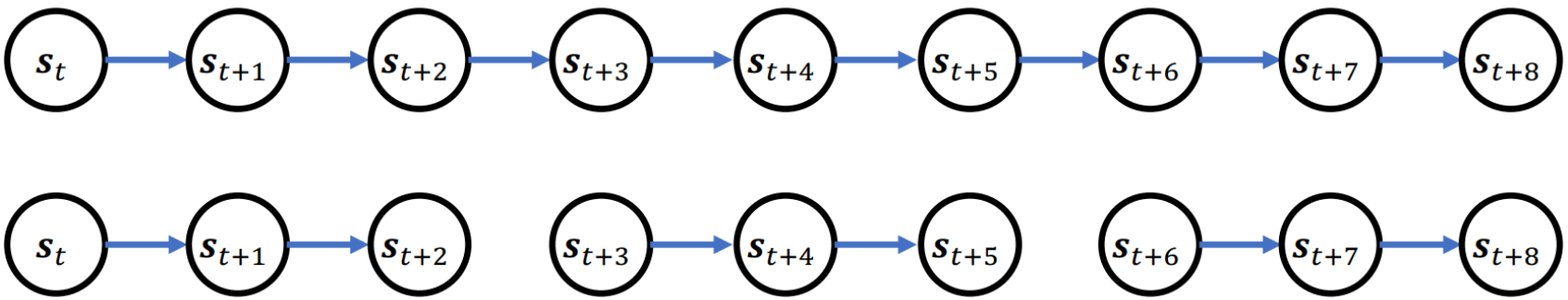
Suppose noise $(\boldsymbol{w}_t)_{t\geq 0}$ and excitation $(\boldsymbol{z}_t)_{t\geq 0}$ are i.i.d and $\boldsymbol{s}_0 = 0$.

- Leverage mixing-time argument with stability

Mixing-time Approach

Suppose noise $(\boldsymbol{w}_t)_{t\geq 0}$ and excitation $(\boldsymbol{z}_t)_{t\geq 0}$ are i.i.d and $\boldsymbol{s}_0 = 0$.

- Leverage mixing-time argument with stability



Related works

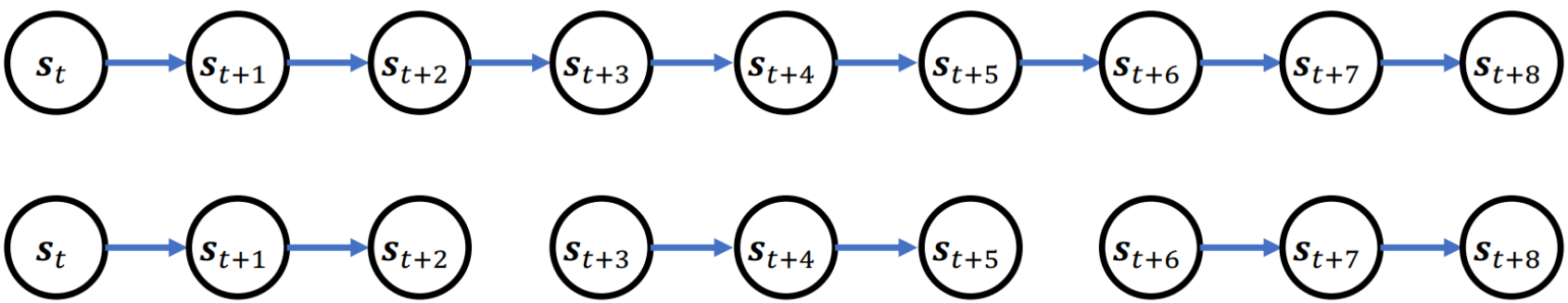
- Mixing-time/stability: Yu (Annals of Prob.'94), Mohri & Rostamizadeh (JMLR'10), Miller & Hardt (ICLR'19), Oymak (COLT'19)
- Martingale-based: Recht, Rakhlin, Tewari & coauthors

Mixing-time Approach

Definition (Multiple trajectory loss)

Let $\boldsymbol{s}_{t+1,L}, \boldsymbol{s}_{t,L-1}$ be L-truncated and $L-1$-truncated states at time $t+1$ and t respectively. We define the truncated (empirical) risk as

$$\hat{\mathcal{L}}^{\text{tr}}(\boldsymbol{\theta}) := \frac{1}{2(T-L)} \sum_{t=L}^{T-1} \|\boldsymbol{s}_{t+1,L} - \tilde{\phi}(\boldsymbol{s}_{t,L-1}, \boldsymbol{z}_t; \boldsymbol{\theta})\|_{\ell_2}^2.$$



Mixing-time Approach

Theorem (Difference between single and multiple trajectories)

Under ρ-stability and boundedness assumptions, with probability at least $1 - p_0$, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, we have

$$\|\nabla \hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \hat{\mathcal{L}}^{tr}(\boldsymbol{\theta})\|_{\ell_2} \lesssim C_\rho \rho^{L-1}(c_w \sigma + C_{\tilde{\phi}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}).$$

Optimization Landscape

To concretely show how stability helps, we define the following loss function, obtained from i.i.d. samples at time $L-1$ and can be used as a proxy for $\mathbb{E}[\hat{\mathcal{L}}]$.

