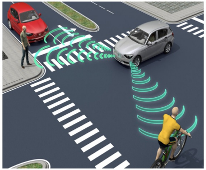
Non-asymptotic and Accurate Learning of Nonlinear Dynamical Systems

Yahya Sattar   Samet Oymak

Department of ECE, University of California Riverside

Motivation

- **Dynamical systems** appear in
  - physical systems,
  - reinforcement learning & control
  - natural language processing (i.e. RNN, LSTM)

- **Goal:** Efficient learning guarantees for nonlinear systems
- **Challenge:** spatio-temporal dependencies, nonlinear state equation, single trajectory . . .

State Equation

Nonlinear systems with state observations

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

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

Example: A nonlinear linear dynamical system
State equation: $\boldsymbol{h}_{t+1} = \phi(\boldsymbol{A}\boldsymbol{h}_t + \boldsymbol{B}\boldsymbol{u}_t) + \boldsymbol{w}_t$.

Learning from Finite Data

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

Learning from Finite Data

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

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

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

Definition (ρ-stabilized system)

(1) Pick inputs $\boldsymbol{u}_t = \boldsymbol{\pi}(\boldsymbol{h}_t) + \boldsymbol{z}_t$. 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 $\boldsymbol{h}_0 = \boldsymbol{\alpha}$ by $\boldsymbol{h}_t(\boldsymbol{\alpha})$.
(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

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

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

Nonlinear Stability

Nonlinear systems: Use ρ-stability

Definition (ρ-stabilized system)

(1) Pick inputs $\boldsymbol{u}_t = \boldsymbol{\pi}(\boldsymbol{h}_t) + \boldsymbol{z}_t$. 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 $\boldsymbol{h}_0 = \boldsymbol{\alpha}$ by $\boldsymbol{h}_t(\boldsymbol{\alpha})$.
(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

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

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

Key observation: System forgets the past quickly

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.

- a linear dynamical system is ρ-stable if spectral radius $\rho(\boldsymbol{A}_\star) < 1$.

Assumptions on the System and Inputs

Assumption (Stability)

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

- a linear dynamical system is ρ-stable if spectral radius $\rho(\boldsymbol{A}_\star) < 1$.
- a ρ-stable nonlinear system is a contractive system.

Assumptions on the System and Inputs

Assumption (Stability)

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

- a linear dynamical system is ρ-stable if spectral radius $\rho(\boldsymbol{A}_\star) < 1$.
- a ρ-stable nonlinear system is a contractive system.

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.

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 (Auxiliary Loss)

Suppose $\boldsymbol{h}_0 = 0$. 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,

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

Optimization Landscape

Assumption (One-point convexity & smoothness (OPCS))

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

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

Optimization Landscape

Assumption (One-point convexity & smoothness (OPCS))

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

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

- aka restricted secant inequality, and implies Polyak-Lojasiewicz condition.

Optimization Landscape

Assumption (One-point convexity & smoothness (OPCS))

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

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

- aka restricted secant inequality, and implies Polyak-Lojasiewicz condition.
- use OPC with one-point smoothness (rather than global smoothness).

Optimization Landscape

Assumption (One-point convexity & smoothness (OPCS))

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

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

- aka restricted secant inequality, and implies Polyak-Lojasiewicz condition.
- use OPC with one-point smoothness (rather than global smoothness).
- example, nonlinear state equation $\boldsymbol{h}_{t+1} = \phi(\boldsymbol{A}_\star \boldsymbol{h}_t + \boldsymbol{B}_\star \boldsymbol{z}_t) + \boldsymbol{w}_t$, with γ-increasing activation (i.e. $\phi'(x) \geq \gamma > 0$ for all $x \in \mathbb{R}$).

Main Result

Theorem (Main result – informal)

Suppose we run gradient descent algorithm, $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_{\tau} - \eta \nabla \hat{\mathcal{L}}(\boldsymbol{\theta}_{\tau})$ to solve the ERM problem. Suppose $T \gtrsim \frac{d}{\alpha^2(1-\rho)}$ and $r \gtrsim \frac{\sigma}{\alpha}\sqrt{\frac{d}{T(1-\rho)}}$. Under certain assumptions, the following statements hold with high probability over the trajectory.

Main Result

Theorem (Main result – informal)

Suppose we run gradient descent algorithm, $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_{\tau} - \eta\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}_{\tau})$ to solve the ERM problem. Suppose $T \gtrsim \frac{d}{\alpha^2(1-\rho)}$ and $r \gtrsim \frac{\sigma}{\alpha}\sqrt{\frac{d}{T(1-\rho)}}$. Under certain assumptions, the following statements hold with high probability over the trajectory.

- **Uniform convergence of gradient:** *For all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, $\nabla\hat{\mathcal{L}}(\boldsymbol{\theta})$ satisfies*

$$\|\nabla\hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \lesssim (\sigma + \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})\sqrt{\frac{d}{T(1-\rho)}}$$

Main Result

Theorem (Main result – informal)

Suppose we run gradient descent algorithm, $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta \nabla \hat{\mathcal{L}}(\boldsymbol{\theta}_\tau)$ to solve the ERM problem. Suppose $T \gtrsim \frac{d}{\alpha^2(1-\rho)}$ and $r \gtrsim \frac{\sigma}{\alpha}\sqrt{\frac{d}{T(1-\rho)}}$. Under certain assumptions, the following statements hold with high probability over the trajectory.

