An Alternative Perspective

The Offset Basic Inequality and Learning Nonlinear Dynamics

Ingvar Ziemann (Penn)

Introduction

We are interested in learning from 'mixing' (~ stable) time-series data

Focus on the square loss function

Classical results in this area rely on blocking [Yu 1994]

Transforms T dependent data points into $n = T/k$ independent data points

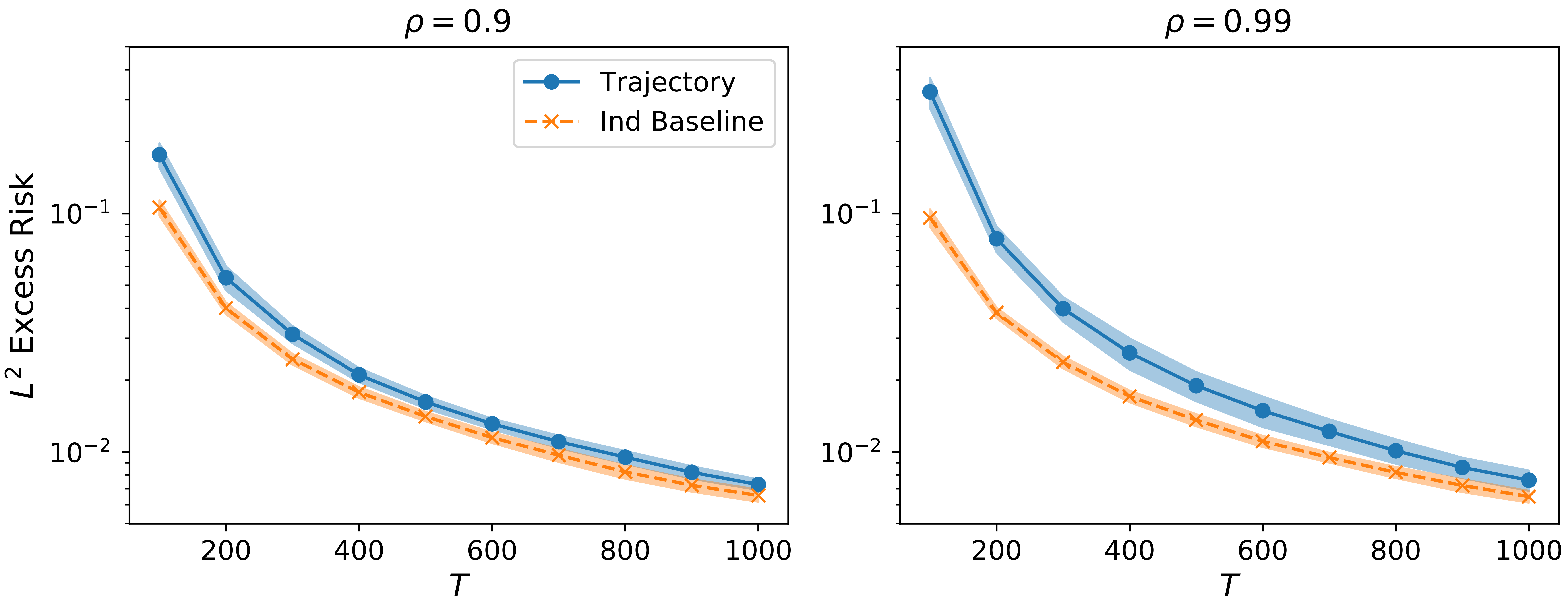
$Z_{1:T} \Rightarrow \tilde{Z}_{1:k}, \tilde{Z}_{k+1:2k}, \tilde{Z}_{2k+1:3k}, \dots$ n independent 'blocks'

Allows us to port iid machine learning results to the dependent setting

Generically employed, deflates rate of convergence by a factor of the mixing time

Dependency Deflation?

Consider an autoregressive GLM: $X_{t+1} = \sigma(\theta^{\star} X_t) + V_t$



Independent base: same marginals as trajectory but decoupled

$\rho \in (0,1)$: degree of dependence ($\rho = 1$ does not mix!)