Optimization Landscape

To concretely show how stability helps, we define the following loss function, obtained from i.i.d. samples at time $L-1$ and can be used as a proxy for $\mathbb{E}[\hat{\mathcal{L}}]$.

Definition (Population Loss)

Suppose $\boldsymbol{s}_0 = 0$. Let $(\boldsymbol{z}_t)_{t\geq 0} \overset{\text{i.i.d.}}{\sim} \mathcal{D}_z$ and $(\boldsymbol{w}_t)_{t\geq 0} \overset{\text{i.i.d.}}{\sim} \mathcal{D}_w$. The auxiliary loss is defined as the expected loss at timestamp $L-1$, that is,

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{2}\,\mathbb{E}[\|\boldsymbol{s}_L - \tilde{\phi}(\boldsymbol{s}_{L-1}, \boldsymbol{z}_{L-1}; \boldsymbol{\theta})\|_{\ell_2}^2].$$

Uniform Convergence of Empirical Gradient

Define the empirical and population losses,

$$\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}_i) \quad \text{and} \quad \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})].$$

Uniform Convergence of Empirical Gradient

Define the empirical and population losses,

$$\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}_i) \quad \text{and} \quad \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})].$$

Assumption (Lipschitz gradients)

There exist numbers $L_{\mathcal{D}}, p_0 > 0$ such that with probability at least $1 - p_0$ over the generation of data, for all pairs $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$,

$$\max(\|\nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}')\|_{\ell_2}, \|\nabla \hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla \hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}')\|_{\ell_2}) \leq L_{\mathcal{D}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_{\ell_2}.$$

Uniform Convergence of Empirical Gradient

Define the empirical and population losses,

$$\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}_i) \quad \text{and} \quad \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})].$$

Assumption (Subexponential gradient noise)

There exist scalars $K, \sigma_0 > 0$ such that, given $\boldsymbol{x} := (\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1}) \sim \mathcal{D}$, at any point $\boldsymbol{\theta}$, the subexponential norm of the gradient is upper bounded as a function of the noise level σ_0 and distance to the population minimizer via

$$\|\nabla \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}) - \mathbb{E}[\nabla \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})]\|_{\psi_1} \leq \sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}.$$

Uniform Convergence of Empirical Gradient

Define the empirical and population losses,

$$\hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}_i) \quad \text{and} \quad \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})].$$

Assumption (Subexponential gradient noise)

There exist scalars $K, \sigma_0 > 0$ such that, given $\boldsymbol{x} := (\boldsymbol{h}_L, \boldsymbol{h}_{L-1}, \boldsymbol{z}_{L-1}) \sim \mathcal{D}$, at any point $\boldsymbol{\theta}$, the subexponential norm of the gradient is upper bounded as a function of the noise level σ_0 and distance to the population minimizer via

$$\|\nabla \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x}) - \mathbb{E}[\nabla \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{x})]\|_{\psi_1} \leq \sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}.$$

For a random variable $X \in \mathbb{R}$, $\|X\|_{\psi_1} := \sup_{k \geq 1} \frac{(\mathbb{E}[|X|^k])^{1/k}}{k}$
For a random vector $\boldsymbol{x} \in \mathbb{R}^n$ is $\|\boldsymbol{x}\|_{\psi_1} := \sup_{\boldsymbol{v} \in \mathcal{S}^{n-1}} \|\boldsymbol{v}^\top \boldsymbol{x}\|_{\psi_1}$

Uniform Convergence of Empirical Gradient

Theorem (Uniform gradient convergence)

Suppose the gradients of $\mathcal{L}_{\mathcal{D}}$ and $\mathcal{L}_{\mathcal{S}}$ obey the above two Assumptions. Then, for all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, with probability $1 - p_0 - \log(\frac{Kr}{\sigma_0}) \exp(-100d)$, we have

$$\|\nabla \hat{\mathcal{L}}_{\mathcal{S}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \lesssim (\sigma_0 + K\|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}) C_{\log} \sqrt{\frac{d}{N}}.$$

- Here, $C_{\log} = \log(3(L_{\mathcal{D}} N/K + 1))$

Optimization Landscape

A special case of Polyak-Lojasiewicz inequality and provides a generalization of strong convexity to nonconvex functions.