- **Uniform convergence of gradient:** *For all $\boldsymbol{\theta} \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$, $\nabla \hat{\mathcal{L}}(\boldsymbol{\theta})$ satisfies*
$$\|\nabla \hat{\mathcal{L}}(\boldsymbol{\theta}) - \nabla \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{\ell_2} \lesssim (\sigma + \|\boldsymbol{\theta} - \boldsymbol{\theta}_\star\|_{\ell_2})\sqrt{\frac{d}{T(1-\rho)}}$$
- **Convergence of gradient descent:** *Set the learning rate $\eta = \alpha/(16\beta^2)$ and fix $\boldsymbol{\theta}_0 \in \mathcal{B}^d(\boldsymbol{\theta}_\star, r)$. All gradient descent iterates $\boldsymbol{\theta}_\tau$ on $\hat{\mathcal{L}}(\boldsymbol{\theta})$ satisfy*
$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_\star\|_{\ell_2} \lesssim (1 - \frac{\alpha^2}{\beta^2})^\tau \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\star\|_{\ell_2} + \frac{\sigma}{\alpha}\sqrt{\frac{d}{T(1-\rho)}}.$$

Case Study

Entrywise nonlinearity: $\boxed{\boldsymbol{h}_{t+1} = \phi(\boldsymbol{A}_\star \boldsymbol{h}_t) + \boldsymbol{z}_t + \boldsymbol{w}_t}$

- $\boldsymbol{A}_\star = [\boldsymbol{a}_1^\star \ \cdots \ \boldsymbol{a}_n^\star]^T \in \mathbb{R}^{n\times n}$.
- Assume $\phi' \geq \gamma > 0$, $|\phi'|, |\phi''| \leq 1$ and $\phi(0) = 0$.

Theorem (simplified)

- *Suppose the system satisfies ρ-stability.*
- *Let $\boldsymbol{w}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ and $\boldsymbol{z}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$.*
- *Suppose trajectory length T obeys $T \gtrsim n \log(T)/(1-\rho)$*

With proper learning rate and the initialization $\boldsymbol{A}^{(0)} = 0$, all gradient descent iterates satisfy

$$\|\boldsymbol{a}_k^{(\tau)} - \boldsymbol{a}_k^\star\|_{\ell_2} \lesssim \left(1 - \frac{\gamma^4(1-\rho)^4}{C_\rho^4 n^2}\right)^\tau \|\boldsymbol{a}_k^{(0)} - \boldsymbol{a}_k^\star\|_{\ell_2} + \frac{\sigma}{\gamma^2}\sqrt{\frac{n}{T(1-\rho)}}.$$

Case Study

Linear Dynamical System: $\boxed{\boldsymbol{h}_{t+1} = \boldsymbol{A}_\star \boldsymbol{h}_t + \boldsymbol{B}_\star \boldsymbol{z}_t + \boldsymbol{w}_t}$

- $[\boldsymbol{A}_\star \ \boldsymbol{B}_\star] = [\boldsymbol{\theta}_1^\star \ \cdots \ \boldsymbol{\theta}_n^\star]^T \in \mathbb{R}^{n \times (n+p)}$.
- $\gamma_- := 1 \wedge \lambda_{\min}(\boldsymbol{\Gamma}_L^{\boldsymbol{B}_\star} + \sigma^2 \boldsymbol{\Gamma}_L)$, and $\gamma_+ := 1 \vee \lambda_{\max}(\boldsymbol{\Gamma}_L^{\boldsymbol{B}_\star} + \sigma^2 \boldsymbol{\Gamma}_L)$.

Theorem (simplified)

- *Suppose* $\rho(\boldsymbol{A}_\star) < 1$.
- *Let* $\boldsymbol{w}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ *and* $\boldsymbol{z}_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 \boldsymbol{I}_p)$.
- *Suppose trajectory length* T *obeys* $T \gtrsim (n+p)\log(T)/(1-\rho)$

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

$$\|\boldsymbol{\theta}_k^{(\tau)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} \lesssim \left(1 - \frac{\gamma_-^2}{\gamma_+^2}\right)^\tau \|\boldsymbol{\theta}_k^{(0)} - \boldsymbol{\theta}_k^\star\|_{\ell_2} + \frac{\sigma\sqrt{\gamma_+}}{\gamma_-}\sqrt{\frac{n+p}{T(1-\rho)}}.$$

Possible Extensions:

- Partial state observations $\boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{h}_t$
- NARMAX: $\boldsymbol{y}_{t+1} = \phi(\boldsymbol{y}_t, \ldots, \boldsymbol{y}_{t-T}, \boldsymbol{u}_t, \ldots, \boldsymbol{u}_{t-T}; \boldsymbol{\theta}_\star)$.
- Better dependence on spectral radius ρ (e.g. by using martingales)