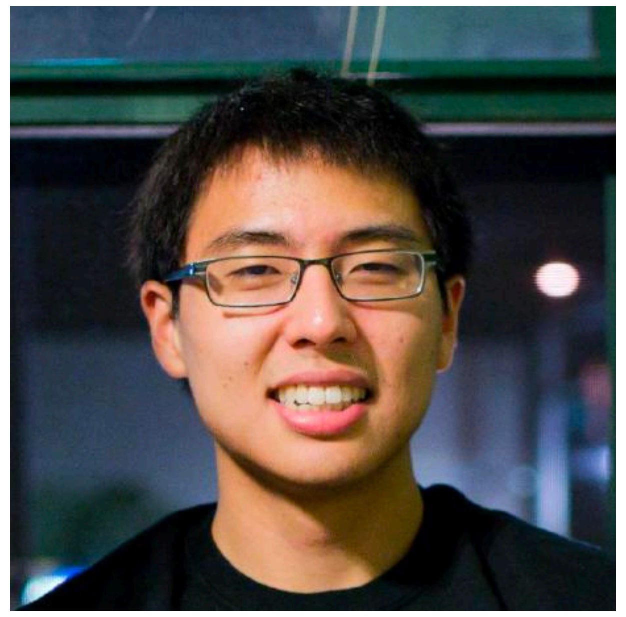
Sharp Rates in Dependent Learning Theory

https://arxiv.org/abs/2402.05928

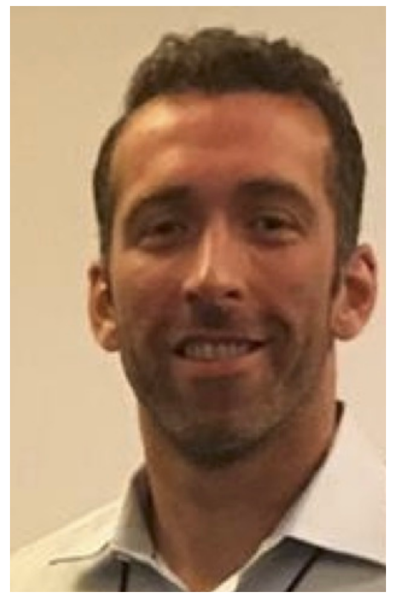
Ingvar Ziemann (UPenn), talk at PGM

Collaborators

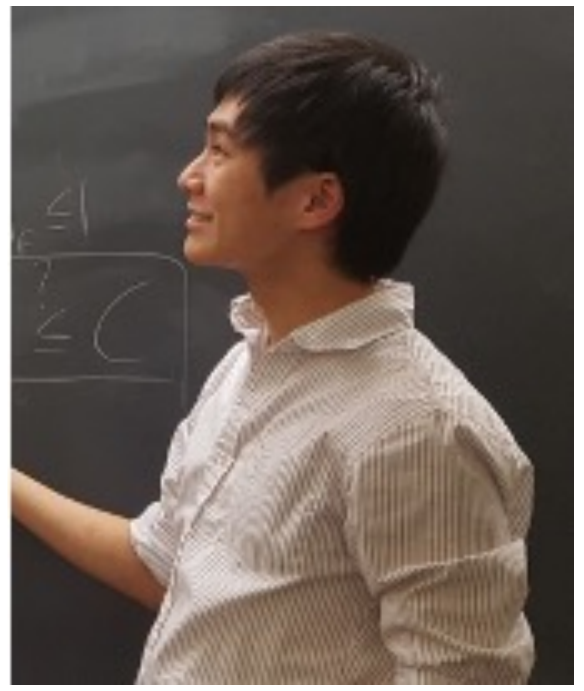


Stephen Tu (USC)

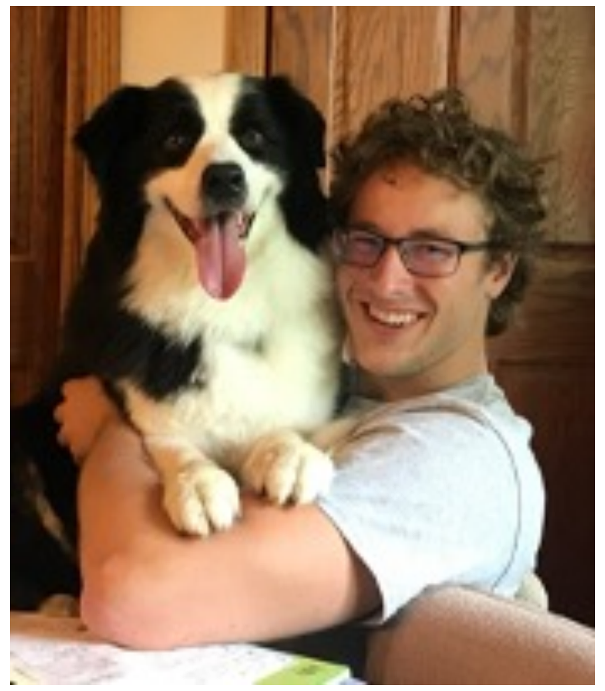
George J. Pappas (UPenn)



