Learning with little mixing

I. Ziemann (KTH) and Stephen Tu (Google)

(Appeared at NeurIPS'22)

Overview

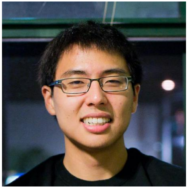
In collaboration

Overview

In collaboration
with Stephen Tu

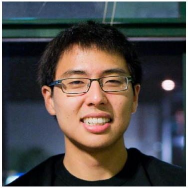
Overview

In collaboration
with Stephen Tu



Overview

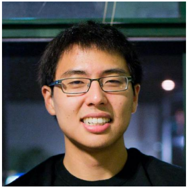
In collaboration
with Stephen Tu



@ Google Brain Robotics

Overview

In collaboration
with Stephen Tu

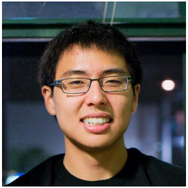


@ Google Brain Robotics

Outline:

Overview

In collaboration
with Stephen Tu



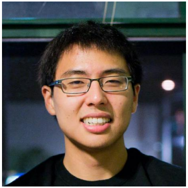
@ Google Brain Robotics

Outline:

Introduction

Overview

In collaboration
with Stephen Tu



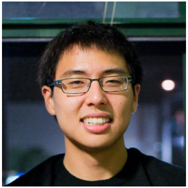
@ Google Brain Robotics

Outline:

- Introduction
- Lit Review

Overview

In collaboration
with Stephen Tu



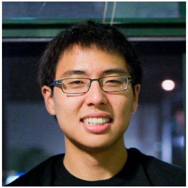
@ Google Brain Robotics

Outline:

- Introduction
- Lit Review
- Contribution

Overview

In collaboration
with Stephen Tu



@ Google Brain Robotics

Outline:

Introduction

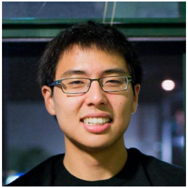
Lit Review

Contribution

Challenges & Proof Outline

Overview

In collaboration
with Stephen Tu



@ Google Brain Robotics

Outline:

Introduction

Lit Review

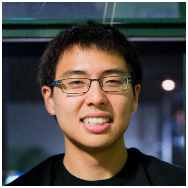
Contribution

Challenges & Proof Outline

Lower isometry

Overview

In collaboration
with Stephen Tu

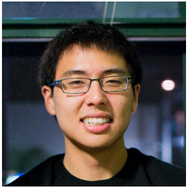


@ Google Brain Robotics

Outline:

Overview

In collaboration
with Stephen Tu

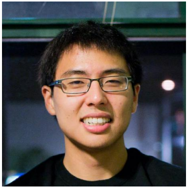


@ Google Brain Robotics

Outline:

Overview

In collaboration
with Stephen Tu



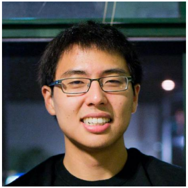
@ Google Brain Robotics

Outline:

- Introduction
- Lit Review
- Contribution
- Challenges & Proof Outline
  - Lower isometry
  - Localization
- Main Result
- Examples

Overview

In collaboration
with Stephen Tu



@ Google Brain Robotics

Outline:

- Introduction
- Lit Review
- Contribution
- Challenges & Proof Outline
  - Lower isometry
  - Localization
- Main Result
- Examples
- (An open problem ☕)

Introduction

What are the key properties dynamical (or control) systems need to possess for learning to be feasible?

Introduction

What are the key properties dynamical (or control) systems need to possess for learning to be feasible?

Relatively clear picture has emerged for perfectly observed **linear** dynamical systems ($X_{t+1} = A_\star X_t + W_t$):

Introduction

What are the key properties dynamical (or control) systems need to possess for learning to be feasible?

Relatively clear picture has emerged for perfectly observed **linear** dynamical systems ($X_{t+1} = A_\star X_t + W_t$):

Fazel et al. [2018]: Policy gradient methods converge

Introduction

What are the key properties dynamical (or control) systems need to possess for learning to be feasible?

Relatively clear picture has emerged for perfectly observed **linear** dynamical systems ($X_{t+1} = A_\star X_t + W_t$):

Fazel et al. [2018]: Policy gradient methods converge

Simchowitz et al. [2018]: Lack of mixing does not impede conv. for LDS

Introduction

What are the key properties dynamical (or control) systems need to possess for learning to be feasible?

Relatively clear picture has emerged for perfectly observed **linear** dynamical systems ($X_{t+1} = A_\star X_t + W_t$):

- Fazel et al. [2018]: Policy gradient methods converge
- Simchowitz et al. [2018]: Lack of mixing does not impede conv. for LDS
- Simchowitz and Foster [2020]: Optimal (dim) rates for LQR Regret

Introduction

What are the key properties dynamical (or control) systems need to possess for learning to be feasible?

Relatively clear picture has emerged for perfectly observed **linear** dynamical systems ($X_{t+1} = A_\star X_t + W_t$):

- Fazel et al. [2018]: Policy gradient methods converge
- Simchowitz et al. [2018]: Lack of mixing does not impede conv. for LDS
- Simchowitz and Foster [2020]: Optimal (dim) rates for LQR Regret
- Tsiamis et al. [2022a]: Exponential hardness results for LQR regret

Introduction

What are the key properties dynamical (or control) systems need to possess for learning to be feasible?

Relatively clear picture has emerged for perfectly observed **linear** dynamical systems ($X_{t+1} = A_\star X_t + W_t$):

- Fazel et al. [2018]: Policy gradient methods converge
- Simchowitz et al. [2018]: Lack of mixing does not impede conv. for LDS
- Simchowitz and Foster [2020]: Optimal (dim) rates for LQR Regret
- Tsiamis et al. [2022a]: Exponential hardness results for LQR regret

Would like to pursue more realistic models!

Introduction

What are the key properties dynamical (or control) systems need to possess for learning to be feasible?

Relatively clear picture has emerged for perfectly observed **linear** dynamical systems ($X_{t+1} = A_\star X_t + W_t$):

Fazel et al. [2018]: Policy gradient methods converge

Simchowitz et al. [2018]: Lack of mixing does not impede conv. for LDS

Simchowitz and Foster [2020]: Optimal (dim) rates for LQR Regret

Tsiamis et al. [2022a]: Exponential hardness results for LQR regret

Would like to pursue more realistic models!

Today: discuss the above question in terms of **nonlinear** time-series

Introduction

What are the key properties dynamical (or control) systems need to possess for learning to be feasible?

Relatively clear picture has emerged for perfectly observed **linear** dynamical systems ($X_{t+1} = A_\star X_t + W_t$):

Fazel et al. [2018]: Policy gradient methods converge

Simchowitz et al. [2018]: Lack of mixing does not impede conv. for LDS

Simchowitz and Foster [2020]: Optimal (dim) rates for LQR Regret

Tsiamis et al. [2022a]: Exponential hardness results for LQR regret

Would like to pursue more realistic models!

Today: discuss the above question in terms of **nonlinear** time-series

Q: What is the effect of mixing on the rate of convergence of the ERM?

Lit Review

Tsiamis et al. [2022b]: Recent survey in the linear setting
https://arxiv.org/abs/2209.05423

Statistical Learning Theory for Control

A FINITE SAMPLE PERSPECTIVE

Anastasios Tsiamis*, Ingvar Ziemann*, Nikolai Matni, and George J. Pappas
A. Tsiamis (atsiamis@control.ee.ethz.ch) is with the Dept. of Information Technology and Electrical Engineering, ETH Zürich, Zürich, Switzerland.
I. Ziemann (ziemann@kth.se) is with the Division of Decision and Control Systems, KTH Royal Institute of Technology, Stockholm, Sweden.
N. Matni (nmatni@seas.upenn.edu) and G. J. Pappas (pappasg@seas.upenn.edu) are with the Dept. of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, USA.
*Both authors contributed equally.

Lit Review: Learning without mixing in LDS

For linear dynamical systems

$$X_{t+1} = A_\star X_t + W_t \qquad \Gamma_k \triangleq \sum_{t=0}^{k} A^t (A^t)^\top \qquad \rho(A_\star) \leq 1 \tag{1}$$

Lit Review: Learning without mixing in LDS

For linear dynamical systems

$$X_{t+1} = A_\star X_t + W_t \qquad \Gamma_k \triangleq \sum_{t=0}^{k} A^t (A^t)^\mathsf{T} \qquad \rho(A_\star) \leq 1 \qquad (1)$$

Simchowitz et al. [2018] have shown that ERM satisfies

$$\|\widehat{A} - A_\star\|_{\mathrm{op}} \lesssim \sqrt{\frac{d_\mathsf{X} \log(d_\mathsf{X}/\delta) + \log\det(\Gamma_T \Gamma_k^{-1})}{T \lambda_{\min}(\Gamma_k)}} \qquad (2)$$

Lit Review: Learning without mixing in LDS

For linear dynamical systems

$$X_{t+1} = A_\star X_t + W_t \qquad \Gamma_k \triangleq \sum_{t=0}^{k} A^t (A^t)^\mathsf{T} \qquad \rho(A_\star) \le 1 \tag{1}$$

Simchowitz et al. [2018] have shown that ERM satisfies

$$\|\widehat{A} - A_\star\|_{\mathrm{op}} \lesssim \sqrt{\frac{d_{\mathsf{X}} \log(d_{\mathsf{X}}/\delta) + \log\det(\Gamma_T \Gamma_k^{-1})}{T \lambda_{\min}(\Gamma_k)}} \tag{2}$$

Takeaway: dependence does not impede convergence in LDS

Lit Review: Learning without mixing in LDS

For linear dynamical systems

$$X_{t+1} = A_\star X_t + W_t \qquad \Gamma_k \triangleq \sum_{t=0}^{k} A^t (A^t)^\top \qquad \rho(A_\star) \leq 1 \tag{1}$$

Simchowitz et al. [2018] have shown that ERM satisfies

$$\|\widehat{A} - A_\star\|_{\mathrm{op}} \lesssim \sqrt{\frac{d_{\mathsf{X}} \log(d_{\mathsf{X}}/\delta) + \log\det(\Gamma_T \Gamma_k^{-1})}{T \lambda_{\min}(\Gamma_k)}} \tag{2}$$

Takeaway: dependence does not impede convergence in LDS

Lit Review: Learning without mixing in LDS

For linear dynamical systems

$$X_{t+1} = A_\star X_t + W_t \qquad \Gamma_k \triangleq \sum_{t=0}^{k} A^t (A^t)^\top \qquad \rho(A_\star) \leq 1 \tag{1}$$

Simchowitz et al. [2018] have shown that ERM satisfies

$$\|\widehat{A} - A_\star\|_{\mathrm{op}} \lesssim \sqrt{\frac{d_X \log(d_X/\delta) + \log\det(\Gamma_T \Gamma_k^{-1})}{T \lambda_{\min}(\Gamma_k)}} \tag{2}$$

Takeaway: dependence does not impede convergence in LDS

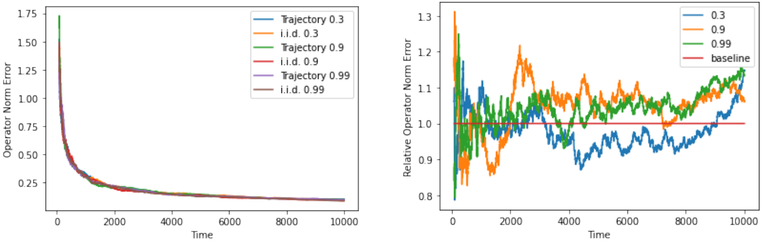


Figure: The spectral radius of the matrix $A_\star$ has (almost) no impact on the rate of convergence; $\rho(A_\star) \in \{0.3, 0.9, 0.99\}$ and $\sigma_{\min}(A_\star) \approx 0$

Lit Review: beyond perfectly observed LDS?

Nagaraj et al. [2020]: mix. dep. burn-in unavoidable in the worst case

Lit Review: beyond perfectly observed LDS?

Nagaraj et al. [2020]: mix. dep. burn-in unavoidable in the worst case

GLM: $X_{t+1} = \phi(A_\star X_t) + W_t$

Lit Review: beyond perfectly observed LDS?

Nagaraj et al. [2020]: mix. dep. burn-in unavoidable in the worst case

GLM: $X_{t+1} = \phi(A_\star X_t) + W_t$

rates with mixing: Sattar and Oymak [2022], Foster et al. [2020]

Lit Review: beyond perfectly observed LDS?