The Statistical Model

Setup

MDS + subG

$$Y_t = f^\star(X_t) + V_t, \quad t = 1, \ldots, T$$

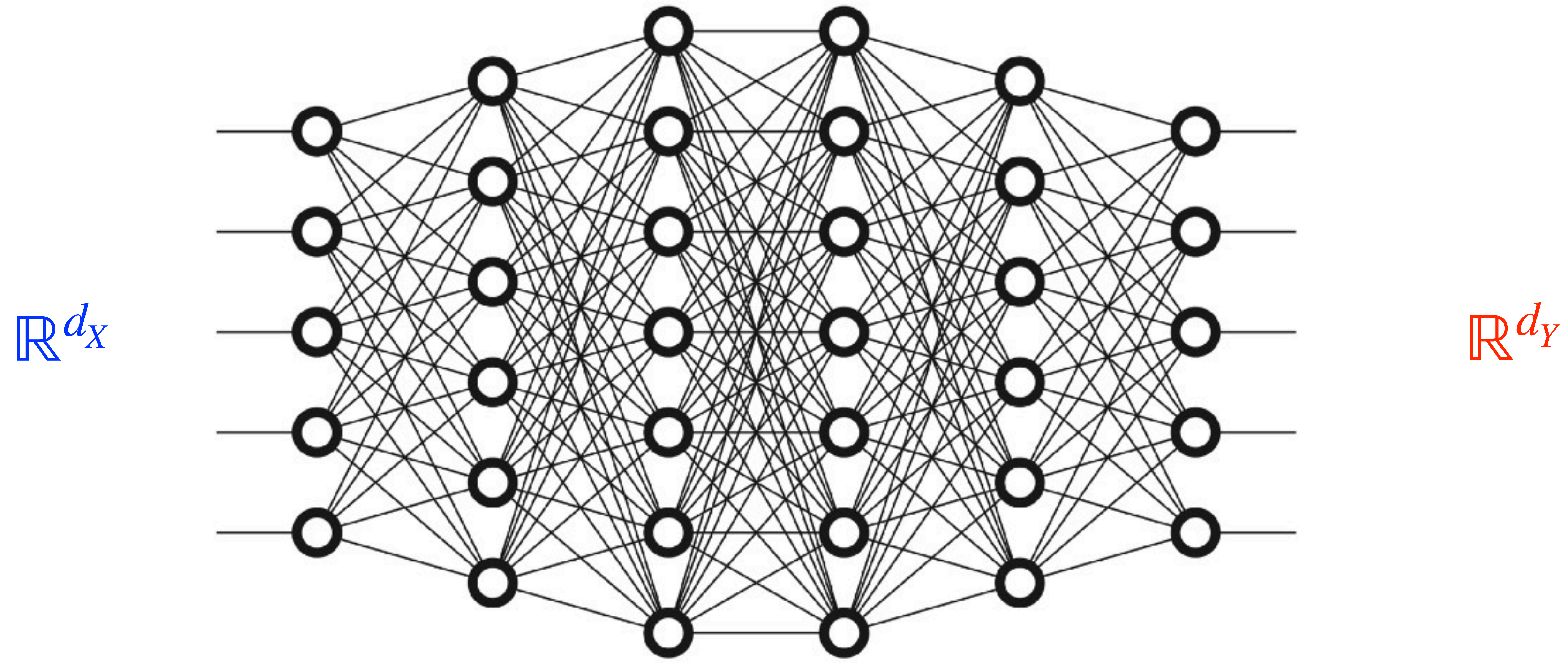
Where:

- Y_t - Outputs in $\mathbb{R}^{d_Y}$
- X_t - Covariates in $\mathbb{R}^{d_X}$
- V_t - Noise in $\mathbb{R}^{d_Y}$
- $f^\star$- Unknown Function in $\mathscr{F}$