Nikolai Matni (UPenn)

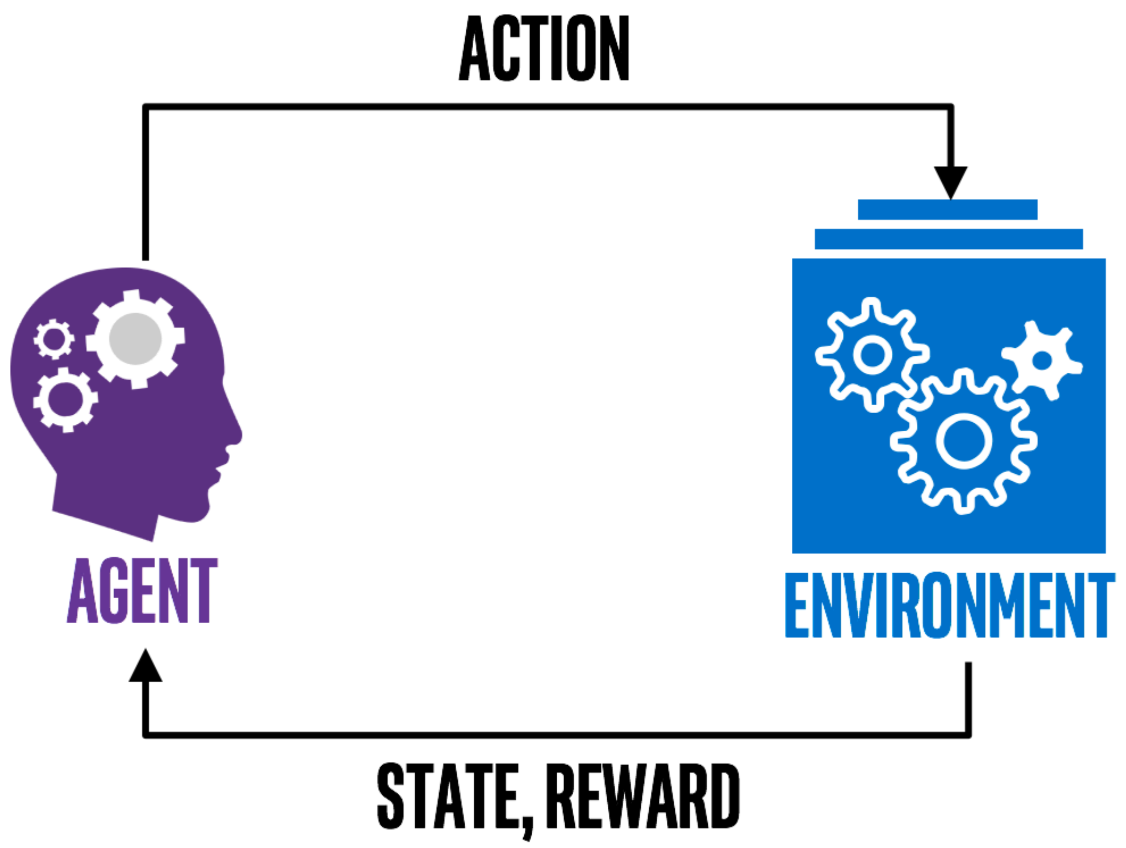


Thomas Zhang (UPenn)



Bruce Lee (UPenn)