Nagaraj et al. [2020]: mix. dep. burn-in unavoidable in the worst case

GLM: $X_{t+1} = \phi(A_\star X_t) + W_t$

rates with mixing: Sattar and Oymak [2022], Foster et al. [2020]

rates without mixing: Kowshik et al. [2021]

Lit Review: beyond perfectly observed LDS?

Nagaraj et al. [2020]: mix. dep. burn-in unavoidable in the worst case

GLM: $X_{t+1} = \phi(A_\star X_t) + W_t$

rates with mixing: Sattar and Oymak [2022], Foster et al. [2020]

rates without mixing: Kowshik et al. [2021]

expansive link fcn, SGD achieves: $\|\widehat{A} - A_\star\|_F^2 = \tilde{O}(d_X^2/(T\lambda_{\min}(\Sigma_W)))$

Lit Review: beyond perfectly observed LDS?

Nagaraj et al. [2020]: mix. dep. burn-in unavoidable in the worst case

GLM: $X_{t+1} = \phi(A_\star X_t) + W_t$

rates with mixing: Sattar and Oymak [2022], Foster et al. [2020]

rates without mixing: Kowshik et al. [2021]

expansive link fcn, SGD achieves: $\|\widehat{A} - A_\star\|_F^2 = \tilde{O}(d_{\mathsf{X}}^2/(T\lambda_{\min}(\Sigma_W)))$

general nonlinear rates w/ mixing: Roy et al. [2021], Ziemann et al. [2022]

Lit Review: beyond perfectly observed LDS?

Nagaraj et al. [2020]: mix. dep. burn-in unavoidable in the worst case

GLM: $X_{t+1} = \phi(A_\star X_t) + W_t$

rates with mixing: Sattar and Oymak [2022], Foster et al. [2020]

rates without mixing: Kowshik et al. [2021]

expansive link fcn, SGD achieves: $\|\widehat{A} - A_\star\|_F^2 = \tilde{O}(d_{\mathsf{X}}^2/(T\lambda_{\min}(\Sigma_W)))$

general nonlinear rates w/ mixing: Roy et al. [2021], Ziemann et al. [2022]

Bilinear dyn. sys. rates without mixing: Sattar et al. [2022]

Lit Review: beyond perfectly observed LDS?

Nagaraj et al. [2020]: mix. dep. burn-in unavoidable in the worst case

GLM: $X_{t+1} = \phi(A_\star X_t) + W_t$

rates with mixing: Sattar and Oymak [2022], Foster et al. [2020]

rates without mixing: Kowshik et al. [2021]

expansive link fcn, SGD achieves: $\|\widehat{A} - A_\star\|_F^2 = \tilde{O}(d_X^2/(T\lambda_{\min}(\Sigma_W)))$

general nonlinear rates w/ mixing: Roy et al. [2021], Ziemann et al. [2022]

Bilinear dyn. sys. rates without mixing: Sattar et al. [2022]

PO & MJS: Oymak and Ozay [2019], Tsiamis and Pappas [2019], Sarkar and Rakhlin [2019], Lee [2022], Djehiche and Mazhar [2022], Sun et al. [2022], Sattar et al. [2021]

Lit Review: beyond perfectly observed LDS?

Nagaraj et al. [2020]: mix. dep. burn-in unavoidable in the worst case

GLM: $X_{t+1} = \phi(A_\star X_t) + W_t$

rates with mixing: Sattar and Oymak [2022], Foster et al. [2020]

rates without mixing: Kowshik et al. [2021]

expansive link fcn, SGD achieves: $\|\widehat{A} - A_\star\|_F^2 = \tilde{O}(d_X^2/(T\lambda_{\min}(\Sigma_W)))$

general nonlinear rates w/ mixing: Roy et al. [2021], Ziemann et al. [2022]

Bilinear dyn. sys. rates without mixing: Sattar et al. [2022]

PO & MJS: Oymak and Ozay [2019], Tsiamis and Pappas [2019], Sarkar and Rakhlin [2019], Lee [2022], Djehiche and Mazhar [2022], Sun et al. [2022], Sattar et al. [2021]

Lit Review: beyond perfectly observed LDS?

Nagaraj et al. [2020]: mix. dep. burn-in unavoidable in the worst case

GLM: $X_{t+1} = \phi(A_\star X_t) + W_t$

rates with mixing: Sattar and Oymak [2022], Foster et al. [2020]

rates without mixing: Kowshik et al. [2021]

expansive link fcn, SGD achieves: $\|\widehat{A} - A_\star\|_F^2 = \tilde{O}(d_X^2/(T\lambda_{\min}(\Sigma_W)))$

general nonlinear rates w/ mixing: Roy et al. [2021], Ziemann et al. [2022]

Bilinear dyn. sys. rates without mixing: Sattar et al. [2022]

PO & MJS: Oymak and Ozay [2019], Tsiamis and Pappas [2019], Sarkar and Rakhlin [2019], Lee [2022], Djehiche and Mazhar [2022], Sun et al. [2022], Sattar et al. [2021]

two types of rates: iid rate or iid rate × dependency deflation

Q: when do we get the iid rate?

Problem Formulation

Interested in nonlinear time-series / dynamical system ($Y_t = X_{t+1}$)

$$\underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\cap}{Y_t}} = f_\star(\underset{\mathsf{X}\subset\mathbb{R}^{d_\mathsf{X}}}{\underset{\cap}{X_t}}) + \underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\cap}{W_t}} \qquad f_\star \in \mathscr{F} \tag{3}$$

Problem Formulation

Interested in nonlinear time-series / dynamical system ($Y_t = X_{t+1}$)

$$\underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\cap}{Y_t}} = f_\star(\underset{\mathsf{X}\subset\mathbb{R}^{d_\mathsf{X}}}{\underset{\cap}{X_t}}) + \underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\cap}{W_t}} \qquad f_\star \in \mathscr{F} \tag{3}$$

$\mathscr{F}$: hypothesis class of functions

Problem Formulation

Interested in nonlinear time-series / dynamical system ($Y_t = X_{t+1}$)

$$\underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\cap}{Y_t}} = f_\star(\underset{\mathsf{X}\subset\mathbb{R}^{d_\mathsf{X}}}{\underset{\cap}{X_t}}) + \underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\cap}{W_t}} \qquad f_\star \in \mathscr{F} \tag{3}$$

$\mathscr{F}$: hypothesis class of functions

$\mathscr{F}_\star \triangleq \mathscr{F} - \{f_\star\}$ "shifted/centered class"

Problem Formulation

Interested in nonlinear time-series / dynamical system ($Y_t = X_{t+1}$)

$$\underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\Cap}{Y_t}} = f_\star(\underset{\mathsf{X}\subset\mathbb{R}^{d_\mathsf{X}}}{\underset{\Cap}{X_t}}) + \underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\Cap}{W_t}} \qquad f_\star \in \mathscr{F} \tag{3}$$

$\mathscr{F}$: hypothesis class of functions

$\mathscr{F}_\star \triangleq \mathscr{F} - \{f_\star\}$ "shifted/centered class"

$(X_t)_{t=0}^{T-1} \sim \mathsf{P}_X$: covariate process

Problem Formulation

Interested in nonlinear time-series / dynamical system ($Y_t = X_{t+1}$)

$$\underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\Cap}{Y_t}} = f_\star(\underset{\mathsf{X}\subset\mathbb{R}^{d_\mathsf{X}}}{\underset{\Cap}{X_t}}) + \underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\Cap}{W_t}} \qquad f_\star \in \mathscr{F} \tag{3}$$

$\mathscr{F}$: hypothesis class of functions

$\mathscr{F}_\star \triangleq \mathscr{F} - \{f_\star\}$ "shifted/centered class"

$(X_t)_{t=0}^{T-1} \sim \mathsf{P}_X$: covariate process

$(X_t, Y_t)_{t=0}^{T-1}$: data available to the learner

Problem Formulation

Interested in nonlinear time-series / dynamical system ($Y_t = X_{t+1}$)

$$\underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\cap}{Y_t}} = f_\star(\underset{\mathsf{X}\subset\mathbb{R}^{d_\mathsf{X}}}{\underset{\cap}{X_t}}) + \underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\cap}{W_t}} \qquad f_\star \in \mathscr{F} \tag{3}$$

$\mathscr{F}$: hypothesis class of functions

$\mathscr{F}_\star \triangleq \mathscr{F} - \{f_\star\}$ "shifted/centered class"

$(X_t)_{t=0}^{T-1} \sim \mathsf{P}_X$: covariate process

$(X_t, Y_t)_{t=0}^{T-1}$: data available to the learner

$(W_t)_{t=0}^{T-1}$: martingale difference noise

Problem Formulation

Interested in nonlinear time-series / dynamical system ($Y_t = X_{t+1}$)

$$\underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\Cap}{Y_t}} = f_\star(\underset{\mathsf{X}\subset\mathbb{R}^{d_\mathsf{X}}}{\underset{\Cap}{X_t}}) + \underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\Cap}{W_t}} \qquad f_\star \in \mathscr{F} \tag{3}$$

$\mathscr{F}$: hypothesis class of functions

$\mathscr{F}_\star \triangleq \mathscr{F} - \{f_\star\}$ "shifted/centered class"

$(X_t)_{t=0}^{T-1} \sim \mathsf{P}_X$: covariate process

$(X_t, Y_t)_{t=0}^{T-1}$: data available to the learner

$(W_t)_{t=0}^{T-1}$: martingale difference noise

Interested in the performance of ERM:

$$\widehat{f} \in \operatorname{argmin}_{f\in\mathscr{F}} \sum_{t=0}^{T-1} \|Y_t - f(X_t)\|_2^2$$

Problem Formulation

Interested in nonlinear time-series / dynamical system ($Y_t = X_{t+1}$)

$$\underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\cap}{Y_t}} = f_\star(\underset{\mathsf{X}\subset\mathbb{R}^{d_\mathsf{X}}}{\underset{\cap}{X_t}}) + \underset{\mathsf{Y}\subset\mathbb{R}^{d_\mathsf{Y}}}{\underset{\cap}{W_t}} \qquad f_\star \in \mathscr{F} \tag{3}$$

$\mathscr{F}$: hypothesis class of functions

$\mathscr{F}_\star \triangleq \mathscr{F} - \{f_\star\}$ "shifted/centered class"

$(X_t)_{t=0}^{T-1} \sim \mathsf{P}_X$: covariate process

$(X_t, Y_t)_{t=0}^{T-1}$: data available to the learner

$(W_t)_{t=0}^{T-1}$: martingale difference noise

Interested in the performance of ERM:

$$\widehat{f} \in \operatorname{argmin}_{f\in\mathscr{F}} \sum_{t=0}^{T-1} \|Y_t - f(X_t)\|_2^2$$

in terms of square-loss excess risk:

$$\|f - f_\star\|_{L^2}^2 \triangleq \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t) - f_\star(X_t)\|_2^2 \qquad (f \in \mathscr{F})$$

Contribution

Study ERM under two assumptions

Contribution

Study ERM under two assumptions

A1. Trajectory Hypercontractivity (identifiability/small-ball)

Contribution

Study ERM under two assumptions

A1. Trajectory Hypercontractivity (identifiability/small-ball)

A2. Mixing

Contribution

Study ERM under two assumptions

A1. Trajectory Hypercontractivity (identifiability/small-ball)

A2. Mixing

Main result: Informally, under A1-A2, ERM $\widehat{f}$ satisfies:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \lesssim \left(\frac{\text{dimensional factors} \times \sigma_W^2}{T}\right)^{\mathrm{comp}(\mathscr{F})} + \text{higher order } o(t_{\mathrm{mix}}/T^{\mathrm{comp}(\mathscr{F})}) \text{ terms} \quad (4)$$

Contribution

Study ERM under two assumptions

A1. Trajectory Hypercontractivity (identifiability/small-ball)

A2. Mixing

Main result: Informally, under A1-A2, ERM $\widehat{f}$ satisfies:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \lesssim \left(\frac{\text{dimensional factors} \times \sigma_W^2}{T}\right)^{\text{comp}(\mathscr{F})} + \text{higher order } o(t_{\text{mix}}/T^{\text{comp}(\mathscr{F})}) \text{ terms} \quad (4)$$

comp($\mathscr{F}$): (inverse) measure of complexity

Contribution

Study ERM under two assumptions

A1. Trajectory Hypercontractivity (identifiability/small-ball)