Assumption (One-point convexity & smoothness)

There exist scalars $\beta \geq \alpha > 0$ *such that the auxiliary loss* $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ *satisfies*

$$\langle \boldsymbol{\theta} - \boldsymbol{\theta}_\star, \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) \rangle \geq \alpha \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}^2,$$
$$\|\nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \leq \beta \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2}.$$

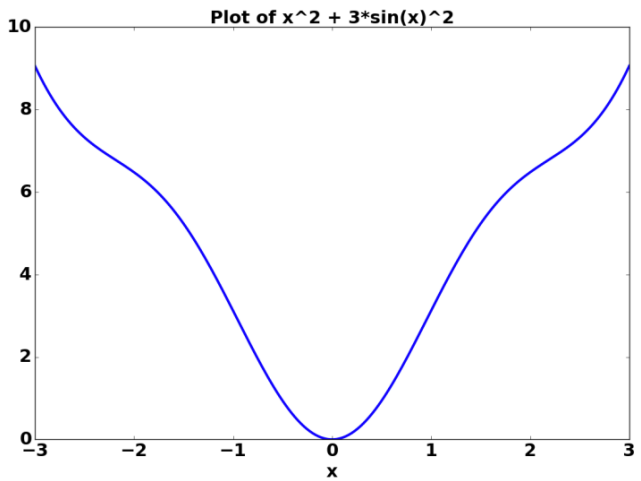


Figure 1: Not a convex function. But it also does not look challenging to optimize. Polyak-Lojasiewicz condition can capture such functions (and beyond).

Main Result

Theorem (Finite sample learning)

- *Suppose* $\rho < 1$ *and Assumptions above hold*

Main Result

Theorem (Finite sample learning)

- *Suppose* $\rho < 1$ *and Assumptions above hold*
- *Suppose* $\mathcal{L}_{\mathcal{D}}$ *satisfies OPCS with* $\alpha, \beta > 0$

Main Result

Theorem (Finite sample learning)

- *Suppose* $\rho < 1$ *and Assumptions above hold*
- *Suppose* $\mathcal{L}_{\mathcal{D}}$ *satisfies OPCS with* $\alpha, \beta > 0$
- *Suppose trajectory length* T *obeys*

$$T \gtrsim L(N+1), \quad \textit{where} \quad N \propto d \log(N), \quad L \propto 1 + \log(N)/\log(\rho^{-1})$$

Main Result

Theorem (Finite sample learning)

- *Suppose* $\rho < 1$ *and Assumptions above hold*
- *Suppose* $\mathcal{L}_{\mathcal{D}}$ *satisfies OPCS with* $\alpha, \beta > 0$
- *Suppose trajectory length* T *obeys*

$$T \gtrsim L(N+1), \quad \textit{where} \quad N \propto d\log(N), \quad L \propto 1 + \log(N)/\log(\rho^{-1})$$

Pick a proper constant learning rate and starting from $\boldsymbol{\theta}_0 \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$*, run gradient iterates* $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta \nabla \hat{\mathcal{L}}(\boldsymbol{\theta}_\tau)$*. All iterates satisfy*

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\star\|_{\ell_2} \leq \underbrace{(1 - \frac{\alpha^2}{128\beta^2})^\tau \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_{\ell_2}}_{\textit{fast convergence}} + \underbrace{\frac{c\sigma_0}{\alpha}\sqrt{\frac{d}{N}}}_{\textit{statistical error}} .$$

Main Result

Theorem (Finite sample learning)

- *Suppose* $\rho < 1$ *and Assumptions above hold*
- *Suppose* $\mathcal{L}_{\mathcal{D}}$ *satisfies OPCS with* $\alpha, \beta > 0$
- *Suppose trajectory length* T *obeys*

$$T \gtrsim L(N+1), \quad \textit{where} \quad N \propto d\log(N), \quad L \propto 1 + \log(N)/\log(\rho^{-1})$$

Pick a proper constant learning rate and starting from $\boldsymbol{\theta}_0 \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$*, run gradient iterates* $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta \nabla \hat{\mathcal{L}}(\boldsymbol{\theta}_\tau)$*. All iterates satisfy*

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\star\|_{\ell_2} \leq \underbrace{(1 - \frac{\alpha^2}{128\beta^2})^\tau \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_{\ell_2}}_{\textit{fast convergence}} + \underbrace{\frac{c\sigma_0}{\alpha}\sqrt{\frac{d}{N}}}_{\textit{statistical error}} .$$

Remark: Need to calculate α, β, σ_0 etc. for specific scenario e.g.

$$\boldsymbol{s}_{t+1} = \phi(\boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t) + \boldsymbol{w}_t$$

Case Study

Linear Dynamical System: $\boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t}$

$[\boldsymbol{F}\ \boldsymbol{G}] = [\boldsymbol{\theta}_1\ \cdots\ \boldsymbol{\theta}_n]^T \in \mathbb{R}^{n\times(n+m)}$

Case Study

Linear Dynamical System: $\boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t}$

$$[\boldsymbol{F}\ \boldsymbol{G}] = [\boldsymbol{\theta}_1\ \cdots\ \boldsymbol{\theta}_n]^T \in \mathbb{R}^{n\times(n+m)}$$

$$\boldsymbol{s}_t = \boldsymbol{F}^t\boldsymbol{s}_0 + \sum_{\tau=0}^{t-1}\boldsymbol{F}^{t-\tau-1}\boldsymbol{G}\boldsymbol{u}_\tau + \sum_{\tau=0}^{t-1}\boldsymbol{F}^{t-\tau-1}\boldsymbol{w}_\tau$$