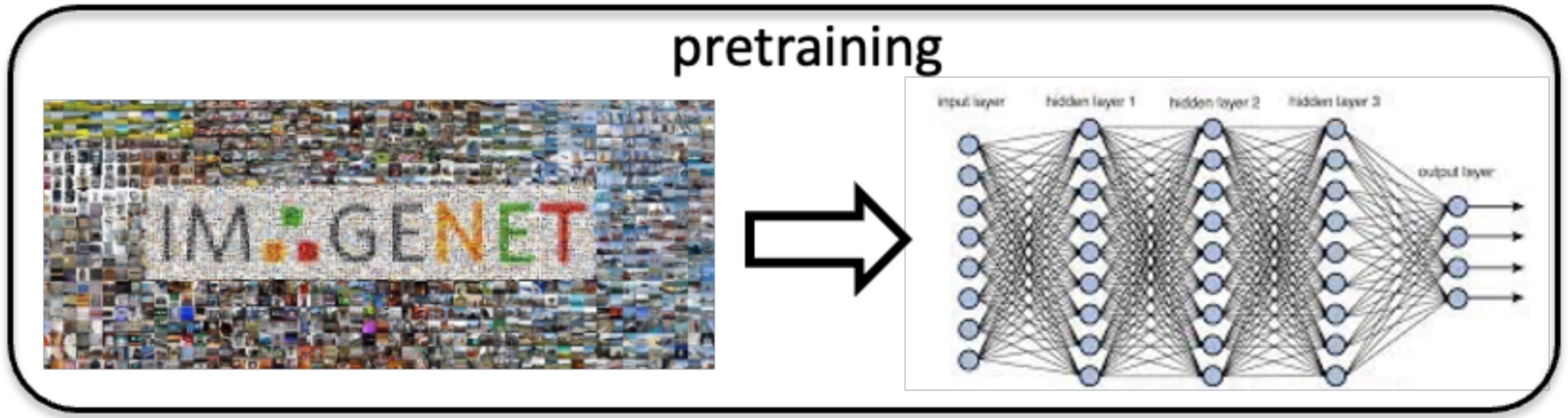
Empirical Risk Minimization

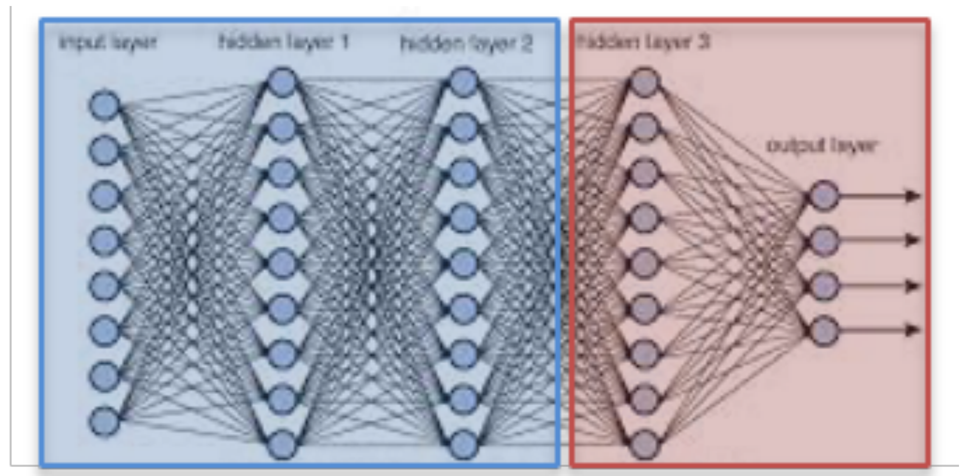
$$\hat{f} \in \operatorname{argmin}_{f \in \mathscr{F}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \| Y_t - f(X_t) \|_2^2 \right\}$$