A2. Mixing

Main result: Informally, under A1-A2, ERM $\widehat{f}$ satisfies:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \lesssim \left(\frac{\text{dimensional factors} \times \sigma_W^2}{T}\right)^{\mathrm{comp}(\mathscr{F})} + \text{higher order } o(t_{\mathrm{mix}}/T^{\mathrm{comp}(\mathscr{F})}) \text{ terms} \quad (4)$$

$\mathrm{comp}(\mathscr{F})$: (inverse) measure of complexity

Takeaway: after a burn-in, slow mixing does not impede convergence for a large class of problems

Contribution

Study ERM under two assumptions

A1. Trajectory Hypercontractivity (identifiability/small-ball)

A2. Mixing

Main result: Informally, under A1-A2, ERM $\widehat{f}$ satisfies:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \lesssim \left(\frac{\text{dimensional factors} \times \sigma_W^2}{T}\right)^{\mathrm{comp}(\mathscr{F})} + \text{higher order } o(t_{\mathrm{mix}}/T^{\mathrm{comp}(\mathscr{F})}) \text{ terms} \quad (4)$$

comp($\mathscr{F}$): (inverse) measure of complexity

Takeaway: after a burn-in, slow mixing does not impede convergence for a large class of problems

$\Rightarrow$ we match the iid rate

Contribution

Study ERM under two assumptions

A1. Trajectory Hypercontractivity (identifiability/small-ball)

A2. Mixing

Main result: Informally, under A1-A2, ERM $\widehat{f}$ satisfies:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \lesssim \left(\frac{\text{dimensional factors} \times \sigma_W^2}{T}\right)^{\text{comp}(\mathscr{F})} + \text{higher order } o(t_{\text{mix}}/T^{\text{comp}(\mathscr{F})}) \text{ terms} \quad (4)$$

comp($\mathscr{F}$): (inverse) measure of complexity

Takeaway: after a burn-in, slow mixing does not impede convergence for a large class of problems

$\Rightarrow$ we match the iid rate

Examples: LDS, GLM, RKHS, finite hyp. classes, ergodic finite state MC

So how do we get there?

Two Technical Challenges to Overcome

Notation LDS: $X_{t+1} = A_\star X_t + W_t$ $f(x) = Ax$

Two Technical Challenges to Overcome

Notation LDS: $X_{t+1} = A_\star X_t + W_t$ $f(x) = Ax$

First, in the linear setting we have

$$\widehat{A} - A_\star = \left(\sum_{t=0}^{T-1} W_t X_t^\top\right)\left(\sum_{t=0}^{T-1} X_t X_t^\top\right)^\dagger \tag{5}$$

Two Technical Challenges to Overcome

Notation LDS: $X_{t+1} = A_\star X_t + W_t$ $f(x) = Ax$

First, in the linear setting we have

$$\widehat{A} - A_\star = \left(\sum_{t=0}^{T-1} W_t X_t^\top \right) \left(\sum_{t=0}^{T-1} X_t X_t^\top \right)^\dagger \tag{5}$$

Can be controlled by self-normalized martingale bound [Abbasi-Yadkori and Szepesvári, 2011]

Two Technical Challenges to Overcome

Notation LDS: $X_{t+1} = A_\star X_t + W_t$ $f(x) = Ax$

First, in the linear setting we have

$$\widehat{A} - A_\star = \left(\sum_{t=0}^{T-1} W_t X_t^\top\right)\left(\sum_{t=0}^{T-1} X_t X_t^\top\right)^\dagger \tag{5}$$

Can be controlled by self-normalized martingale bound [Abbasi-Yadkori and Szepesvári, 2011]

(5) does not hold beyond linear classes ☹

Two Technical Challenges to Overcome

Notation LDS: $X_{t+1} = A_\star X_t + W_t$ $f(x) = Ax$

First, in the linear setting we have

$$\widehat{A} - A_\star = \left(\sum_{t=0}^{T-1} W_t X_t^\top\right)\left(\sum_{t=0}^{T-1} X_t X_t^\top\right)^\dagger \tag{5}$$

Can be controlled by self-normalized martingale bound [Abbasi-Yadkori and Szepesvári, 2011]

(5) does not hold beyond linear classes ☹

⇒ Challenge: a nonlinear *localization* analogue of (5) is needed

Two Technical Challenges to Overcome

Notation LDS: $X_{t+1} = A_\star X_t + W_t$ $\quad f(x) = Ax$

First, in the linear setting we have

$$\widehat{A} - A_\star = \left(\sum_{t=0}^{T-1} W_t X_t^\mathsf{T} \right) \left(\sum_{t=0}^{T-1} X_t X_t^\mathsf{T} \right)^\dagger \tag{5}$$

Can be controlled by self-normalized martingale bound [Abbasi-Yadkori and Szepesvári, 2011]

(5) does not hold beyond linear classes ☹

⇒ Challenge: a nonlinear *localization* analogue of (5) is needed

Second, in the LDS setting, can adapt Mendelson [2014] to control

$$\lambda_{\min} \left(\sum_{t=0}^{T-1} X_t X_t^\mathsf{T} \right) \gtrsim \lambda_{\min} \left(\mathbf{E} \sum_{t=0}^{T-1} X_t X_t^\mathsf{T} \right) \qquad (w.h.p.) \tag{6}$$

Two Technical Challenges to Overcome

Notation LDS: $X_{t+1} = A_\star X_t + W_t \qquad f(x) = Ax$

First, in the linear setting we have

$$\widehat{A} - A_\star = \left(\sum_{t=0}^{T-1} W_t X_t^\mathsf{T} \right) \left(\sum_{t=0}^{T-1} X_t X_t^\mathsf{T} \right)^\dagger \tag{5}$$

Can be controlled by self-normalized martingale bound [Abbasi-Yadkori and Szepesvári, 2011]

(5) does not hold beyond linear classes ☹

$\Rightarrow$ Challenge: a nonlinear *localization* analogue of (5) is needed

Second, in the LDS setting, can adapt Mendelson [2014] to control

$$\lambda_{\min} \left(\sum_{t=0}^{T-1} X_t X_t^\mathsf{T} \right) \gtrsim \lambda_{\min} \left(\mathbf{E} \sum_{t=0}^{T-1} X_t X_t^\mathsf{T} \right) \qquad (w.h.p.) \tag{6}$$

$\Rightarrow$ Challenge: we also require a nonlinear *lower-isometry* analogue of (6)

High-Level Proof Strategy

First challenge: Prove a high probability lower isometry result

$$\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)-f_\star(X_t)\|_2^2 \gtrsim \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)-f_\star(X_t)\|_2^2 \qquad (\text{unif.}\forall f\in\mathscr{F}) \tag{7}$$

High-Level Proof Strategy

First challenge: Prove a high probability lower isometry result

$$\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t) - f_\star(X_t)\|_2^2 \gtrsim \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t) - f_\star(X_t)\|_2^2 \qquad (\text{unif.} \forall f \in \mathscr{F}) \tag{7}$$

use mixing + "small-ball" (traj. hyp.)

High-Level Proof Strategy

First challenge: Prove a high probability lower isometry result

$$\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)-f_\star(X_t)\|_2^2 \gtrsim \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)-f_\star(X_t)\|_2^2 \qquad (\text{unif.}\forall f \in \mathscr{F}) \tag{7}$$

use mixing + "small-ball" (traj. hyp.)

insight from Mendelson [2014]: lower isometry w/ small-ball is cheap (only affects burn-in) so we can use some mixing here w/o affecting the rate ☺

High-Level Proof Strategy

First challenge: Prove a high probability lower isometry result

$$\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)-f_\star(X_t)\|_2^2 \gtrsim \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)-f_\star(X_t)\|_2^2 \qquad (\text{unif.}\forall f\in\mathscr{F}) \tag{7}$$

use mixing + "small-ball" (traj. hyp.)

insight from Mendelson [2014]: lower isometry w/ small-ball is cheap (only affects burn-in) so we can use some mixing here w/o affecting the rate ☺

Second challenge: Combine with *offset* basic ineq [Liang et al., 2015]:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\widehat{f}(X_t)-f_\star(X_t)\|_2^2 \le \underbrace{\sup_{f\in\mathscr{F}_\star}\frac{1}{T}\sum_{t=0}^{T-1}4\langle W_t, f(X_t)\rangle-\|f(X_t)\|_2^2}_{\triangleq \mathsf{M}_T(\mathscr{F}_\star)\text{ "martingale offset complexity"}} \tag{8}$$

High-Level Proof Strategy

First challenge: Prove a high probability lower isometry result

$$\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)-f_\star(X_t)\|_2^2 \gtrsim \frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)-f_\star(X_t)\|_2^2 \qquad (\text{unif.}\forall f\in\mathscr{F}) \tag{7}$$

use mixing + "small-ball" (traj. hyp.)

insight from Mendelson [2014]: lower isometry w/ small-ball is cheap (only affects burn-in) so we can use some mixing here w/o affecting the rate ☺

Second challenge: Combine with *offset* basic ineq [Liang et al., 2015]:

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\widehat{f}(X_t)-f_\star(X_t)\|_2^2 \le \underbrace{\sup_{f\in\mathscr{F}_\star}\frac{1}{T}\sum_{t=0}^{T-1}4\langle W_t, f(X_t)\rangle-\|f(X_t)\|_2^2}_{\triangleq \mathsf{M}_T(\mathscr{F}_\star)\text{ "martingale offset complexity"}} \tag{8}$$

Quadratic penalization in (8) gives free localization/self-normalization ☺

Localization: Martingale Offset Complexity

combining (7) and (8): $\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|\widehat{f}(X_t) - f_\star(X_t)\|_2^2 \lesssim \mathsf{M}_T(\mathscr{F}_\star)$

Localization: Martingale Offset Complexity

combining (7) and (8): $\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|\widehat{f}(X_t) - f_\star(X_t)\|_2^2 \lesssim \mathsf{M}_T(\mathscr{F}_\star)$

Definition (Martingale Offset Complexity [Liang et al., 2015])

$$\mathsf{M}_T(\mathscr{F}_\star) \triangleq \sup_{f\in\mathscr{F}_\star} \frac{1}{T}\sum_{t=0}^{T-1} 4\langle W_t, f(X_t)\rangle - \|f(X_t)\|_2^2$$

Localization: Martingale Offset Complexity

combining (7) and (8): $\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|\widehat{f}(X_t) - f_\star(X_t)\|_2^2 \lesssim \mathsf{M}_T(\mathscr{F}_\star)$

Definition (Martingale Offset Complexity [Liang et al., 2015])

$$\mathsf{M}_T(\mathscr{F}_\star) \triangleq \sup_{f\in\mathscr{F}_\star} \frac{1}{T}\sum_{t=0}^{T-1} 4\langle W_t, f(X_t)\rangle - \|f(X_t)\|_2^2$$

$\Rightarrow$ do not pay complexity for $\mathscr{F}$ but only for those hypotheses near $f_\star$

Localization: Martingale Offset Complexity

combining (7) and (8): $\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|\widehat{f}(X_t) - f_\star(X_t)\|_2^2 \lesssim \mathsf{M}_T(\mathscr{F}_\star)$

Definition (Martingale Offset Complexity [Liang et al., 2015])

$$\mathsf{M}_T(\mathscr{F}_\star) \triangleq \sup_{f\in\mathscr{F}_\star} \frac{1}{T}\sum_{t=0}^{T-1} 4\langle W_t, f(X_t)\rangle - \|f(X_t)\|_2^2$$

$\Rightarrow$ do not pay complexity for $\mathscr{F}$ but only for those hypotheses near $f_\star$

Behaves like a *local* complexity

Localization: Martingale Offset Complexity

combining (7) and (8): $\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{E}\|\widehat{f}(X_t) - f_\star(X_t)\|_2^2 \lesssim \mathsf{M}_T(\mathscr{F}_\star)$

Definition (Martingale Offset Complexity [Liang et al., 2015])

$$\mathsf{M}_T(\mathscr{F}_\star) \triangleq \sup_{f \in \mathscr{F}_\star} \frac{1}{T}\sum_{t=0}^{T-1} 4\langle W_t, f(X_t)\rangle - \|f(X_t)\|_2^2$$

$\Rightarrow$ do not pay complexity for $\mathscr{F}$ but only for those hypotheses near $f_\star$

Behaves like a *local* complexity

$\Rightarrow$ $\mathsf{M}_T(\mathscr{F}_\star)$ reduces to self-normalized martingale for linear hyp.