Case Study

Linear Dynamical System: $\boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t}$

$[\boldsymbol{F}\ \boldsymbol{G}] = [\boldsymbol{\theta}_1\ \cdots\ \boldsymbol{\theta}_n]^T \in \mathbb{R}^{n\times(n+m)}$

$$\boldsymbol{s}_t = \boldsymbol{F}^t\boldsymbol{s}_0 + \sum_{\tau=0}^{t-1}\boldsymbol{F}^{t-\tau-1}\boldsymbol{G}\boldsymbol{u}_\tau + \sum_{\tau=0}^{t-1}\boldsymbol{F}^{t-\tau-1}\boldsymbol{w}_\tau$$

$$= \boldsymbol{F}^t\boldsymbol{s}_0 + \boldsymbol{\Gamma}_t^G \begin{bmatrix} \boldsymbol{u}_0 \\ \vdots \\ \boldsymbol{u}_{T-1} \end{bmatrix} + \boldsymbol{\Gamma}_t \begin{bmatrix} \boldsymbol{w}_0 \\ \vdots \\ \boldsymbol{w}_{T-1} \end{bmatrix}$$

Case Study

Linear Dynamical System: $\boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t}$

$[\boldsymbol{F}\ \boldsymbol{G}] = [\boldsymbol{\theta}_1\ \cdots\ \boldsymbol{\theta}_n]^T \in \mathbb{R}^{n\times(n+m)}$

$$\begin{aligned}\boldsymbol{s}_t &= \boldsymbol{F}^t\boldsymbol{s}_0 + \sum_{\tau=0}^{t-1}\boldsymbol{F}^{t-\tau-1}\boldsymbol{G}\boldsymbol{u}_\tau + \sum_{\tau=0}^{t-1}\boldsymbol{F}^{t-\tau-1}\boldsymbol{w}_\tau \\ &= \boldsymbol{F}^t\boldsymbol{s}_0 + \boldsymbol{\Gamma}_t^G\begin{bmatrix}\boldsymbol{u}_0\\ \vdots \\ \boldsymbol{u}_{T-1}\end{bmatrix} + \boldsymbol{\Gamma}_t\begin{bmatrix}\boldsymbol{w}_0\\ \vdots \\ \boldsymbol{w}_{T-1}\end{bmatrix}\end{aligned}$$

where $\boldsymbol{\Gamma}_t^G := [\boldsymbol{F}^{t-1}\boldsymbol{G}\ \boldsymbol{F}^{t-2}\boldsymbol{G}\ \cdots\ \boldsymbol{G}]$ and $\boldsymbol{\Gamma}_t := [\boldsymbol{F}^{t-1}\ \boldsymbol{F}^{t-2}\ \cdots\ \boldsymbol{I}_n]$

Case Study

Linear Dynamical System: $\boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t}$

$$[\boldsymbol{F}\ \boldsymbol{G}] = [\boldsymbol{\theta}_1\ \cdots\ \boldsymbol{\theta}_n]^T \in \mathbb{R}^{n\times(n+m)}$$

$$\begin{aligned}\boldsymbol{s}_t &= \boldsymbol{F}^t\boldsymbol{s}_0 + \sum_{\tau=0}^{t-1}\boldsymbol{F}^{t-\tau-1}\boldsymbol{G}\boldsymbol{u}_\tau + \sum_{\tau=0}^{t-1}\boldsymbol{F}^{t-\tau-1}\boldsymbol{w}_\tau \\ &= \boldsymbol{F}^t\boldsymbol{s}_0 + \boldsymbol{\Gamma}_t^G\begin{bmatrix}\boldsymbol{u}_0\\ \vdots \\ \boldsymbol{u}_{T-1}\end{bmatrix} + \boldsymbol{\Gamma}_t\begin{bmatrix}\boldsymbol{w}_0\\ \vdots \\ \boldsymbol{w}_{T-1}\end{bmatrix}\end{aligned}$$

where $\boldsymbol{\Gamma}_t^G := [\boldsymbol{F}^{t-1}\boldsymbol{G}\ \boldsymbol{F}^{t-2}\boldsymbol{G}\ \cdots\ \boldsymbol{G}]$ and $\boldsymbol{\Gamma}_t := [\boldsymbol{F}^{t-1}\ \boldsymbol{F}^{t-2}\ \cdots\ \boldsymbol{I}_n]$

$\boldsymbol{\Gamma}_t^G\boldsymbol{\Gamma}_t^{G\top}$ and $\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_t^\top$ are the finite time controllability Gramians for the control and noise inputs, respectively.