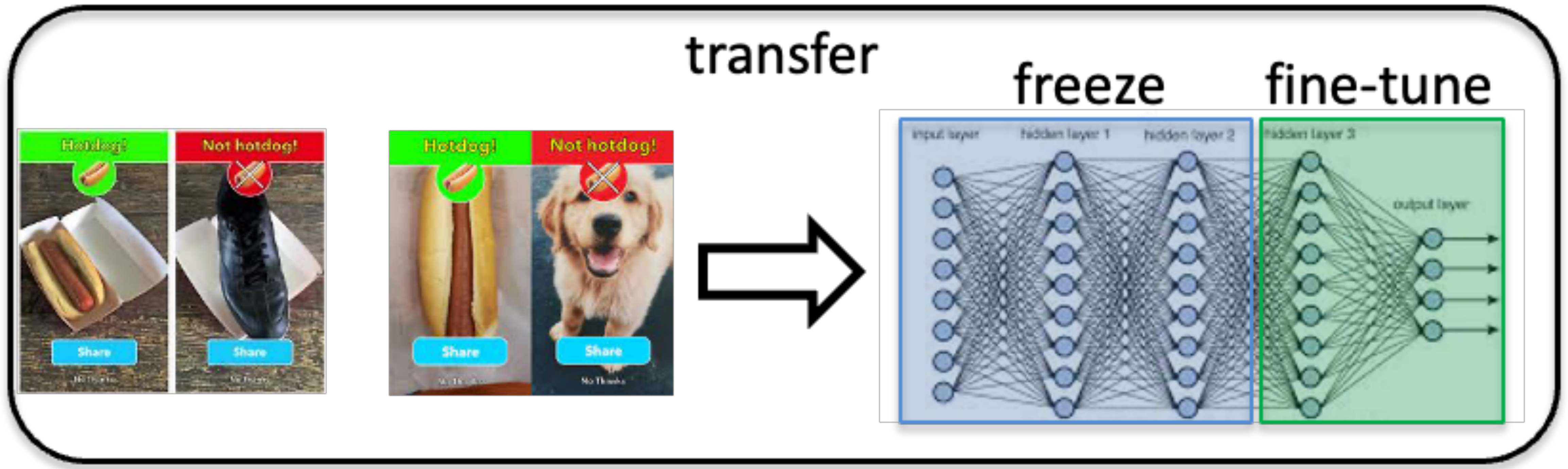
Example, $\mathscr{F}$ a parametric family:



Application: Representation Learning







Abstraction:

$$Y_t^j = \theta^j \phi^\star(X_t^j) + V_t^j, \qquad t \in [T]$$

Different tasks $j \in [H]$

Different final layer $\theta^j, j \in [H]$

Shared representation $\phi^\star$

Requires analysis of a two-stage estimator

but we can do that too using [Ziemann+ 2023]

iid [Du+ 2020], lin.dyn [Zhang+ 2023], gen.dyn [WIP]

~~Ordinary Least Squares~~

Empirical Risk Minimization

$$\cancel{\widehat{\theta} \triangleq \left(\sum_{t=1}^{T} Y_t X_t^\top\right)\left(\sum_{t=1}^{T} X_t X_t^\top\right)^{-1}}$$

$$\cancel{\Rightarrow \quad \|(\widehat{\theta} - \theta^\star)\sqrt{\Sigma_X}\|_F \lesssim (\text{noise scale}) \times \sqrt{\frac{d_X \times d_Y + \log(1/\delta)}{\text{sample size}}}}$$

Fresh/Test Sample $X'_{1:T}$

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

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\|\hat{f}(X'_t) - f_\star(X'_t)\|_2^2}$$

$$\Rightarrow \quad \|\hat{f} - f^\star\|_{L^2_X} \lesssim (\text{noise scale}) \times \sqrt{\frac{\text{complexity}(\mathscr{F}) + \log(1/\delta)}{\text{sample size}}}$$