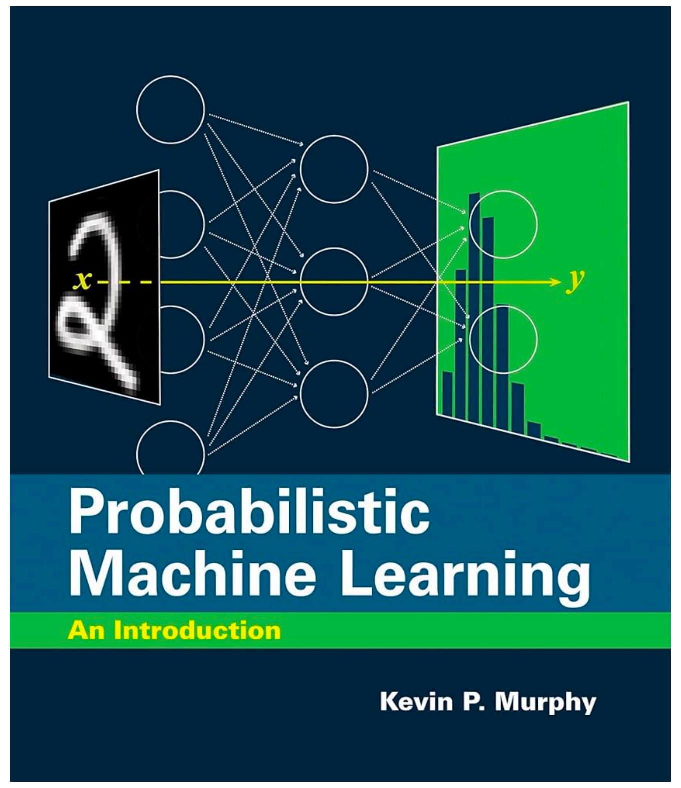
Dependent data is everywhere

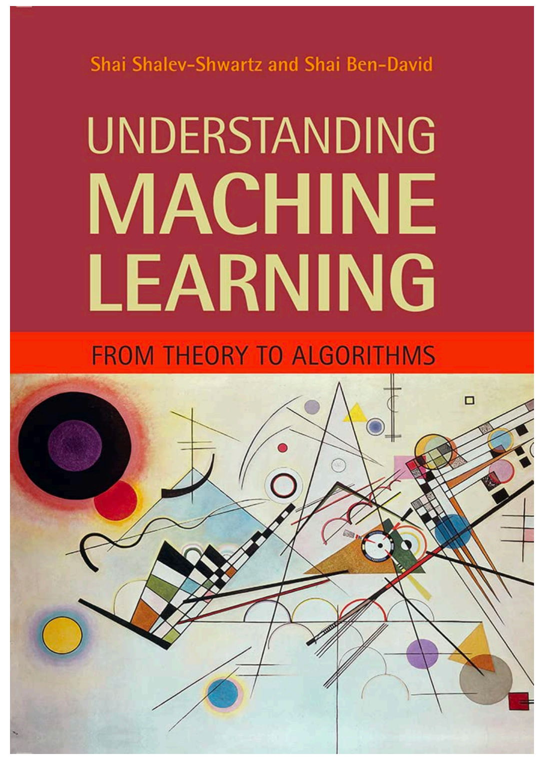


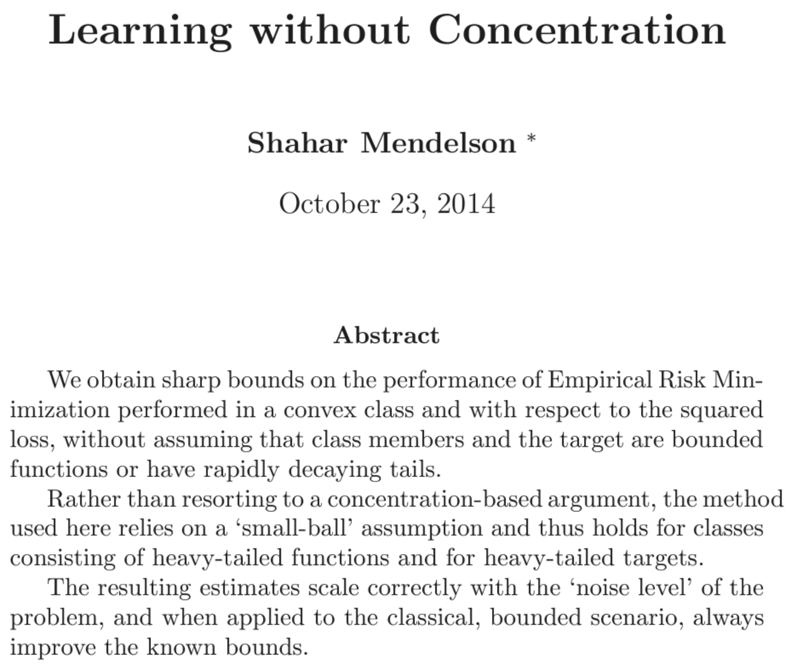
We understand iid learning very well

Uniform convergence, PAC, etc

Instance optimal (non-)asymptotics






Learning without Concentration

Shahar Mendelson *

October 23, 2014

Abstract

We obtain sharp bounds on the performance of Empirical Risk Minimization performed in a convex class and with respect to the squared loss, without assuming that class members and the target are bounded functions or have rapidly decaying tails.