Case Study

Linear Dynamical System: $\boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t}$

- Suppose $\boldsymbol{s}_0 = 0$

Case Study

Linear Dynamical System: $\boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t}$

- Suppose $\boldsymbol{s}_0 = 0$
- Suppose $\boldsymbol{u}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_m)$ and $\boldsymbol{w}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_w^2 \boldsymbol{I}_n)$

Case Study

Linear Dynamical System: $\boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t}$

- Suppose $\boldsymbol{s}_0 = 0$
- Suppose $\boldsymbol{u}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_m)$ and $\boldsymbol{w}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_w^2 \boldsymbol{I}_n)$
- $\boxed{\boldsymbol{s}_t \sim \mathcal{N}(0, \boldsymbol{\Gamma}_t^G \boldsymbol{\Gamma}_t^{G\top} + \sigma_w^2 \boldsymbol{\Gamma}_t \boldsymbol{\Gamma}_t^\top)}$

Case Study

Linear Dynamical System: $\boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t}$

Theorem (simplified)

- *Suppose* $\rho(\boldsymbol{F}) < 1$.
- *Suppose* $\boldsymbol{u}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$ *and* $\boldsymbol{w}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2 \boldsymbol{I}_n)$.
- *Suppose trajectory length* T *obeys*

$$T \gtrsim L(N+1), \text{ where } N \propto (n+m)\log(N),\ L \propto 1 + \log(N)/\log(\rho^{-1})$$

Case Study

Linear Dynamical System: $\boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}\boldsymbol{s}_t + \boldsymbol{G}\boldsymbol{u}_t + \boldsymbol{w}_t}$

Theorem (simplified)

- *Suppose* $\rho(\boldsymbol{F}) < 1$.
- *Suppose* $\boldsymbol{u}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$ *and* $\boldsymbol{w}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_w^2 \boldsymbol{I}_n)$.
- *Suppose trajectory length* T *obeys*

$$T \gtrsim L(N+1), \text{ where } N \propto (n+m)\log(N),\ L \propto 1 + \log(N)/\log(\rho^{-1})$$

With proper learning rate and the initialization $[\boldsymbol{F}^{(0)}\ \boldsymbol{G}^{(0)}] = 0$, *all gradient descent iterates satisfy*

$$\|\boldsymbol{\theta}_k^{(\tau)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} \leq (1 - \frac{\gamma_-^2}{128\gamma_+^2})^\tau \|\boldsymbol{\theta}_k^{(0)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \frac{c\sigma}{\gamma_-}\sqrt{\gamma_+}\log(9N)\sqrt{\frac{n+m}{N}}.$$

Another Interesting Nonlinear Dynamical System

Bilinear dynamical systems with state observations

- state $\boldsymbol{s}_t \in \mathbb{R}^n$
- input $\boldsymbol{u}_t \in \mathbb{R}^m$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}_0 \boldsymbol{x}_t + \sum_{k=1}^{m} \boldsymbol{u}_t[k] \boldsymbol{F}_k \boldsymbol{x}_t + \boldsymbol{w}_{t+1}$$

- noise $\boldsymbol{w}_t \in \mathbb{R}^n$
- system dynamics $\{\boldsymbol{F}_k\}_{k=0}^m \in (\mathbb{R}^{n \times n})^{m+1}$

Another Interesting Nonlinear Dynamical System

Bilinear dynamical systems with state observations

- state $\boldsymbol{s}_t \in \mathbb{R}^n$
- input $\boldsymbol{u}_t \in \mathbb{R}^m$

$$\boldsymbol{s}_{t+1} = \boldsymbol{F}_0 \boldsymbol{x}_t + \sum_{k=1}^{m} \boldsymbol{u}_t[k] \boldsymbol{F}_k \boldsymbol{x}_t + \boldsymbol{w}_{t+1}$$

- noise $\boldsymbol{w}_t \in \mathbb{R}^n$
- system dynamics $\{\boldsymbol{F}_k\}_{k=0}^m \in (\mathbb{R}^{n \times n})^{m+1}$

Goal: Learning the dynamics $\{\boldsymbol{F}_k\}_{k=0}^m$ from data

Learning Without Mixing

Run the system until time T, collect a single $(\boldsymbol{s}_t, \boldsymbol{u}_t)_{t=1}^{T}$

Learning Without Mixing

Run the system until time T, collect a single $(\boldsymbol{s}_t, \boldsymbol{u}_t)_{t=1}^{T}$

- System dynamics $\boldsymbol{F}^\star := \begin{bmatrix} \boldsymbol{F}_0 & \sigma_{\boldsymbol{u}} \boldsymbol{F}_1 & \cdots & \sigma_{\boldsymbol{u}} \boldsymbol{F}_m \end{bmatrix} \in \mathbb{R}^{n \times n(m+1)}$

Learning Without Mixing

Run the system until time T, collect a single $(\boldsymbol{s}_t, \boldsymbol{u}_t)_{t=1}^{T}$