Variational Form of the Empirical Risk

$$\hat{\theta} - \theta^\star = \left(\sum_{t=1}^{T} V_t X_t^\top\right)\left(\sum_{t=1}^{T} X_t X_t^\top\right)^{-1}$$

$$\hat{f} \in \operatorname{argmin}_{f\in\mathscr{F}} \left\{ \frac{1}{T}\sum_{t=1}^{T} \|Y_t - f(X_t)\|_2^2 \right\} \qquad \Rightarrow \frac{1}{T}\sum_{t=1}^{T} \|Y_t - \hat{f}(X_t)\|_2^2 \le \frac{1}{T}\sum_{t=1}^{T} \|Y_t - f^\star(X_t)\|_2^2 \qquad \star$$

$$Y_t = f^\star(X_t) + V_t, \quad t = 1,\ldots,T \qquad \star\star$$

$$\star + \star\star \Rightarrow \quad \frac{1}{T}\sum_{t=1}^{T} \|\hat{f}(X_t) - f^\star(X_t)\|_2^2 \le \sup_{f\in\mathscr{F}-\{f^\star\}} \frac{1}{T}\left(\sum_{t=1}^{T} 4\langle V_t, f(X_t)\rangle - \sum_{t=1}^{T} \|f(X_t)\|_2^2\right).$$

$$\text{linear model} \Rightarrow \qquad = \frac{4}{T}\left\| \left(\sum_{t=1}^{T} V_t X_t^\top\right)\left(\sum_{t=1}^{T} X_t X_t^\top\right)^{-1/2} \right\|_F^2$$

A Theorem

Suppose:

A1. Finite hypothesis class $|\mathscr{F}| < \infty$

A2. We have access to T/k independent stationary trajectories of length k

A3. $\mathrm{cond}_{\mathscr{F}} \triangleq \max_{f \in \mathscr{F}_\star} \max_{t \in T} \frac{\sqrt{\mathbf{E}\|f(X_t)\|_2^4}}{\mathbf{E}\|f(X_t)\|_2^2}$ is finite

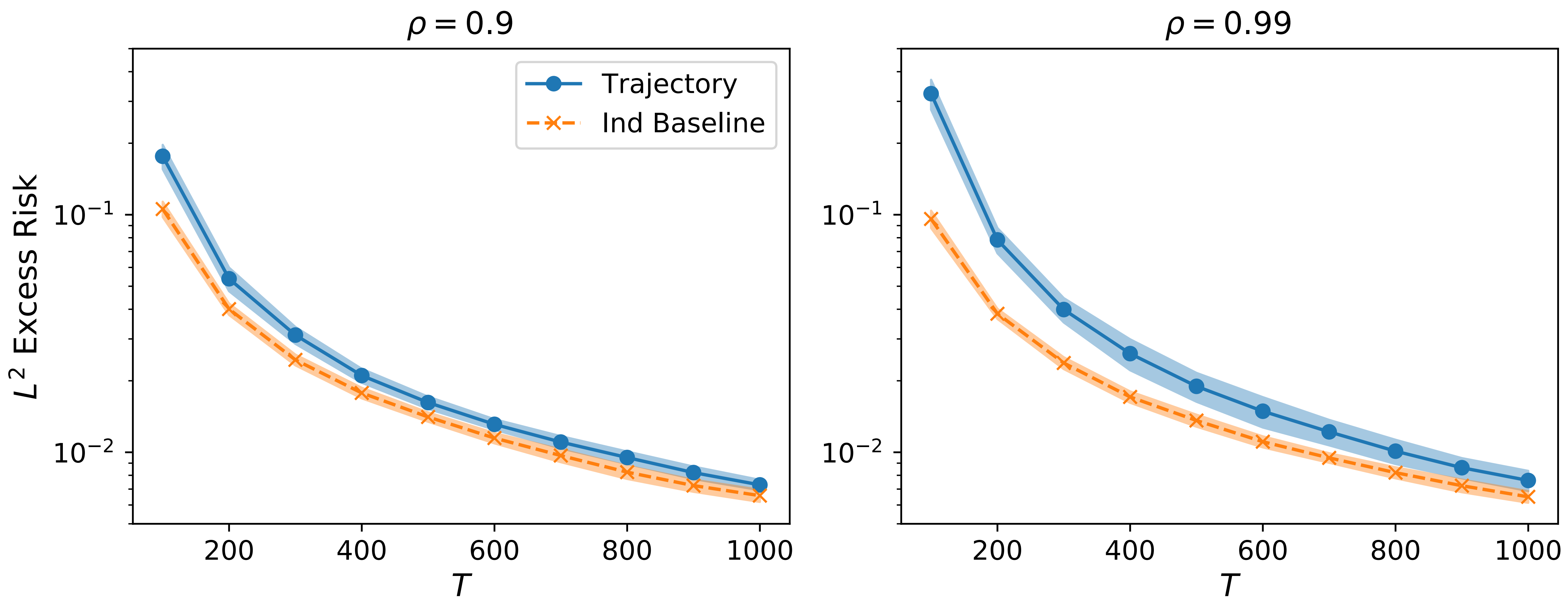
A4. For $t \in [T]$, $V_t \,|\, X_{1:t}$ is σ^2-conditionally sub-Gaussian and mean zero