Localization: Martingale Offset Complexity

combining (7) and (8): $\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|\widehat{f}(X_t) - f_\star(X_t)\|_2^2 \lesssim \mathsf{M}_T(\mathscr{F}_\star)$

Definition (Martingale Offset Complexity [Liang et al., 2015])

$$\mathsf{M}_T(\mathscr{F}_\star) \triangleq \sup_{f \in \mathscr{F}_\star} \frac{1}{T}\sum_{t=0}^{T-1} 4\langle W_t, f(X_t)\rangle - \|f(X_t)\|_2^2$$

$\Rightarrow$ do not pay complexity for $\mathscr{F}$ but only for those hypotheses near $f_\star$

Behaves like a *local* complexity

$\Rightarrow$ $\mathsf{M}_T(\mathscr{F}_\star)$ reduces to self-normalized martingale for linear hyp.

$\Rightarrow$ $\mathsf{M}_T(\mathscr{F}_\star)$ can be bounded by chaining to give $\tilde{O}$(iid rate):

$$\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) \lesssim \inf_{\gamma>0}\left\{\frac{\sigma_W^2 \log \mathcal{N}_\infty(\mathscr{F}_\star, \gamma)}{T} + \frac{\sigma_W}{\sqrt{T}}\int_0^\gamma \sqrt{\log \mathcal{N}_\infty(\mathscr{F}_\star, s)}ds + \gamma^2\right\}.$$

Localization: Martingale Offset Complexity

combining (7) and (8): $\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|\widehat{f}(X_t) - f_\star(X_t)\|_2^2 \lesssim \mathsf{M}_T(\mathscr{F}_\star)$

Definition (Martingale Offset Complexity [Liang et al., 2015])

$$\mathsf{M}_T(\mathscr{F}_\star) \triangleq \sup_{f \in \mathscr{F}_\star} \frac{1}{T} \sum_{t=0}^{T-1} 4\langle W_t, f(X_t)\rangle - \|f(X_t)\|_2^2$$

$\Rightarrow$ do not pay complexity for $\mathscr{F}$ but only for those hypotheses near $f_\star$

Behaves like a *local* complexity

$\Rightarrow$ $\mathsf{M}_T(\mathscr{F}_\star)$ reduces to self-normalized martingale for linear hyp.

$\Rightarrow$ $\mathsf{M}_T(\mathscr{F}_\star)$ can be bounded by chaining to give $\tilde{O}$(iid rate):

$$\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star) \lesssim \inf_{\gamma > 0} \left\{ \frac{\sigma_W^2 \log \mathcal{N}_\infty(\mathscr{F}_\star, \gamma)}{T} + \frac{\sigma_W}{\sqrt{T}} \int_0^\gamma \sqrt{\log \mathcal{N}_\infty(\mathscr{F}_\star, s)} ds + \gamma^2 \right\}.$$

$\Rightarrow$ know how to control empirical excess risk — need lower iso!

Lower Isometry: Mixing

The following Bernstein-type inequality is key

Theorem (Samson [2000, Theorem 2])

Let $g : \mathsf{X} \to \mathbb{R}$ be non-negative. Then for any $\lambda \geq 0$ we have that:

$$\mathbf{E} \exp\left(-\lambda \sum_{t=0}^{T-1} g(X_t)\right) \leq \exp\left(-\lambda \sum_{t=0}^{T-1} \mathbf{E} g(X_t) + \frac{\lambda^2 \|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2 \sum_{t=0}^{T-1} \mathbf{E} g^2(X_t)}{2}\right). \tag{9}$$

where $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}$ can be bounded as

$\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}} = O(1)$ if P_X is geo ϕ-mixing

(!) However, $\|\Gamma_{\mathsf{dep}}(\mathsf{P}_X)\|_{\mathsf{op}}^2 = o(T)$ is sufficient for us to obtain interesting results

Lower Isometry: Mixing

The following Bernstein-type inequality is key

Theorem (Samson [2000, Theorem 2])

Let $g : \mathsf{X} \to \mathbb{R}$ be non-negative. Then for any $\lambda \geq 0$ we have that:

$$\mathbf{E} \exp\left(-\lambda \sum_{t=0}^{T-1} g(X_t)\right) \leq \exp\left(-\lambda \sum_{t=0}^{T-1} \mathbf{E} g(X_t) + \frac{\lambda^2 \|\Gamma_{\mathrm{dep}}(\mathsf{P}_X)\|_{\mathrm{op}}^2 \sum_{t=0}^{T-1} \mathbf{E} g^2(X_t)}{2}\right). \tag{10}$$

where $\|\Gamma_{\mathrm{dep}}(\mathsf{P}_X)\|_{\mathrm{op}}$ is given by

Definition (Dependency matrix 1, Samson [2000, Section 2])

The *dependency matrix* of a process $X_{0:T-1}$ with distribution P_X is the (upper-triangular) matrix $\Gamma_{\mathrm{dep}}(\mathsf{P}_X) = \{\Gamma_{ij}\}_{i,j=0}^{T-1} \in \mathbb{R}^{T \times T}$ defined as follows. Let $\mathcal{X}_{0:i}$ denote the σ-field generated by $\{X_t\}_{t=0}^{i}$. For indices $i < j$, let

$$\Gamma_{ij} = \sqrt{2 \sup_{A \in \mathcal{X}_{0:i}} \|\mathsf{P}_{X_{j:T-1}}(\cdot \mid A) - \mathsf{P}_{X_{j:T-1}}\|_{\mathsf{TV}}}. \tag{11}$$

For the remaining indices $i \geq j$, let $\Gamma_{ii} = 1$ and $\Gamma_{ij} = 0$ when $i > j$ (below the diagonal).

Lower Isometry: Trajectory Hypercontractivity

Mixing does not seem to be sufficient. We also need:

Lower Isometry: Trajectory Hypercontractivity

Mixing does not seem to be sufficient. We also need:

Definition (Trajectory (C, α)-hypercontractivity)

Fix constants $C > 0$ and $\alpha \in [1, 2]$. We say that the tuple $(\mathscr{F}, \mathsf{P}_X)$ satisfies the *trajectory* (C, α)*-hypercontractivity* condition if

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^4\right] \leq C\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right]\right)^{\alpha} \text{ for all } f \in \mathscr{F}. \quad (12)$$

Here, the expectation is with respect to P_X, the joint law of $X_{0:T-1}$.

Lower Isometry: Trajectory Hypercontractivity

Mixing does not seem to be sufficient. We also need:

Definition (Trajectory (C, α)-hypercontractivity)

Fix constants $C > 0$ and $\alpha \in [1, 2]$. We say that the tuple $(\mathscr{F}, \mathsf{P}_X)$ satisfies the *trajectory* (C, α)*-hypercontractivity* condition if

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^4\right] \le C\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right]\right)^{\alpha} \text{ for all } f \in \mathscr{F}. \quad (12)$$

Here, the expectation is with respect to P_X, the joint law of $X_{0:T-1}$.

Can be thought of as a small-ball type condition (Paley-Zygmund)

Lower Isometry: Trajectory Hypercontractivity

Mixing does not seem to be sufficient. We also need:

Definition (Trajectory (C, α)-hypercontractivity)

Fix constants $C > 0$ and $\alpha \in [1, 2]$. We say that the tuple $(\mathscr{F}, \mathsf{P}_X)$ satisfies the *trajectory* (C, α)*-hypercontractivity* condition if

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^4\right] \leq C\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right]\right)^{\alpha} \text{ for all } f \in \mathscr{F}. \qquad (12)$$

Here, the expectation is with respect to P_X, the joint law of $X_{0:T-1}$.

Can be thought of as a small-ball type condition (Paley-Zygmund)

Examples satisfying traj. hyp.:

Lower Isometry: Trajectory Hypercontractivity

Mixing does not seem to be sufficient. We also need:

Definition (Trajectory (C, α)-hypercontractivity)

Fix constants $C > 0$ and $\alpha \in [1, 2]$. We say that the tuple $(\mathscr{F}, \mathsf{P}_X)$ satisfies the *trajectory* (C, α)*-hypercontractivity* condition if

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^4\right] \leq C\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right]\right)^{\alpha} \text{ for all } f \in \mathscr{F}. \quad (12)$$

Here, the expectation is with respect to P_X, the joint law of $X_{0:T-1}$.

Can be thought of as a small-ball type condition (Paley-Zygmund)

Examples satisfying traj. hyp.:

- All finite hyp-classes

Lower Isometry: Trajectory Hypercontractivity

Mixing does not seem to be sufficient. We also need:

Definition (Trajectory (C, α)-hypercontractivity)

Fix constants $C > 0$ and $\alpha \in [1, 2]$. We say that the tuple $(\mathscr{F}, \mathsf{P}_X)$ satisfies the *trajectory* (C, α)*-hypercontractivity* condition if

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^4\right] \le C\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right]\right)^{\alpha} \text{ for all } f \in \mathscr{F}. \qquad (12)$$

Here, the expectation is with respect to P_X, the joint law of $X_{0:T-1}$.

Can be thought of as a small-ball type condition (Paley-Zygmund)

Examples satisfying traj. hyp.:

- All finite hyp-classes
- LDS with log-concave noise (using Carbery and Wright [2001])

Lower Isometry: Trajectory Hypercontractivity

Mixing does not seem to be sufficient. We also need:

Definition (Trajectory (C, α)-hypercontractivity)

Fix constants $C > 0$ and $\alpha \in [1, 2]$. We say that the tuple $(\mathscr{F}, \mathsf{P}_X)$ satisfies the *trajectory* (C, α)*-hypercontractivity* condition if

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^4\right] \leq C\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right]\right)^{\alpha} \text{ for all } f \in \mathscr{F}. \tag{12}$$

Here, the expectation is with respect to P_X, the joint law of $X_{0:T-1}$.

Can be thought of as a small-ball type condition (Paley-Zygmund)

Examples satisfying traj. hyp.:

- All finite hyp-classes
- LDS with log-concave noise (using Carbery and Wright [2001])
- GLM with expansive link function

Lower Isometry: Trajectory Hypercontractivity

Mixing does not seem to be sufficient. We also need:

Definition (Trajectory (C,α)-hypercontractivity)

Fix constants $C > 0$ and $\alpha \in [1,2]$. We say that the tuple $(\mathscr{F}, \mathsf{P}_X)$ satisfies the *trajectory* (C,α)*-hypercontractivity* condition if

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^4\right] \leq C\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right]\right)^{\alpha} \text{ for all } f \in \mathscr{F}. \quad (12)$$

Here, the expectation is with respect to P_X, the joint law of $X_{0:T-1}$.

Can be thought of as a small-ball type condition (Paley-Zygmund)

Examples satisfying traj. hyp.:

- All finite hyp-classes
- LDS with log-concave noise (using Carbery and Wright [2001])
- GLM with expansive link function
- Ergodic Finite State MC (arbitrary hyp class)

Lower Isometry: Trajectory Hypercontractivity

Mixing does not seem to be sufficient. We also need:

Definition (Trajectory (C, α)-hypercontractivity)

Fix constants $C > 0$ and $\alpha \in [1, 2]$. We say that the tuple $(\mathscr{F}, \mathsf{P}_X)$ satisfies the *trajectory* (C, α)*-hypercontractivity* condition if

$$\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^4\right] \leq C\left(\mathbf{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right]\right)^\alpha \text{ for all } f \in \mathscr{F}. \qquad (12)$$

Here, the expectation is with respect to P_X, the joint law of $X_{0:T-1}$.