Rather than resorting to a concentration-based argument, the method used here relies on a 'small-ball' assumption and thus holds for classes consisting of heavy-tailed functions and for heavy-tailed targets.

The resulting estimates scale correctly with the 'noise level' of the problem, and when applied to the classical, bounded scenario, always improve the known bounds.

Dependent data is less well understood

1: Correct notion of dependence?

2: Optimal rates for some reasonable notion of dependence?

Today: make some headway on 2

Focus on supervised learning with

square loss $\quad l_{\mathrm{sq}}(f, x, y) = \|y - f(x)\|^2$

Notion of Weak Dependence: Mixing

We will consider the **mixing** case:

- $\mathbf{E}_{Z_{1:t}}[\mathrm{dist}(\mathbf{P}(Z_{t+k} \in \cdot \mid Z_{1:t}), \mathbf{P}(Z_{t+k} \in \cdot\,))] \leqslant \beta(k) \to 0$ as $k \to \infty$.
- Often we also assume that $\{Z_t\}$ converges to a stationary measure.

Mixing time is defined as:

$$t_{\mathrm{mix}}(\varepsilon) := \min\{k \in \mathbb{N}_+ \mid \beta(k) \leqslant \varepsilon\}$$

Example, Linear Dynamical Systems:

$$Z_{t+1} = aZ_t + W_t, W_t \sim N(0,1) \quad \Rightarrow \quad t_{\text{mix}} \approx \frac{1}{1 - |a|} \text{ if } |a| < 1$$

The Classical Proof Approach: Blocking

Classical results in supervised learning rely on *blocking* [Yu1994, Bernstein1927,…]

Transforms n dep. samples into $m = n/t_{\text{mix}}$ independent "blocks"

$$Z_{1:n} \Rightarrow \tilde{Z}_{1:t_{\text{mix}}}, \tilde{Z}_{k+1:2t_{\text{mix}}}, \tilde{Z}_{2k+1:3t_{\text{mix}}}, \ldots \text{ (m independent blocks)}$$

Can now apply standard results for independent data to the "blocks"

Generically employed, deflates rate of converge by a factor of the mixing time t_{mix}…

Classical asymptotics tell us this is **not optimal**!

Goal: an instance-optimal non-asymptotic theory of learning from dependent data

How Many Samples Do I Need?

Overview of today's talk

Focus on supervised learning with

square loss $l_{\mathrm{sq}}(f, x, y) = \|y - f(x)\|^2$

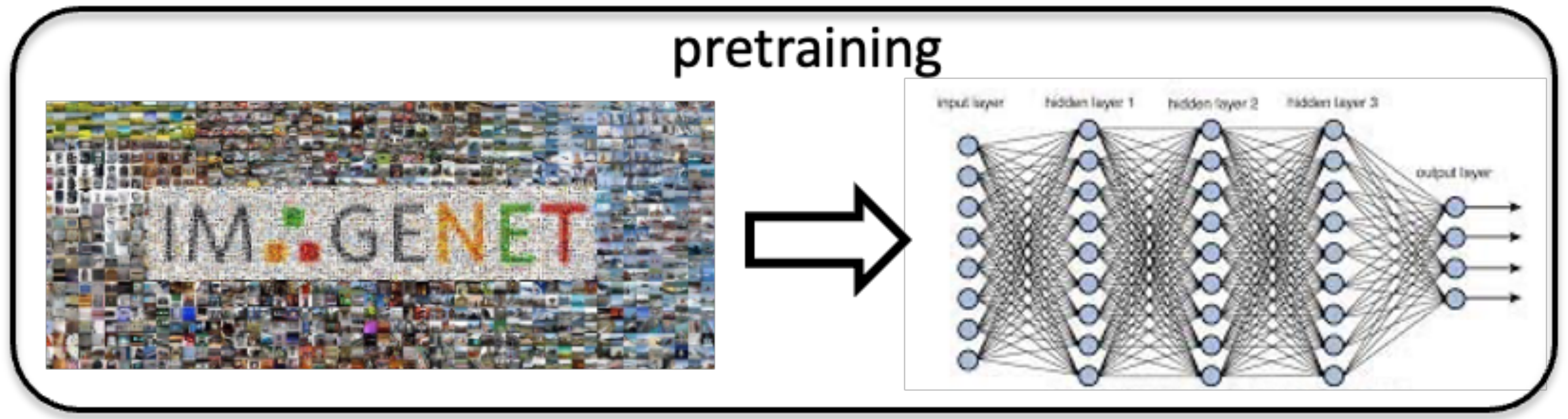
Given: Dependent data $Z_{1:n} = ({\color{blue}X}, {\color{red}Y})_{1:n}$ and a hypothesis class ${\color{green}\mathcal{F}}$