Then:

$$\|\hat{f} - f^\star\|_{L^2_X}^2 \leq 16\sigma^2 \times \frac{\log(|\mathscr{F}|) + \log(1/\delta)}{T}$$

As long as:

$$T/k \geq 4\mathrm{cond}_{\mathscr{F}}^2 \left(\log|\mathscr{F}| + \log(2/\delta)\right)$$



$\Rightarrow$ consistent with above figure*

*A1 and A2 can be relaxed to:

All finite state Markov Chains,

GLM, RKHS, and compact

subsets of L^∞

Ziemann, Ingvar, and Stephen Tu. "Learning with little mixing." *Advances in Neural Information Processing Systems* 35 (2022): 4626-4637.

Proof Sketch

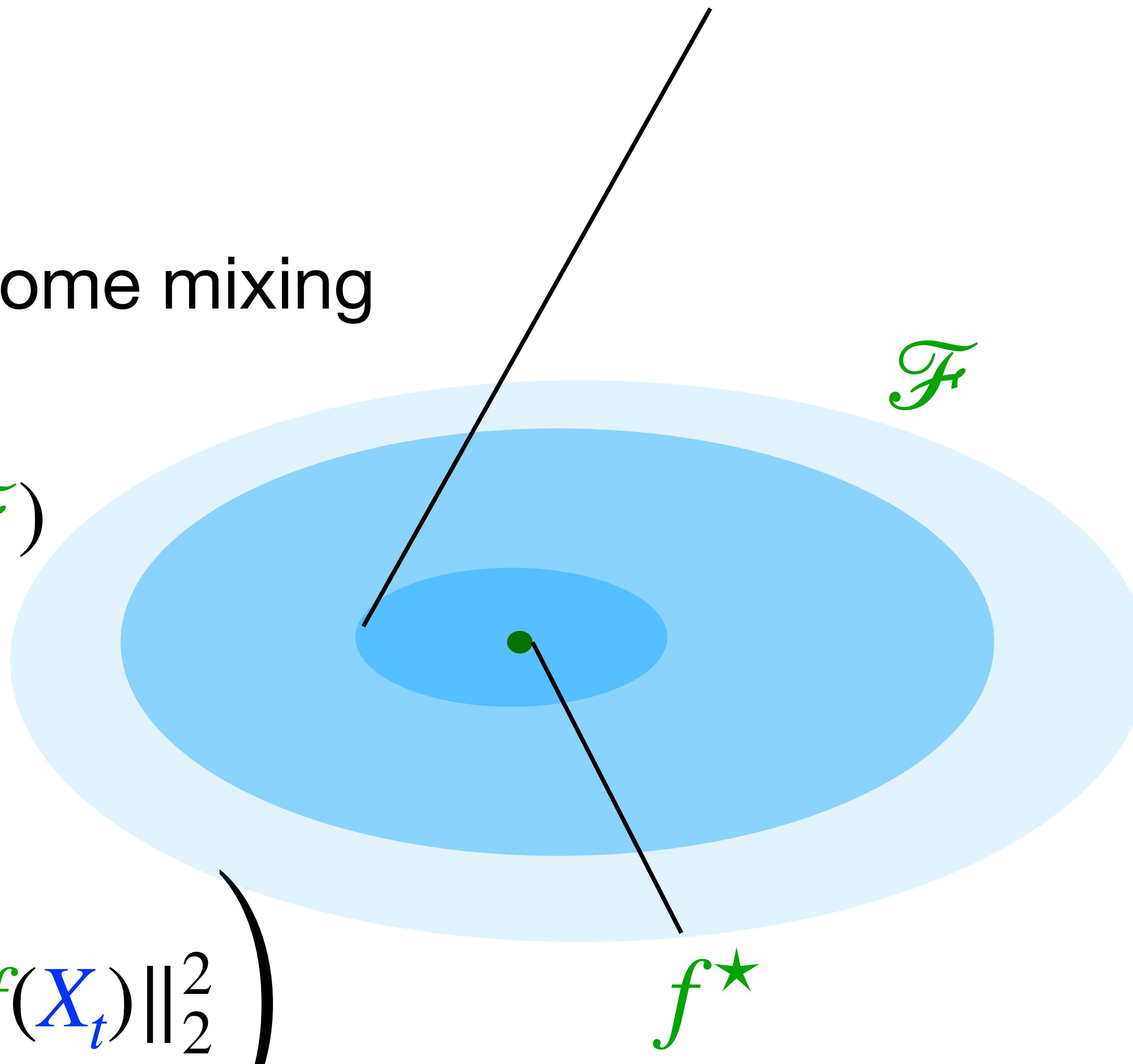
Step 1: prove a lower uniform estimate

Insight from [Mendelson 2014], lower uniform estimates are cheap—can use some mixing

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\|f(X_t')-f_\star(X_t')\|_2^2 \lesssim \frac{1}{T}\sum_{t=1}^{T} \|f(X_t')-f_\star(X_t')\|_2^2 \qquad (\forall f \in \mathscr{F})$$

population error

empirical error



Step 2: localize using the offset basic inequality [Liang+ 2015]

$$\frac{1}{T}\sum_{t=1}^{T} \|\hat{f}(X_t)-f^\star(X_t)\|_2^2 \le \sup_{f\in\mathscr{F}-\{f^\star\}} \frac{1}{T}\left(\sum_{t=1}^{T} 4\langle V_t, f(X_t)\rangle - \sum_{t=1}^{T} \|f(X_t)\|_2^2\right)$$