- Can be thought of as a small-ball type condition (Paley-Zygmund)
- Examples satisfying traj. hyp.:
  - All finite hyp-classes
  - LDS with log-concave noise (using Carbery and Wright [2001])
  - GLM with expansive link function
  - Ergodic Finite State MC (arbitrary hyp class)
  - Ellipsoids in $\ell^2(\mathbb{N})$, i.e., RKHS

Lower Isometry: Sketch

$$\mathbf{P}\left(\sum_{t=0}^{T-1} \|f(X_t)\|_2^2 \leq \frac{1}{2} \sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2\right)$$

Lower Isometry: Sketch

$$\mathbf{P}\left(\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 \le \frac{1}{2}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^2\right)$$
$$\le \inf_{\lambda\ge 0}\mathbf{E}\exp\left(\frac{\lambda}{2}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^2 - \lambda\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right) \qquad \text{(Chernoff)}$$

Lower Isometry: Sketch

$$\mathbf{P}\left(\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 \leq \frac{1}{2}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^2\right)$$

$$\leq \inf_{\lambda\geq 0}\mathbf{E}\exp\left(\frac{\lambda}{2}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^2 - \lambda\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right) \qquad \text{(Chernoff)}$$

$$\leq \inf_{\lambda\geq 0}\exp\left(-\frac{\lambda}{2}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^2 + \frac{\lambda^2\|\Gamma_{\mathrm{dep}}(\mathsf{P}_X)\|_{\mathrm{op}}^2\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^4}{2}\right) \qquad \text{(Samson's)}$$

Lower Isometry: Sketch

$$
\begin{aligned}
&\mathbf{P}\left(\sum_{t=0}^{T-1} \|f(X_t)\|_2^2 \le \frac{1}{2}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2\right) \\
&\le \inf_{\lambda\ge 0} \mathbf{E}\exp\left(\frac{\lambda}{2}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 - \lambda\sum_{t=0}^{T-1} \|f(X_t)\|_2^2\right) && \text{(Chernoff)} \\
&\le \inf_{\lambda\ge 0} \exp\left(-\frac{\lambda}{2}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 + \frac{\lambda^2\|\Gamma_{\mathrm{dep}}(\mathsf{P}_X)\|_{\mathrm{op}}^2 \sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^4}{2}\right) && \text{(Samson's)} \\
&\le \exp\left(-\frac{T}{8C\|\Gamma_{\mathrm{dep}}(\mathsf{P}_X)\|_{\mathrm{op}}^2} \times \left(\frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2\right)^{2-\alpha}\right), && \text{(hyp. con.)}
\end{aligned}
$$

Lower Isometry: Sketch

$$
\begin{aligned}
&\mathbf{P}\left(\textstyle\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 \le \frac{1}{2}\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^2\right) \\
&\le \inf_{\lambda\ge 0}\mathbf{E}\exp\left(\tfrac{\lambda}{2}\textstyle\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^2 - \lambda\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right) \qquad \text{(Chernoff)} \\
&\le \inf_{\lambda\ge 0}\exp\left(-\tfrac{\lambda}{2}\textstyle\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^2 + \frac{\lambda^2\|\Gamma_{\mathrm{dep}}(\mathsf{P}_X)\|_{\mathrm{op}}^2\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^4}{2}\right) \qquad \text{(Samson's)} \\
&\le \exp\left(-\frac{T}{8C\|\Gamma_{\mathrm{dep}}(\mathsf{P}_X)\|_{\mathrm{op}}^2}\times\left(\tfrac{1}{T}\textstyle\sum_{t=0}^{T-1}\mathbf{E}\|f(X_t)\|_2^2\right)^{2-\alpha}\right), \qquad \text{(hyp. con.)}
\end{aligned}
$$

assume star-shaped + use a union bound:

$$
\mathbf{P}\left(\sup_{f\in\mathscr{F}_\star\setminus\{\|f\|_{L^2}\le r\}}\left\{\frac{1}{T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2 - \mathbf{E}\frac{1}{8T}\sum_{t=0}^{T-1}\|f(X_t)\|_2^2\right\}\le 0\right) \le |\mathscr{F}_r|\exp\left(\frac{-Tr^{4-2\alpha}}{8C\|\Gamma_{\mathrm{dep}}(\mathsf{P}_X)\|_{\mathrm{op}}^2}\right).
$$

Main Result: Simplified

$$B(r) \triangleq \left\{ f \in \mathscr{F}_\star \,\middle|\, \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 \le r^2 \right\}, \quad \partial B(r) \triangleq \left\{ f \in \mathscr{F}_\star \,\middle|\, \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 = r^2 \right\}$$

Theorem

Fix $B > 0$, $C \in \mathbb{R}_+$, $r \in (0, B]$. Suppose:

- *that $\mathscr{F}_\star$ is star-shaped and B-bounded;*

Main Result: Simplified

$$B(r) \triangleq \left\{ f \in \mathscr{F}_\star \;\middle|\; \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 \le r^2 \right\}, \quad \partial B(r) \triangleq \left\{ f \in \mathscr{F}_\star \;\middle|\; \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 = r^2 \right\}$$

Theorem

Fix $B > 0$, $C \in \mathbb{R}_+$, $r \in (0, B]$. Suppose:

- *that $\mathscr{F}_\star$ is star-shaped and B-bounded;*
- *that $\mathscr{F}_r \subset \mathscr{F}_\star$ is an $r/\sqrt{8}$-net of $\partial B(r)$ in $\|\cdot\|_\infty$ such that*

Main Result: Simplified

$$B(r) \triangleq \left\{ f \in \mathscr{F}_\star \;\middle|\; \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 \le r^2 \right\}, \;\; \partial B(r) \triangleq \left\{ f \in \mathscr{F}_\star \;\middle|\; \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 = r^2 \right\}$$

Theorem

Fix $B > 0$, $C \in \mathbb{R}_+$, $r \in (0, B]$. Suppose:

- *that $\mathscr{F}_\star$ is star-shaped and B-bounded;*
- *that $\mathscr{F}_r \subset \mathscr{F}_\star$ is an $r/\sqrt{8}$-net of $\partial B(r)$ in $\|\cdot\|_\infty$ such that*
  - *$(\mathscr{F}_r, \mathsf{P}_X)$ is $(C, 2)$-trajectory hypercontractive*

Main Result: Simplified

$$B(r) \triangleq \left\{ f \in \mathscr{F}_\star \;\middle|\; \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 \leq r^2 \right\}, \quad \partial B(r) \triangleq \left\{ f \in \mathscr{F}_\star \;\middle|\; \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 = r^2 \right\}$$

Theorem

Fix $B > 0$, $C \in \mathbb{R}_+$, $r \in (0, B]$. Suppose:

- *that $\mathscr{F}_\star$ is star-shaped and B-bounded;*
- *that $\mathscr{F}_r \subset \mathscr{F}_\star$ is an $r/\sqrt{8}$-net of $\partial B(r)$ in $\|\cdot\|_\infty$ such that*
  - *$(\mathscr{F}_r, \mathsf{P}_X)$ is $(C, 2)$-trajectory hypercontractive*

Main Result: Simplified

$$B(r) \triangleq \left\{ f \in \mathscr{F}_\star \;\middle|\; \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 \le r^2 \right\}, \quad \partial B(r) \triangleq \left\{ f \in \mathscr{F}_\star \;\middle|\; \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 = r^2 \right\}$$

Theorem

Fix $B > 0$, $C \in \mathbb{R}_+$, $r \in (0, B]$. Suppose:

- *that $\mathscr{F}_\star$ is star-shaped and B-bounded;*
- *that $\mathscr{F}_r \subset \mathscr{F}_\star$ is an $r/\sqrt{8}$-net of $\partial B(r)$ in $\|\cdot\|_\infty$ such that*
  - *$(\mathscr{F}_r, \mathsf{P}_X)$ is $(C, 2)$-trajectory hypercontractive*

Then:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \le 8\,\underbrace{\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)}_{\text{"iid rate"}} + r^2 + B^2 \underbrace{|\mathscr{F}_r|}_{\lesssim \mathcal{N}_\infty(\mathscr{F}_\star, r)} \exp\left(\frac{-T}{8C\|\Gamma_{\mathrm{dep}}(\mathsf{P}_X)\|_{\mathrm{op}}^2} \right) \quad (13)$$

Main Result: Simplified

$$B(r) \triangleq \left\{ f \in \mathscr{F}_\star \;\middle|\; \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 \le r^2 \right\}, \quad \partial B(r) \triangleq \left\{ f \in \mathscr{F}_\star \;\middle|\; \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{E}\|f(X_t)\|_2^2 = r^2 \right\}$$

Theorem

Fix $B > 0$, $C \in \mathbb{R}_+$, $r \in (0, B]$. Suppose:

- *that $\mathscr{F}_\star$ is star-shaped and B-bounded;*
- *that $\mathscr{F}_r \subset \mathscr{F}_\star$ is an $r/\sqrt{8}$-net of $\partial B(r)$ in $\|\cdot\|_\infty$ such that*
  - $(\mathscr{F}_r, \mathsf{P}_X)$ *is $(C, 2)$-trajectory hypercontractive*

Then:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \le 8\underbrace{\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)}_{\text{"iid rate"}} + r^2 + B^2 \underbrace{|\mathscr{F}_r|}_{\lesssim \mathcal{N}_\infty(\mathscr{F}_\star, r)} \exp\left(\frac{-T}{8C\|\Gamma_{\mathrm{dep}}(\mathsf{P}_X)\|_{\mathrm{op}}^2} \right) \quad (13)$$

choose $r^2 \asymp \mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$

suppose $\|\Gamma_{\mathrm{dep}}(\mathsf{P}_X)\|_{\mathrm{op}}^2 = O(1)$

$\mathcal{N}_\infty(\mathscr{F}_\star, \mathbf{E}\mathsf{M}_T(\mathscr{F}_\star))$ grows slower than the neg. exp. term ☺

$\Rightarrow$ dominant term in (13) is $\mathbf{E}\mathsf{M}_T(\mathscr{F}_\star)$

$\Rightarrow$ iid rate after a burn-in ☺

Examples

Let's do some examples ☕

Examples

Let's do some examples ☕

Stable LDS

Examples

Let's do some examples ☕

Stable LDS

Stable and expansive GLM

Examples

Let's do some examples ☕

Stable LDS

Stable and expansive GLM

$\ell^2(\mathbb{N})$-ellipsoids ("RKHS")

Example: Stable LDS

LDS: $X_{t+1} = A_\star X_t + HV_t,\ \ X_0 = HV_0,\ \ V_t \sim N(0, I)$

[1] Technically, we verify hyp.con. and mix. for a truncated noise process and then couple

Example: Stable LDS

LDS: $X_{t+1} = A_\star X_t + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

lin hyp: $\mathscr{F} \triangleq \{f(x) = Ax \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \leq B\}$

[1] Technically, we verify hyp.con. and mix. for a truncated noise process and then couple

Example: Stable LDS

LDS: $X_{t+1} = A_\star X_t + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

lin hyp: $\mathscr{F} \triangleq \{f(x) = Ax \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \leq B\}$

$(A_\star, H)$ *k-step cont.*; rank $\left(\begin{bmatrix} H & A_\star H & A_\star^2 H & \dots & A_\star^{k-1} H \end{bmatrix}\right) = d_X$

[1] Technically, we verify hyp.con. and mix. for a truncated noise process and then couple

Example: Stable LDS

LDS: $X_{t+1} = A_\star X_t + HV_t,\ \ X_0 = HV_0,\ \ V_t \sim N(0, I)$

lin hyp: $\mathscr{F} \triangleq \{f(x) = Ax \mid A \in \mathbb{R}^{d_X \times d_X},\ \ \|A\|_F \leq B\}$

$(A_\star, H)$ *k-step cont.*; $\operatorname{rank}\left(\begin{bmatrix} H & A_\star H & A_\star^2 H & \dots & A_\star^{k-1} H \end{bmatrix}\right) = d_X$

$A_\star$ (τ, ρ)-*stable*; for all $k \in \mathbb{N}$ we have $\|A_\star^k\|_{\text{op}} \leq \tau \rho^k$ $(\rho \in (0,1))$

[1] Technically, we verify hyp.con. and mix. for a truncated noise process and then couple

Example: Stable LDS

LDS: $X_{t+1} = A_\star X_t + HV_t,\ \ X_0 = HV_0,\ \ V_t \sim N(0, I)$

lin hyp: $\mathscr{F} \triangleq \{f(x) = Ax \mid A \in \mathbb{R}^{d_X \times d_X},\ \ \|A\|_F \leq B\}$

$(A_\star, H)$ *k-step cont.*; rank $\left(\begin{bmatrix} H & A_\star H & A_\star^2 H & \dots & A_\star^{k-1} H \end{bmatrix}\right) = d_X$

$A_\star$ (τ, ρ)*-stable*; for all $k \in \mathbb{N}$ we have $\|A_\star^k\|_{\mathrm{op}} \leq \tau\rho^k$ $(\rho \in (0,1))$

$\Rightarrow$ $(C_{\mathrm{LDS}}, 2)$-traj. hyp. with $C_{\mathrm{LDS}} \lesssim \frac{\tau^4 \|H\|_{\mathrm{op}}^4}{(1-\rho)^2 \mu^2}$ where $\mu = \lambda_{\min}(\Gamma_k)$

[1] Technically, we verify hyp.con. and mix. for a truncated noise process and then couple