- System dynamics $\boldsymbol{F}^\star := \begin{bmatrix} \boldsymbol{F}_0 & \sigma_{\boldsymbol{u}} \boldsymbol{F}_1 & \cdots & \sigma_{\boldsymbol{u}} \boldsymbol{F}_m \end{bmatrix} \in \mathbb{R}^{n \times n(m+1)}$
- State $\tilde{\boldsymbol{s}}_t := \begin{bmatrix} \boldsymbol{s}_t^\mathsf{T} & \sigma_{\boldsymbol{u}}^{-1} \boldsymbol{u}_t[1] \boldsymbol{s}_t^\mathsf{T} & \cdots & \sigma_{\boldsymbol{u}}^{-1} \boldsymbol{u}_t[m] \boldsymbol{s}_t^\mathsf{T} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{n(m+1)}$

Learning Without Mixing

Run the system until time T, collect a single $(\boldsymbol{s}_t, \boldsymbol{u}_t)_{t=1}^{T}$

- System dynamics $\boldsymbol{F}^\star := \begin{bmatrix} \boldsymbol{F}_0 & \sigma_{\boldsymbol{u}} \boldsymbol{F}_1 & \cdots & \sigma_{\boldsymbol{u}} \boldsymbol{F}_m \end{bmatrix} \in \mathbb{R}^{n \times n(m+1)}$
- State $\tilde{\boldsymbol{s}}_t := \begin{bmatrix} \boldsymbol{s}_t^\mathsf{T} & \sigma_{\boldsymbol{u}}^{-1} \boldsymbol{u}_t[1] \boldsymbol{s}_t^\mathsf{T} & \cdots & \sigma_{\boldsymbol{u}}^{-1} \boldsymbol{u}_t[m] \boldsymbol{s}_t^\mathsf{T} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{n(m+1)}$
- $\boxed{\boldsymbol{s}_{t+1} = \boldsymbol{F}^\star \tilde{\boldsymbol{s}}_t + \boldsymbol{w}_{t+1}}$

Learning Without Mixing

- Suppose $\{\boldsymbol{u}_t\}_{t=0}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{u}}^2 \boldsymbol{I}_m)$ and $\{\boldsymbol{w}_t\}_{t=1}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_n)$.

Learning Without Mixing

- Suppose $\{\boldsymbol{u}_t\}_{t=0}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{u}}^2 \boldsymbol{I}_m)$ and $\{\boldsymbol{w}_t\}_{t=1}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_n)$.
- Set the loss function $\hat{\mathcal{L}}(\boldsymbol{F}) = \frac{1}{2T} \sum_{t=1}^{T} \|\boldsymbol{s}_{t+1} - \boldsymbol{F}\tilde{\boldsymbol{s}}_t\|_{\ell_2}^2$.

Learning Without Mixing

- Suppose $\{\boldsymbol{u}_t\}_{t=0}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{u}}^2 \boldsymbol{I}_m)$ and $\{\boldsymbol{w}_t\}_{t=1}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_n)$.
- Set the loss function $\hat{\mathcal{L}}(\boldsymbol{F}) = \frac{1}{2T} \sum_{t=1}^{T} \|\boldsymbol{s}_{t+1} - \boldsymbol{F}\tilde{\boldsymbol{s}}_t\|_{\ell_2}^2$.
- Over-determined problem, solution $\hat{\boldsymbol{F}} = \boldsymbol{S}_+^{\mathsf{T}} \boldsymbol{S}_- (\boldsymbol{S}_-^{\mathsf{T}} \boldsymbol{S}_-)^{-1}$.
- where $\boldsymbol{S}_+ := \begin{bmatrix} \boldsymbol{s}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{s}_{T+1}^{\mathsf{T}} \end{bmatrix}$, $\boldsymbol{S}_- := \begin{bmatrix} \tilde{\boldsymbol{s}}_1^{\mathsf{T}} \\ \vdots \\ \tilde{\boldsymbol{s}}_T^{\mathsf{T}} \end{bmatrix}$, $\boldsymbol{W} := \begin{bmatrix} \boldsymbol{w}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{w}_{T+1}^{\mathsf{T}} \end{bmatrix}$.

Learning Without Mixing

- Suppose $\{\boldsymbol{u}_t\}_{t=0}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{u}}^2 \boldsymbol{I}_m)$ and $\{\boldsymbol{w}_t\}_{t=1}^{\infty} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I}_n)$.
- Set the loss function $\hat{\mathcal{L}}(\boldsymbol{F}) = \frac{1}{2T} \sum_{t=1}^{T} \|\boldsymbol{s}_{t+1} - \boldsymbol{F}\tilde{\boldsymbol{s}}_t\|_{\ell_2}^2$.
- Over-determined problem, solution $\hat{\boldsymbol{F}} = \boldsymbol{S}_+^{\mathsf{T}} \boldsymbol{S}_- (\boldsymbol{S}_-^{\mathsf{T}} \boldsymbol{S}_-)^{-1}$.
- where $\boldsymbol{S}_+ := \begin{bmatrix} \boldsymbol{s}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{s}_{T+1}^{\mathsf{T}} \end{bmatrix}$, $\boldsymbol{S}_- := \begin{bmatrix} \tilde{\boldsymbol{s}}_1^{\mathsf{T}} \\ \vdots \\ \tilde{\boldsymbol{s}}_T^{\mathsf{T}} \end{bmatrix}$, $\boldsymbol{W} := \begin{bmatrix} \boldsymbol{w}_2^{\mathsf{T}} \\ \vdots \\ \boldsymbol{w}_{T+1}^{\mathsf{T}} \end{bmatrix}$.
- Estimation error $\|\hat{\boldsymbol{F}} - \boldsymbol{F}^{\star}\| \leq \frac{\|\boldsymbol{W}^{\mathsf{T}} \boldsymbol{S}_-\|}{\lambda_{\min}\left(\boldsymbol{S}_-^{\mathsf{T}} \boldsymbol{S}_-\right)}$.