variational form of the empirical error

Step 3: combine

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{E}\|\hat{f}(X_t')-f_\star(X_t')\|_2^2 \lesssim \sup_{f\in\mathscr{F}-\{f^\star\}} \frac{1}{T}\left(\sum_{t=1}^{T} 4\langle V_t, f(X_t)\rangle - \sum_{t=1}^{T} \|f(X_t)\|_2^2\right) \lesssim (\text{noise scale})^2 \times \frac{\text{complexity}(\mathscr{F}) + \log(1/\delta)}{\text{sample size}}$$

Proof Sketch

Lower Uniform Estimate (step 1)

Lemma: for every $\lambda \in \mathbb{R}_+$

$$\mathbf{E}\exp\left(-\lambda\sum_{t=1}^{T}\|f(X_t)\|_2^2\right) \le \exp\left(-\lambda\sum_{t=1}^{T}\mathbf{E}\|f(X_t)\|_2^2 + \frac{\lambda^2 T}{2k}\mathbf{E}\left(\sum_{t=jT/k+1}^{(j+1)T/k}\|f(X_t)\|_2^2\right)^2\right)$$

Proof: integrate $e^{-x} \le 1 - x + x^2/2$ and $1 + x \le e^x$.

Proposition:

$$\mathbf{P}\left(\exists f \in \mathcal{F}_\star \ : \ \sum_{t=1}^{T}\|f(X_t)\|_2^2 < \frac{1}{2}\sum_{t=1}^{T}\mathbf{E}\|f(X_t)\|_2^2\right) \le |\mathcal{F}_\star|\exp\left(-\frac{T}{4k \times \mathrm{cond}_{\mathcal{F}}^2}\right)$$

Proof: Chernoff Bound, Tower Property and Union Bound the above lemma.