Example: Stable LDS

LDS: $X_{t+1} = A_\star X_t + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

lin hyp: $\mathscr{F} \triangleq \{f(x) = Ax \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \leq B\}$

$(A_\star, H)$ *k-step cont.*; $\operatorname{rank}\left(\begin{bmatrix} H & A_\star H & A_\star^2 H & \dots & A_\star^{k-1} H \end{bmatrix}\right) = d_X$

$A_\star$ (τ, ρ)*-stable*; for all $k \in \mathbb{N}$ we have $\|A_\star^k\|_{\mathrm{op}} \leq \tau \rho^k$ $(\rho \in (0,1))$

$\Rightarrow$ $(C_{\mathrm{LDS}}, 2)$-traj. hyp. with $C_{\mathrm{LDS}} \lesssim \frac{\tau^4 \|H\|_{\mathrm{op}}^4}{(1-\rho)^2 \mu^2}$ where $\mu = \lambda_{\min}(\Gamma_k)$

$\Rightarrow$ can also control dependency matrix by stability

[1] Technically, we verify hyp.con. and mix. for a truncated noise process and then couple

Example: Stable LDS

LDS: $X_{t+1} = A_\star X_t + HV_t,\;\; X_0 = HV_0,\;\; V_t \sim N(0, I)$

lin hyp: $\mathscr{F} \triangleq \{f(x) = Ax \mid A \in \mathbb{R}^{d_X \times d_X},\;\; \|A\|_F \le B\}$

$(A_\star, H)$ *k-step cont.*; $\operatorname{rank}\left(\begin{bmatrix} H & A_\star H & A_\star^2 H & \ldots & A_\star^{k-1} H \end{bmatrix}\right) = d_X$

$A_\star$ (τ, ρ)*-stable*; for all $k \in \mathbb{N}$ we have $\|A_\star^k\|_{\mathrm{op}} \le \tau \rho^k$ $(\rho \in (0,1))$

$\Rightarrow$ $(C_{\mathrm{LDS}}, 2)$-traj. hyp. with $C_{\mathrm{LDS}} \lesssim \frac{\tau^4 \|H\|_{\mathrm{op}}^4}{(1-\rho)^2 \mu^2}$ where $\mu = \lambda_{\min}(\Gamma_k)$

$\Rightarrow$ can also control dependency matrix by stability

Use our main theorem + truncation[1]:

$$\mathbf{E}\|(\widehat{A} - A_\star)\sqrt{\Sigma_X}\|_F^2 \lesssim \frac{\|H\|_{\mathrm{op}}^2 d_X^2}{T} \qquad (T \ge \mathrm{poly}(\mathtt{params}))$$

[1] Technically, we verify hyp.con. and mix. for a truncated noise process and then couple

Example: Stable LDS

LDS: $X_{t+1} = A_\star X_t + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

lin hyp: $\mathscr{F} \triangleq \{f(x) = Ax \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \leq B\}$

$(A_\star, H)$ *k-step cont.*; $\operatorname{rank}\left(\begin{bmatrix} H & A_\star H & A_\star^2 H & \dots & A_\star^{k-1} H \end{bmatrix}\right) = d_X$

$A_\star$ (τ, ρ)*-stable*; for all $k \in \mathbb{N}$ we have $\|A_\star^k\|_{\mathrm{op}} \leq \tau \rho^k$ $(\rho \in (0,1))$

$\Rightarrow$ $(C_{\mathrm{LDS}}, 2)$-traj. hyp. with $C_{\mathrm{LDS}} \lesssim \frac{\tau^4 \|H\|_{\mathrm{op}}^4}{(1-\rho)^2 \mu^2}$ where $\mu = \lambda_{\min}(\Gamma_k)$

$\Rightarrow$ can also control dependency matrix by stability

Use our main theorem + truncation[1]:

$$\mathbf{E}\|(\widehat{A} - A_\star)\sqrt{\Sigma_X}\|_F^2 \lesssim \frac{\|H\|_{\mathrm{op}}^2 d_X^2}{T} \qquad (T \geq \mathrm{poly}(\mathtt{params}))$$

matches the iid minimax rate after a burn-in

[1] Technically, we verify hyp.con. and mix. for a truncated noise process and then couple

Example: Stable LDS

LDS: $X_{t+1} = A_\star X_t + HV_t,\ \ X_0 = HV_0,\ \ V_t \sim N(0, I)$

lin hyp: $\mathscr{F} \triangleq \{f(x) = Ax \mid A \in \mathbb{R}^{d_X \times d_X},\ \ \|A\|_F \le B\}$

$(A_\star, H)$ *k-step cont.*; $\operatorname{rank}\left(\begin{bmatrix} H & A_\star H & A_\star^2 H & \dots & A_\star^{k-1} H \end{bmatrix}\right) = d_X$

$A_\star$ (τ, ρ)*-stable*; for all $k \in \mathbb{N}$ we have $\|A_\star^k\|_{\mathrm{op}} \le \tau\rho^k$ $(\rho \in (0,1))$

$\Rightarrow$ $(C_{\mathrm{LDS}}, 2)$-traj. hyp. with $C_{\mathrm{LDS}} \lesssim \frac{\tau^4 \|H\|_{\mathrm{op}}^4}{(1-\rho)^2\mu^2}$ where $\mu = \lambda_{\min}(\Gamma_k)$

$\Rightarrow$ can also control dependency matrix by stability

Use our main theorem + truncation[1]:

$$\mathbf{E}\|(\widehat{A} - A_\star)\sqrt{\Sigma_X}\|_F^2 \lesssim \frac{\|H\|_{\mathrm{op}}^2 d_X^2}{T} \qquad (T \ge \mathrm{poly}(\texttt{params}))$$

matches the iid minimax rate after a burn-in

relies on a bound from Tu et al. [2022] on the RHS of

$$\mathbf{E}\mathrm{M}_T(\mathscr{F}_\star) \le \frac{4}{T}\mathbf{E}\left\|\left(\sum_{t=0}^{T-1} X_t X_t^\top\right)^{-1/2} \sum_{t=0}^{T-1} X_t V_t^\top H^\top\right\|_F^2$$

[1] Technically, we verify hyp.con. and mix. for a truncated noise process and then couple

Example: Stable GLM

GLM: $X_{t+1} = \sigma(A_\star X_t) + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

Example: Stable GLM

GLM: $X_{t+1} = \sigma(A_\star X_t) + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

$$\mathscr{F} \triangleq \{f(x) = \sigma(Ax) \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \leq B\}$$

Example: Stable GLM

GLM: $X_{t+1} = \sigma(A_\star X_t) + HV_t,\ \ X_0 = HV_0,\ \ V_t \sim N(0, I)$

$\mathscr{F} \triangleq \{f(x) = \sigma(Ax) \mid A \in \mathbb{R}^{d_X \times d_X},\ \ \|A\|_F \leq B\}$

1-step-cont.;$H \in \mathbb{R}^{d_X \times d_X}$ is full rank

Example: Stable GLM

GLM: $X_{t+1} = \sigma(A_\star X_t) + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

$\mathscr{F} \triangleq \{f(x) = \sigma(Ax) \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \leq B\}$

1-step-cont.;$H \in \mathbb{R}^{d_X \times d_X}$ is full rank

σ is 1-lip

Example: Stable GLM

GLM: $X_{t+1} = \sigma(A_\star X_t) + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

$\mathscr{F} \triangleq \{f(x) = \sigma(Ax) \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \leq B\}$

1-step-cont.;$H \in \mathbb{R}^{d_X \times d_X}$ is full rank

σ is 1-lip

σ is expansive; $\exists \zeta \in (0, 1] : |\sigma(x) - \sigma(y)| \geq \zeta |x - y|$ for all $x, y \in \mathbb{R}$

Example: Stable GLM

GLM: $X_{t+1} = \sigma(A_\star X_t) + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

$\mathscr{F} \triangleq \{f(x) = \sigma(Ax) \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \leq B\}$

1-step-cont.;$H \in \mathbb{R}^{d_X \times d_X}$ is full rank

σ is 1-lip

σ is expansive; $\exists \zeta \in (0, 1] : \; |\sigma(x) - \sigma(y)| \geq \zeta |x - y|$ for all $x, y \in \mathbb{R}$

$\exists$ diagonal $P_\star \in \mathbb{R}^{d_X \times d_X}$ w/ $P_\star \succcurlyeq I$, $\rho \in (0, 1)$ with $A_\star^\top P_\star A_\star \preccurlyeq \rho P_\star$

Example: Stable GLM

GLM: $X_{t+1} = \sigma(A_\star X_t) + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

$\mathscr{F} \triangleq \{f(x) = \sigma(Ax) \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \leq B\}$

1-step-cont.;$H \in \mathbb{R}^{d_X \times d_X}$ is full rank

σ is 1-lip

σ is expansive; $\exists \zeta \in (0, 1] : |\sigma(x) - \sigma(y)| \geq \zeta |x - y|$ for all $x, y \in \mathbb{R}$

$\exists$ diagonal $P_\star \in \mathbb{R}^{d_X \times d_X}$ w/ $P_\star \succcurlyeq I$, $\rho \in (0, 1)$ with $A_\star^\top P_\star A_\star \preccurlyeq \rho P_\star$

$\Rightarrow$ $(C_{\mathrm{GLM}}, 2)$-traj. hyp. with $C_{\mathrm{GLM}} \lesssim \frac{B_X^4}{\sigma_{\min}(H)^4 \zeta^4}$ with

$B_X = \frac{\|H\|_{\mathrm{op}} \|P_\star\|_{\mathrm{op}}^{1/2} \sqrt{d_X}}{1-\rho}$

Example: Stable GLM

GLM: $X_{t+1} = \sigma(A_\star X_t) + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

$\mathscr{F} \triangleq \{f(x) = \sigma(Ax) \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \leq B\}$

1-step-cont.;$H \in \mathbb{R}^{d_X \times d_X}$ is full rank

σ is 1-lip

σ is expansive; $\exists \zeta \in (0, 1] : \; |\sigma(x) - \sigma(y)| \geq \zeta |x - y|$ for all $x, y \in \mathbb{R}$

$\exists$ diagonal $P_\star \in \mathbb{R}^{d_X \times d_X}$ w/ $P_\star \succcurlyeq I$, $\rho \in (0, 1)$ with $A_\star^\top P_\star A_\star \preccurlyeq \rho P_\star$

$\Rightarrow$ $(C_{\mathrm{GLM}}, 2)$-traj. hyp. with $C_{\mathrm{GLM}} \lesssim \frac{B_X^4}{\sigma_{\min}(H)^4 \zeta^4}$ with

$$B_X = \frac{\|H\|_{\mathrm{op}} \|P_\star\|_{\mathrm{op}}^{1/2} \sqrt{d_X}}{1 - \rho}$$

$\Rightarrow$ can also control dependency matrix by stability

Example: Stable GLM

GLM: $X_{t+1} = \sigma(A_\star X_t) + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

$\mathscr{F} \triangleq \{f(x) = \sigma(Ax) \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \le B\}$

1-step-cont.;$H \in \mathbb{R}^{d_X \times d_X}$ is full rank

σ is 1-lip

σ is expansive; $\exists \zeta \in (0,1] : \; |\sigma(x) - \sigma(y)| \ge \zeta |x - y|$ for all $x, y \in \mathbb{R}$

$\exists$ diagonal $P_\star \in \mathbb{R}^{d_X \times d_X}$ w/ $P_\star \succcurlyeq I$, $\rho \in (0,1)$ with $A_\star^\top P_\star A_\star \preccurlyeq \rho P_\star$

$\Rightarrow$ $(C_{\mathrm{GLM}}, 2)$-traj. hyp. with $C_{\mathrm{GLM}} \lesssim \frac{B_X^4}{\sigma_{\min}(H)^4 \zeta^4}$ with

$$B_X = \frac{\|H\|_{\mathrm{op}} \|P_\star\|_{\mathrm{op}}^{1/2} \sqrt{d_X}}{1-\rho}$$

$\Rightarrow$ can also control dependency matrix by stability

Use our main result + truncation:

$$\mathbf{E}\|\sigma(\widehat{A}\cdot) - \sigma(A_\star \cdot)\|_{L^2}^2 \lesssim \frac{\|H\|_{\mathrm{op}}^2 d_X^2}{T} \log\left(\max\left\{T, B, d_X, \|P_\star\|_{\mathrm{op}}, \|H\|_{\mathrm{op}}, \frac{1}{1-\rho}\right\}\right)$$