Learning Without Mixing

Challenges:

Learning Without Mixing

Challenges:

- randomness in dynamical behavior

Learning Without Mixing

Challenges:

- randomness in dynamical behavior
- distribution of states

Learning Without Mixing

Challenges:

- randomness in dynamical behavior
- distribution of states
- temporal dependence
- finite samples

Stability in Bilinear Dynamical Systems

State equation: $\boldsymbol{s}_{t+1} = \boldsymbol{F}_0 \boldsymbol{x}_t + \sum_{k=1}^{m} \boldsymbol{u}_t[k] \boldsymbol{F}_k \boldsymbol{x}_t + \boldsymbol{w}_{t+1}$

Definition (Mean-square stability)

The bilinear system is mean-square stable (MSS) if there exists $\boldsymbol{s}_\infty \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}_\infty \in \mathbb{R}_+^{n \times n}$, such that for any initial state $\boldsymbol{x}_0$, as $t \to \infty$, we have

$$\|\mathbb{E}[\boldsymbol{s}_t] - \boldsymbol{s}_\infty\|_{\ell_2} \to 0, \quad \|\mathbb{E}[\boldsymbol{s}_t \boldsymbol{s}_t^\top] - \boldsymbol{\Sigma}_\infty\| \to 0. \tag{1}$$

Here the expectation is over the input sequence $\{\boldsymbol{u}_t\}_{t=0}^{\infty}$, the noise process $\{\boldsymbol{w}_t\}_{t=1}^{\infty}$ and the initial state $\boldsymbol{x}_0$. In the noise free case ($\boldsymbol{w}_t = 0$), we have $\boldsymbol{x}_\infty = 0$ and $\boldsymbol{\Sigma}_\infty = 0$.

Related works

1. C. Kubrusly and O. Costa, "Mean square stability conditions for discrete stochastic bilinear systems," IEEE Transactions on Automatic Control, vol. 30, no. 11, pp. 1082-1087, 1985.

BMSB Condition

- Does the random process $\{\tilde{\boldsymbol{s}}_t\}_{t\geq 1}$ satisfies the martingale small-ball condition?

BMSB Condition

- Does the random process $\{\tilde{\boldsymbol{s}}_t\}_{t\geq 1}$ satisfies the martingale small-ball condition?

Definition (Martingale small-ball)

- Let $\{\mathcal{F}_t\}_{t\geq 1}$ denotes a filtration.

BMSB Condition

- Does the random process $\{\tilde{\boldsymbol{s}}_t\}_{t\geq 1}$ satisfies the martingale small-ball condition?

Definition (Martingale small-ball)

- Let $\{\mathcal{F}_t\}_{t\geq 1}$ denotes a filtration.
- Let $\{Z_t\}_{t\geq 1}$ be $\{\mathcal{F}_t\}_{t\geq 1}$-adapted random process taking values in $\mathbb{R}$.

BMSB Condition

- Does the random process $\{\tilde{\boldsymbol{s}}_t\}_{t\geq 1}$ satisfies the martingale small-ball condition?

Definition (Martingale small-ball)

- Let $\{\mathcal{F}_t\}_{t\geq 1}$ denotes a filtration.
- Let $\{Z_t\}_{t\geq 1}$ be $\{\mathcal{F}_t\}_{t\geq 1}$-adapted random process taking values in $\mathbb{R}$.
- $\{Z_t\}_{t\geq 1}$ satisfies the (k, ν, p)-block martingale small-ball (BMSB) condition if, for any $j \geq 0$, one has $\frac{1}{k}\sum_{i=1}^{k} \mathbb{P}\left(|Z_{j+i}| \geq \nu \;\middle|\; \mathcal{F}_j\right) \geq p$ almost surely.

BMSB Condition

- Does the random process $\{\tilde{\boldsymbol{s}}_t\}_{t\geq 1}$ satisfies the martingale small-ball condition?