Proof Sketch

Controlling the supremum (step 3)

Lemma: $\forall \lambda \in [0,1/8\sigma^2]$ $\quad \mathbf{E} \exp\left(\lambda\left(\sum_{t=1}^{T} 4\langle V_t, f(X_t)\rangle - \sum_{t=1}^{T} \|f(X_t)\|_2^2\right)\right) \leq 1$

Proof: Use the tower property and that: $\left.\mathbf{E}_{t|1:t-1} \exp\left(\lambda\left(\langle V_t, f(X_t)\rangle - \|f(X_t)\|_2^2\right)\right)\right] \leq 1$

Lemma: $\Pr\left(\max_{f\in\mathscr{F}_\star}\left\{\sum_{t=1}^{T} 4\langle V_t, f(X_t)\rangle - \sum_{t=1}^{T} \|f(X_t)\|_2^2\right\} > u\right) \leq |\mathscr{F}_\star| \exp\left(\frac{-u}{8\sigma^2}\right)$

Proof: First lemma, Chernoff argument + union bound

Summary

- Gave an overview of recent advances in non-asymptotics for linear system identification
  - Tools from: machine learning, high-dimensional statistics/probability
- Provide a streamlined proof approach
  - Establish a lower uniform estimate (on the empirical covariance)
  - Combine with an upper bound on a self-normalized process
- Showed how this yields non-asymptotic guarantees for ARX(p,q) ID
- Extended the above program to nonlinear ID problems

Outlook

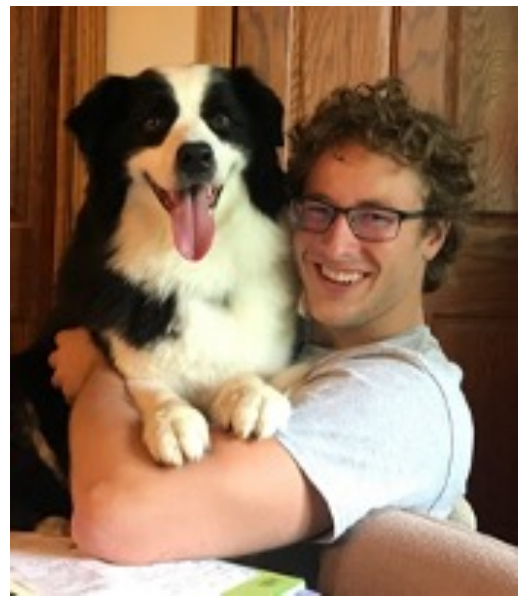
- Learning for control
  - Adaptive Control, Imitation Learning, Identification for Control
- Experiment Design
  - Some progress in the (semi-)linear setting [Wagenmaker+ 2021, 2023]— what about nonlinear?
- Co-design
  - Build systems that are easy to learn (to control)?
- Lack of realizability
  - Some progress in the linear setting [Ziemann+ 2023]
- Deep learning in the loop
  - Effect of SGD, implicit regularization and benign overfitting

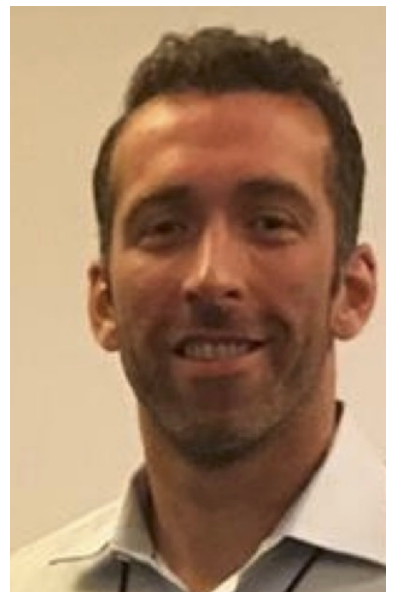
Thanks for listening!

https://arxiv.org/abs/2309.03873

contact: ingvarz@seas.upenn.edu








Ingvar Ziemann (Penn), Anastasios Tsiamis (ETH), Bruce Lee (Penn), Yassir Jedra (MIT), Nikolai Matni (Penn), George J. Pappas (Penn)