Example: Stable GLM

GLM: $X_{t+1} = \sigma(A_\star X_t) + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

$\mathscr{F} \triangleq \{f(x) = \sigma(Ax) \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \leq B\}$

1-step-cont.;$H \in \mathbb{R}^{d_X \times d_X}$ is full rank

σ is 1-lip

σ is expansive; $\exists \zeta \in (0,1] : |\sigma(x) - \sigma(y)| \geq \zeta |x - y|$ for all $x, y \in \mathbb{R}$

$\exists$ diagonal $P_\star \in \mathbb{R}^{d_X \times d_X}$ w/ $P_\star \succcurlyeq I$, $\rho \in (0,1)$ with $A_\star^\top P_\star A_\star \preccurlyeq \rho P_\star$

$\Rightarrow$ $(C_{\mathrm{GLM}}, 2)$-traj. hyp. with $C_{\mathrm{GLM}} \lesssim \frac{B_X^4}{\sigma_{\min}(H)^4 \zeta^4}$ with

$B_X = \frac{\|H\|_{\mathrm{op}} \|P_\star\|_{\mathrm{op}}^{1/2} \sqrt{d_X}}{1-\rho}$

$\Rightarrow$ can also control dependency matrix by stability

Use our main result + truncation:

$$\mathbf{E}\|\sigma(\widehat{A}\cdot) - \sigma(A_\star \cdot)\|_{L^2}^2 \lesssim \frac{\|H\|_{\mathrm{op}}^2 d_X^2}{T} \log\left(\max\left\{T, B, d_X, \|P_\star\|_{\mathrm{op}}, \|H\|_{\mathrm{op}}, \frac{1}{1-\rho}\right\}\right)$$

Compare Kowshik et al. [2021]: $\|\widehat{A} - A_\star\|_F^2 = \tilde{O}(\|H\|_{\mathrm{op}}^2 d_X^2 / (T \lambda_{\min}(\Sigma_X)))$

Example: Stable GLM

GLM: $X_{t+1} = \sigma(A_\star X_t) + HV_t, \;\; X_0 = HV_0, \;\; V_t \sim N(0, I)$

$\mathscr{F} \triangleq \{f(x) = \sigma(Ax) \mid A \in \mathbb{R}^{d_X \times d_X}, \;\; \|A\|_F \leq B\}$

1-step-cont.;$H \in \mathbb{R}^{d_X \times d_X}$ is full rank

σ is 1-lip

σ is expansive; $\exists \zeta \in (0,1] : \; |\sigma(x) - \sigma(y)| \geq \zeta |x - y|$ for all $x, y \in \mathbb{R}$

$\exists$ diagonal $P_\star \in \mathbb{R}^{d_X \times d_X}$ w/ $P_\star \succcurlyeq I$, $\rho \in (0,1)$ with $A_\star^\top P_\star A_\star \preccurlyeq \rho P_\star$

$\Rightarrow$ $(C_{\mathrm{GLM}}, 2)$-traj. hyp. with $C_{\mathrm{GLM}} \lesssim \frac{B_X^4}{\sigma_{\min}(H)^4 \zeta^4}$ with

$B_X = \frac{\|H\|_{\mathrm{op}} \|P_\star\|_{\mathrm{op}}^{1/2} \sqrt{d_X}}{1-\rho}$

$\Rightarrow$ can also control dependency matrix by stability

Use our main result + truncation:

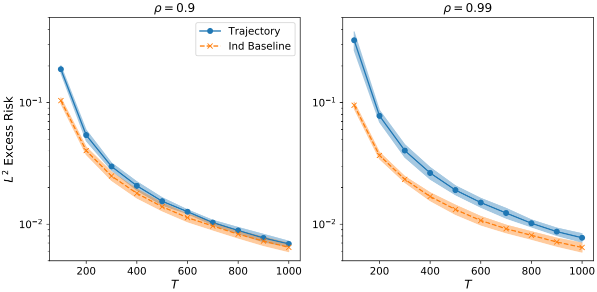
$$\mathbf{E}\|\sigma(\widehat{A}\cdot) - \sigma(A_\star \cdot)\|_{L^2}^2 \lesssim \frac{\|H\|_{\mathrm{op}}^2 d_X^2}{T} \log\left(\max\left\{T, B, d_X, \|P_\star\|_{\mathrm{op}}, \|H\|_{\mathrm{op}}, \frac{1}{1-\rho}\right\}\right)$$

Compare Kowshik et al. [2021]: $\|\widehat{A} - A_\star\|_F^2 = \tilde{O}(\|H\|_{\mathrm{op}}^2 d_X^2 / (T \lambda_{\min}(\Sigma_X)))$

First up-to-logarithms rate-optimal *excess risk bound* 😎

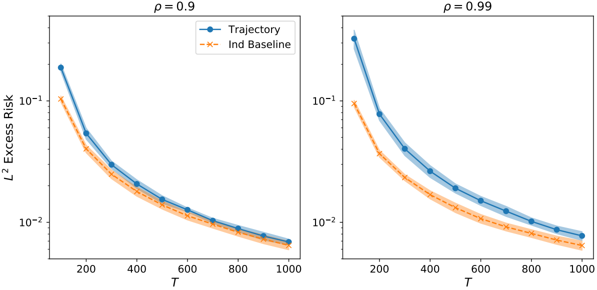
Example: Stable GLM, numerical experiment

LeakyReLU with slope 0.5, i.e., $\sigma(x) = 0.5x\mathbf{1}\{x < 0\} + x\mathbf{1}\{x \geq 0\}$



Example: Stable GLM, numerical experiment

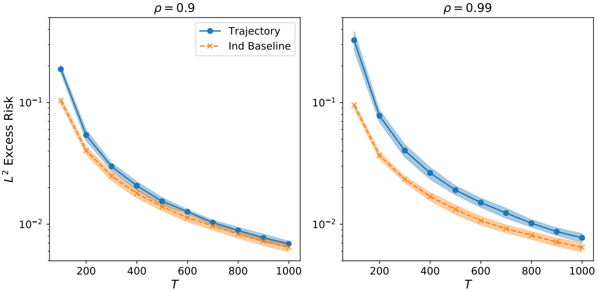
LeakyReLU with slope 0.5, i.e., $\sigma(x) = 0.5x\mathbf{1}\{x < 0\} + x\mathbf{1}\{x \geq 0\}$



L^2 excess risk as a function of dataset length T of ERM

Example: Stable GLM, numerical experiment

LeakyReLU with slope 0.5, i.e., $\sigma(x) = 0.5x\mathbf{1}\{x < 0\} + x\mathbf{1}\{x \geq 0\}$

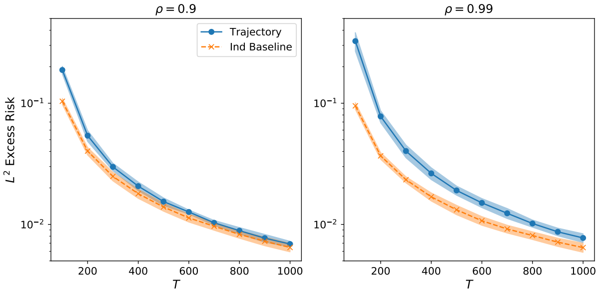


L^2 excess risk as a function of dataset length T of ERM

single trajectory (Trajectory) dataset versus independent baseline (Ind Baseline) dataset

Example: Stable GLM, numerical experiment

LeakyReLU with slope 0.5, i.e., $\sigma(x) = 0.5x\mathbf{1}\{x < 0\} + x\mathbf{1}\{x \geq 0\}$



L^2 excess risk as a function of dataset length T of ERM

single trajectory (Trajectory) dataset versus independent baseline (Ind Baseline) dataset

independent baseline: same marginals but iid

Example: $\ell^2(\mathbb{N})$-ellipsoids and hypercontractivity

Proposition

Fix $\beta, B, K, q, \varepsilon > 0$ *and a base measure* λ *on* X

Example: $\ell^2(\mathbb{N})$-ellipsoids and hypercontractivity

Proposition

Fix $\beta, B, K, q, \varepsilon > 0$ *and a base measure* λ *on* X
$\{\phi_n\}_{n\in\mathbb{N}_+}$*: ONS in* $L^2(\lambda)$ *satisfying* $\|\phi_n\|_\infty \leq Bn^q$, $\forall n \in \mathbb{N}$

Example: $\ell^2(\mathbb{N})$-ellipsoids and hypercontractivity

Proposition

Fix $\beta, B, K, q, \varepsilon > 0$ *and a base measure* λ *on* X
$\{\phi_n\}_{n \in \mathbb{N}_+}$*: ONS in* $L^2(\lambda)$ *satisfying* $\|\phi_n\|_\infty \leq Bn^q, \forall n \in \mathbb{N}$
$\mu_j \leq e^{-2\beta j}$ *and define the ellipsoid:*

$$\mathscr{P} \triangleq \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \middle| \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1 \right\}$$

Example: $\ell^2(\mathbb{N})$-ellipsoids and hypercontractivity

Proposition

Fix $\beta, B, K, q, \varepsilon > 0$ *and a base measure* λ *on* X
$\{\phi_n\}_{n\in\mathbb{N}_+}$*: ONS in* $L^2(\lambda)$ *satisfying* $\|\phi_n\|_\infty \leq Bn^q$, $\forall n \in \mathbb{N}$
$\mu_j \leq e^{-2\beta j}$ *and define the ellipsoid:*

$$\mathscr{P} \triangleq \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \middle| \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1 \right\}$$

Let $P \subset \mathscr{P}$ *be an arbitrary subset*

Example: $\ell^2(\mathbb{N})$-ellipsoids and hypercontractivity

Proposition

Fix $\beta, B, K, q, \varepsilon > 0$ *and a base measure* λ *on* X
$\{\phi_n\}_{n\in\mathbb{N}_+}$*: ONS in* $L^2(\lambda)$ *satisfying* $\|\phi_n\|_\infty \leq Bn^q$, $\forall n \in \mathbb{N}$
$\mu_j \leq e^{-2\beta j}$ *and define the ellipsoid:*

$$\mathscr{P} \triangleq \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \middle| \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1 \right\}$$

Let $P \subset \mathscr{P}$ *be an arbitrary subset*
m_ε *int. solution to* $m \geq \frac{2}{\beta} \left| \log\left(\frac{8B}{\beta\varepsilon}\right) \right|$ *subject to* $\frac{m}{\log m} \geq \frac{q}{\beta}$

Example: $\ell^2(\mathbb{N})$-ellipsoids and hypercontractivity

Proposition

Fix $\beta, B, K, q, \varepsilon > 0$ and a base measure λ on X
$\{\phi_n\}_{n\in\mathbb{N}_+}$: ONS in $L^2(\lambda)$ satisfying $\|\phi_n\|_\infty \le Bn^q$, $\forall n \in \mathbb{N}$
$\mu_j \le e^{-2\beta j}$ and define the ellipsoid:

$$\mathscr{P} \triangleq \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \,\middle|\, \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \le 1 \right\}$$

Let $P \subset \mathscr{P}$ be an arbitrary subset
m_ε int. solution to $m \ge \frac{2}{\beta}\left|\log\left(\frac{8B}{\beta\varepsilon}\right)\right|$ subject to $\frac{m}{\log m} \ge \frac{q}{\beta}$
1. There exists an ε-cover P_ε of P in the $\|\cdot\|_\infty$-norm satisfying:

$$\log|P_\varepsilon| \le m_\varepsilon \log\left(1 + \frac{8Bm_\varepsilon^q}{\varepsilon}\right)$$

Example: $\ell^2(\mathbb{N})$-ellipsoids and hypercontractivity

Proposition

Fix $\beta, B, K, q, \varepsilon > 0$ *and a base measure* λ *on* X
$\{\phi_n\}_{n\in\mathbb{N}_+}$*: ONS in* $L^2(\lambda)$ *satisfying* $\|\phi_n\|_\infty \leq Bn^q$, $\forall n \in \mathbb{N}$
$\mu_j \leq e^{-2\beta j}$ *and define the ellipsoid:*

$$\mathscr{P} \triangleq \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \,\middle|\, \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1 \right\}$$