Definition (Martingale small-ball)

- Let $\{\mathcal{F}_t\}_{t\geq 1}$ denotes a filtration.
- Let $\{Z_t\}_{t\geq 1}$ be $\{\mathcal{F}_t\}_{t\geq 1}$-adapted random process taking values in $\mathbb{R}$.
- $\{Z_t\}_{t\geq 1}$ satisfies the (k, ν, p)-block martingale small-ball (BMSB) condition if, for any $j \geq 0$, one has $\frac{1}{k}\sum_{i=1}^{k} \mathbb{P}\left(|Z_{j+i}| \geq \nu \;\middle|\; \mathcal{F}_j\right) \geq p$ almost surely.
- A process $\{\boldsymbol{x}_t\}_{t\geq 1}$ taking values in $\mathbb{R}^d$ satisfies the $(k, \boldsymbol{\Gamma}_{\mathrm{sb}}, p)$-BMSB condition for $\boldsymbol{\Gamma}_{\mathrm{sb}} \succ 0$ if, for any fixed $\boldsymbol{v} \in \mathcal{S}^{d-1}$, the process $\{Z_t = \langle \boldsymbol{v}, \boldsymbol{x}_t \rangle\}_{t\geq 1}$ satisfies $(k, \sqrt{\boldsymbol{v}^\top \boldsymbol{\Gamma}_{\mathrm{sb}} \boldsymbol{v}}, p)$-BMSB.

Main Result

State equation: $\boldsymbol{s}_{t+1} = \boldsymbol{F}_0 \boldsymbol{x}_t + \sum_{k=1}^{m} \boldsymbol{u}_t[k] \boldsymbol{F}_k \boldsymbol{x}_t + \boldsymbol{w}_{t+1}$

Theorem (Bilinear system identification)

Fix $\delta \in (0, 1)$ and suppose we are given a single trajectory $(\boldsymbol{x}_t, \boldsymbol{u}_t)_{t=1}^{T}$ of the bilinear dynamical system.

Main Result

State equation: $\boldsymbol{s}_{t+1} = \boldsymbol{F}_0 \boldsymbol{x}_t + \sum_{k=1}^{m} \boldsymbol{u}_t[k] \boldsymbol{F}_k \boldsymbol{x}_t + \boldsymbol{w}_{t+1}$

Theorem (Bilinear system identification)

Fix $\delta \in (0, 1)$ and suppose we are given a single trajectory $(\boldsymbol{x}_t, \boldsymbol{u}_t)_{t=1}^{T}$ of the bilinear dynamical system.

- *Suppose the inputs and noise are Gaussian.*

Main Result

State equation: $\boldsymbol{s}_{t+1} = \boldsymbol{F}_0 \boldsymbol{x}_t + \sum_{k=1}^{m} \boldsymbol{u}_t[k] \boldsymbol{F}_k \boldsymbol{x}_t + \boldsymbol{w}_{t+1}$

Theorem (Bilinear system identification)

Fix $\delta \in (0, 1)$ and suppose we are given a single trajectory $(\boldsymbol{x}_t, \boldsymbol{u}_t)_{t=1}^{T}$ of the bilinear dynamical system.

- *Suppose the inputs and noise are Gaussian.*
- *Suppose the bilinear system is marginally mean-square stable.*

Main Result

State equation: $\boldsymbol{s}_{t+1} = \boldsymbol{F}_0 \boldsymbol{x}_t + \sum_{k=1}^{m} \boldsymbol{u}_t[k] \boldsymbol{F}_k \boldsymbol{x}_t + \boldsymbol{w}_{t+1}$

Theorem (Bilinear system identification)

Fix $\delta \in (0,1)$ and suppose we are given a single trajectory $(\boldsymbol{x}_t, \boldsymbol{u}_t)_{t=1}^{T}$ of the bilinear dynamical system.

- *Suppose the inputs and noise are Gaussian.*
- *Suppose the bilinear system is marginally mean-square stable.*
- *Suppose* $T \gtrsim n(m+1) + \log(12\Gamma/(\sigma_w^2 \delta)) + \log(3/\delta)$.

Main Result

State equation: $\boldsymbol{s}_{t+1} = \boldsymbol{F}_0 \boldsymbol{x}_t + \sum_{k=1}^{m} \boldsymbol{u}_t[k] \boldsymbol{F}_k \boldsymbol{x}_t + \boldsymbol{w}_{t+1}$

Theorem (Bilinear system identification)

Fix $\delta \in (0,1)$ and suppose we are given a single trajectory $(\boldsymbol{x}_t, \boldsymbol{u}_t)_{t=1}^{T}$ of the bilinear dynamical system.

- *Suppose the inputs and noise are Gaussian.*
- *Suppose the bilinear system is marginally mean-square stable.*
- *Suppose $T \gtrsim n(m+1) + \log(12\Gamma/(\sigma_w^2 \delta)) + \log(3/\delta)$.*
- *With probability at least $1-\delta$, we have*

$$\max \begin{Bmatrix} \|\hat{\boldsymbol{F}}_0 - \boldsymbol{F}_0\|, \\ \{\sigma_u \|\hat{\boldsymbol{F}}_k - \boldsymbol{F}_k\|\}_{k=1}^{m} \end{Bmatrix} \lesssim \sqrt{\frac{n(m+1) + \log(\frac{12\Gamma}{\sigma_w^2 \delta}) + \log(\frac{3}{\delta})}{T}}.$$

Thanks

Thank you for your attention!