Let $P \subset \mathscr{P}$ *be an arbitrary subset*
m_ε *int. solution to* $m \geq \frac{2}{\beta} \left| \log \left(\frac{8B}{\beta \varepsilon} \right) \right|$ *subject to* $\frac{m}{\log m} \geq \frac{q}{\beta}$
1. There exists an ε*-cover* P_ε *of* P *in the* $\|\cdot\|_\infty$*-norm satisfying:*

$$\log |P_\varepsilon| \leq m_\varepsilon \log \left(1 + \frac{8Bm_\varepsilon^q}{\varepsilon} \right)$$

$\{\mu_t\}_{t=0}^{T-1}$ *marginals of* P_X*: suppose that* $\max_{0\leq t\leq T-1} \max \left\{ \frac{d\mu_t}{d\lambda}, \frac{d\lambda}{d\mu_t} \right\} \leq K$

Example: $\ell^2(\mathbb{N})$-ellipsoids and hypercontractivity

Proposition

Fix $\beta, B, K, q, \varepsilon > 0$ *and a base measure* λ *on* X
$\{\phi_n\}_{n\in\mathbb{N}_+}$*: ONS in* $L^2(\lambda)$ *satisfying* $\|\phi_n\|_\infty \leq Bn^q$, $\forall n \in \mathbb{N}$
$\mu_j \leq e^{-2\beta j}$ *and define the ellipsoid:*

$$\mathscr{P} \triangleq \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \,\middle|\, \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1 \right\}$$

Let $P \subset \mathscr{P}$ *be an arbitrary subset*
m_ε *int. solution to* $m \geq \frac{2}{\beta}\left|\log\left(\frac{8B}{\beta\varepsilon}\right)\right|$ *subject to* $\frac{m}{\log m} \geq \frac{q}{\beta}$
1. There exists an ε*-cover* P_ε *of* P *in the* $\|\cdot\|_\infty$*-norm satisfying:*

$$\log|P_\varepsilon| \leq m_\varepsilon \log\left(1 + \frac{8Bm_\varepsilon^q}{\varepsilon}\right)$$

$\{\mu_t\}_{t=0}^{T-1}$ *marginals of* P_X*: suppose that* $\max_{0\leq t\leq T-1} \max\left\{\frac{d\mu_t}{d\lambda}, \frac{d\lambda}{d\mu_t}\right\} \leq K$
2. as long as $\varepsilon \leq \inf_{f\in P}\|f\|_{L^2(\mathsf{P}_X)}$*:*

Example: $\ell^2(\mathbb{N})$-ellipsoids and hypercontractivity

Proposition

Fix $\beta, B, K, q, \varepsilon > 0$ and a base measure λ on X
$\{\phi_n\}_{n\in\mathbb{N}_+}$*: ONS in $L^2(\lambda)$ satisfying $\|\phi_n\|_\infty \leq Bn^q$, $\forall n \in \mathbb{N}$*
$\mu_j \leq e^{-2\beta j}$ and define the ellipsoid:

$$\mathscr{P} \triangleq \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \,\middle|\, \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \leq 1 \right\}$$

Let $P \subset \mathscr{P}$ be an arbitrary subset
m_ε int. solution to $m \geq \frac{2}{\beta} \left| \log\left(\frac{8B}{\beta\varepsilon}\right) \right|$ subject to $\frac{m}{\log m} \geq \frac{q}{\beta}$
1. There exists an ε-cover P_ε of P in the $\|\cdot\|_\infty$-norm satisfying:

$$\log |P_\varepsilon| \leq m_\varepsilon \log\left(1 + \frac{8Bm_\varepsilon^q}{\varepsilon}\right)$$

$\{\mu_t\}_{t=0}^{T-1}$ marginals of P_X: suppose that $\max_{0\leq t\leq T-1} \max\left\{\frac{d\mu_t}{d\lambda}, \frac{d\lambda}{d\mu_t}\right\} \leq K$
2. as long as $\varepsilon \leq \inf_{f\in P} \|f\|_{L^2(\mathrm{P}_X)}$:

$$(P_\varepsilon, \mathrm{P}_X) \text{ is } (C_\varepsilon, 2)\text{-traj. hyp. with } C_\varepsilon = (1 + 7K^3B^4m_\varepsilon^{4q+2})$$

$\ell^2(\mathbb{N})$-ellipsoids

$$\mathscr{P}_\star \triangleq \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \,\middle|\, \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \le 1 \right\} - \{f_\star\}$$

Under the hypotheses of the previous slide:

exponential eigenvalue decay

bounded ONS growth in $\|\cdot\|_\infty$

M.A.C. marginals

we get for $T \geq \mathrm{poly}(\mathrm{params})$:

$$\mathbf{E}\|\widehat{f} - f_\star\|_{L^2}^2 \lesssim \mathbf{E}\mathsf{M}_T(\mathscr{P}_\star)$$

can bound $\mathbf{E}\mathsf{M}_T(\mathscr{P}_\star) = \tilde{O}(1/T)$ by chaining [Ziemann et al., 2022]

Summarizing

Paper: https://arxiv.org/abs/2206.08269

Summarizing

Paper: https://arxiv.org/abs/2206.08269

Provide a unified approach to learning in nonlinear time-series

Summarizing

Paper: https://arxiv.org/abs/2206.08269

Provide a unified approach to learning in nonlinear time-series

After a burn-in, obtain iid-like excess risk bounds for:

Summarizing

Paper: https://arxiv.org/abs/2206.08269

Provide a unified approach to learning in nonlinear time-series

After a burn-in, obtain iid-like excess risk bounds for:

- LDS, GLM, RKHS, finite hyp class, ergodic finite state MC

Summarizing

Paper: https://arxiv.org/abs/2206.08269

Provide a unified approach to learning in nonlinear time-series

After a burn-in, obtain iid-like excess risk bounds for:

LDS, GLM, RKHS, finite hyp class, ergodic finite state MC

Future directions

Summarizing

Paper: https://arxiv.org/abs/2206.08269

Provide a unified approach to learning in nonlinear time-series

After a burn-in, obtain iid-like excess risk bounds for:

- LDS, GLM, RKHS, finite hyp class, ergodic finite state MC

Future directions

- ☕ find conditions to do this without mixing entirely

Summarizing

Paper: https://arxiv.org/abs/2206.08269

Provide a unified approach to learning in nonlinear time-series

After a burn-in, obtain iid-like excess risk bounds for:

LDS, GLM, RKHS, finite hyp class, ergodic finite state MC

Future directions

- ☕ find conditions to do this without mixing entirely
  - Nagaraj et al. [2020]: burn-in unavoidable *only in the worst case*

Summarizing

Paper: https://arxiv.org/abs/2206.08269

Provide a unified approach to learning in nonlinear time-series

After a burn-in, obtain iid-like excess risk bounds for:

LDS, GLM, RKHS, finite hyp class, ergodic finite state MC

Future directions

- ☕ find conditions to do this without mixing entirely
  - Nagaraj et al. [2020]: burn-in unavoidable *only in the worst case*
- ☕ interplay of mixing (lack thereof) and non-realizability

Summarizing

Paper: https://arxiv.org/abs/2206.08269

Provide a unified approach to learning in nonlinear time-series

After a burn-in, obtain iid-like excess risk bounds for:

LDS, GLM, RKHS, finite hyp class, ergodic finite state MC

Future directions

- ☕ find conditions to do this without mixing entirely
  - Nagaraj et al. [2020]: burn-in unavoidable *only in the worst case*
- ☕ interplay of mixing (lack thereof) and non-realizability
  - Nagaraj et al. [2020]: deflation unavoidable in the *worst case*

Summarizing

Paper: https://arxiv.org/abs/2206.08269

Provide a unified approach to learning in nonlinear time-series

After a burn-in, obtain iid-like excess risk bounds for:

LDS, GLM, RKHS, finite hyp class, ergodic finite state MC

Future directions

- ☕ find conditions to do this without mixing entirely
  - Nagaraj et al. [2020]: burn-in unavoidable *only in the worst case*
- ☕ interplay of mixing (lack thereof) and non-realizability
  - Nagaraj et al. [2020]: deflation unavoidable in the *worst case*
- ☕ Can do this with "classical" regularization but not with "modern"

Thanks for Listening

ziemann@kth.com

Bonus: an open problem

Consider LDS: $X_{t+1} = A_\star X_t + W_t$

but assume $A_\star$ is known to be s-sparse

can invoke our main thm to obtain

$$\mathbf{E}\|(\widehat{A} - A_\star)\sqrt{\Sigma_X}\|_F^2 = \tilde{O}\left(\frac{\sigma_W^2 s \log d}{T}\right)$$

not tractable (search over $\exp(s)$ ERMs) ☹

Known results for LASSO on LDS are linear in the mixing time[2]

$$\|(\widehat{A} - A_\star)\sqrt{\Sigma_X}\|_F^2 \lesssim \frac{t_{\mathrm{mix}}\sigma_W^2 s \log d}{T}$$

tractable ☺

not minimax optimal ☹

Question: What is going on? Is there a trade-off between computation and statistical efficiency, or are existing analyses simply sub-optimal?

More open problems in our survey: Tsiamis et al. [2022b]

[2] Fattahi et al. [2019], Wainwright [2019], Lecué and Mendelson [2018]

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I. Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473. PMLR, 2018.

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

Anastasios Tsiamis, Ingvar M Ziemann, Manfred Morari, Nikolai Matni, and George J Pappas. Learning to control linear systems can be hard. In *Conference on Learning Theory*, pages 3820–3857. PMLR, 2022a.

Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George Pappas. Statistical learning theory for control: A finite sample perspective. *Under review for IEEE Control Systems Magazine*, 2022b.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in Neural Information Processing Systems*, 33, 2020.

Yahya Sattar and Samet Oymak. Non-asymptotic and accurate learning of nonlinear dynamical systems. *Journal of Machine Learning Research*, 23 (140):1–49, 2022.

Dylan Foster, Tuhin Sarkar, and Alexander Rakhlin. Learning nonlinear dynamical systems from a single trajectory. In *Learning for Dynamics and Control*, pages 851–861. PMLR, 2020.

Suhas Kowshik, Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Near-optimal offline and streaming algorithms for learning non-linear dynamical systems. *Advances in Neural Information Processing Systems*, 34, 2021.

Abhishek Roy, Krishnakumar Balasubramanian, and Murat A Erdogdu. On empirical risk minimization with dependent and heavy-tailed data. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Ingvar M Ziemann, Henrik Sandberg, and Nikolai Matni. Single trajectory nonparametric learning of nonlinear dynamics. In *Conference on Learning Theory*, pages 3333–3364. PMLR, 2022.

Yahya Sattar, Samet Oymak, and Necmiye Ozay. Finite sample identification of bilinear dynamical systems. *arXiv preprint arXiv:2208.13915*, 2022.

Samet Oymak and Necmiye Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American control conference (ACC)*, pages 5655–5661. IEEE, 2019.

Anastasios Tsiamis and George J. Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654. IEEE, 2019.

Tuhin Sarkar and Alexander Rakhlin. Near Optimal Finite Time Identification of Arbitrary Linear Dynamical Systems. In *International Conference on Machine Learning*, pages 5610–5618, 2019.

Holden Lee. Improved rates for prediction and identification of partially observed linear dynamical systems. In *International Conference on Algorithmic Learning Theory*, pages 668–698. PMLR, 2022.

Boualem Djehiche and Othmane Mazhar. Efficient learning of hidden state lti state space models of unknown order. *arXiv preprint arXiv:2202.01625*, 2022.

Yue Sun, Samet Oymak, and Maryam Fazel. Finite sample identification of low-order lti systems via nuclear norm regularization. *IEEE Open Journal of Control Systems*, 2022.

Yahya Sattar, Zhe Du, Davoud Ataee Tarzanagh, Laura Balzano, Necmiye Ozay, and Samet Oymak. Identification and adaptive control of markov jump systems: Sample complexity and regret bounds. *arXiv preprint arXiv:2111.07018*, 2021.

Yasin Abbasi-Yadkori and Csaba Szepesvári. Regret Bounds for the Adaptive Control of Linear Quadratic Systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 1–26, 2011.

Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39. PMLR, 2014.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285. PMLR, 2015.

Paul-Marie Samson. Concentration of measure inequalities for markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

Anthony Carbery and James Wright. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8: 233–248, 2001.

Stephen Tu, Roy Frostig, and Mahdi Soltanolkotabi. Learning from many trajectories. *arXiv preprint arXiv:2203.17193*, 2022.

Salar Fattahi, Nikolai Matni, and Somayeh Sojoudi. Learning sparse dynamical systems from a single sample trajectory. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2682–2689. IEEE, 2019.

Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method i: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.