Output: a predictor ${\color{green}\hat{f}} \in {\color{green}\mathcal{F}}$ that "minimizes":

$$\mathbf{ER}({\color{green}\hat{f}}) \triangleq \mathbf{E}_{{\color{blue}X},{\color{red}Y}}\|{\color{green}\hat{f}}({\color{blue}X}) - {\color{red}Y}\|^2 - \min_{{\color{green}f \in \mathcal{F}}} \mathbf{E}_{{\color{blue}X},{\color{red}Y}}\|{\color{green}f}({\color{blue}X}) - {\color{red}Y}\|^2$$

Application: Representation learning

for time-series/dynamical systems

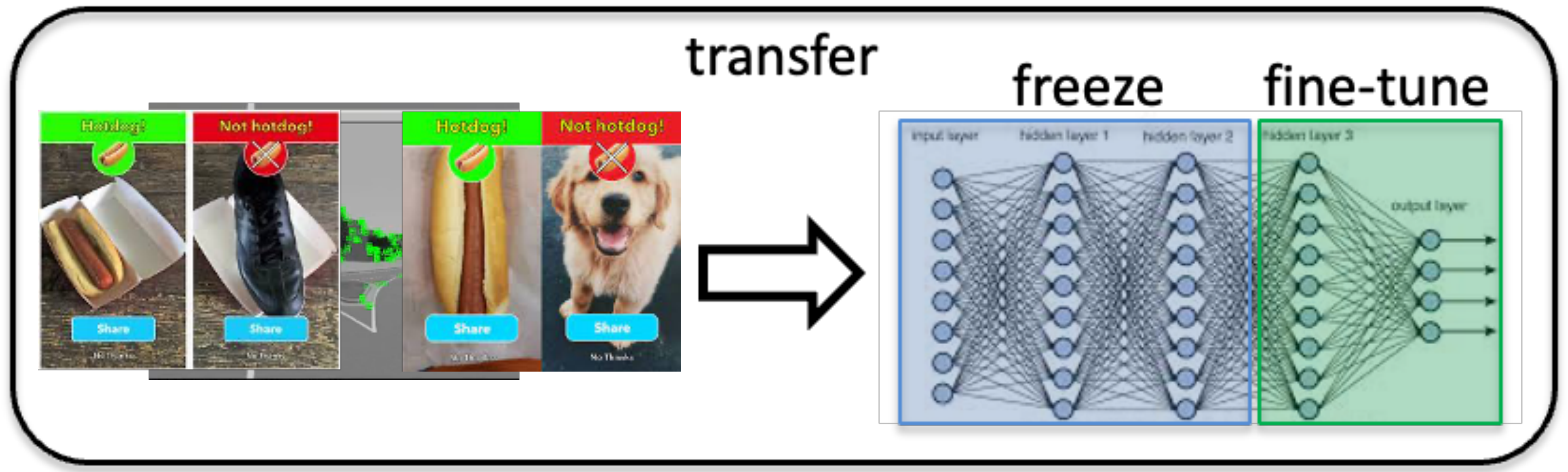


Main: establish instance optimal rates for ERM

Distinguish between:

$${\color{green}\hat{f}} \in \mathrm{argmin}_{{\color{green}f \in \mathcal{F}}} \frac{1}{n} \sum_{i=1}^{n} \|{\color{green}f}({\color{blue}X_i}) - {\color{red}Y_i}\|^2$$

- *realizable data*
- *non-realizable* data



Example: System Identification (Next Slide)

Example: System Identification

ARX(p,q)

$$Y_t = \sum_{i=1}^{p} A_i^\star Y_{t-i} + \sum_{j=1}^{q} B_i^\star U_{t-j} + W_t$$

In other words…

$$X_t = \begin{bmatrix} Y_{t-1:t-p}^\top & U_{t-1:t-q}^\top \end{bmatrix}^\top$$

$$\theta^\star = \begin{bmatrix} A_{1:p}^\star & B_{1:q}^\star \end{bmatrix} \qquad \mathscr{F} \simeq \mathbb{R}^{d_Y \times d_X}$$

$W_t \sim N(0, I)$, drawn iid

ERM = Ordinary Least Squares Estimator:

$$\hat{\theta} \in \operatorname{argmin}_{\theta \in \mathbb{R}^{d_Y \times d_X}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \| Y_t - \theta X_t \|_2^2 \right\}$$

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

$$\mathbf{ER}(\hat{\theta}) = \| (\hat{\theta} - \theta_\star) \sqrt{\Sigma_X} \|^2$$

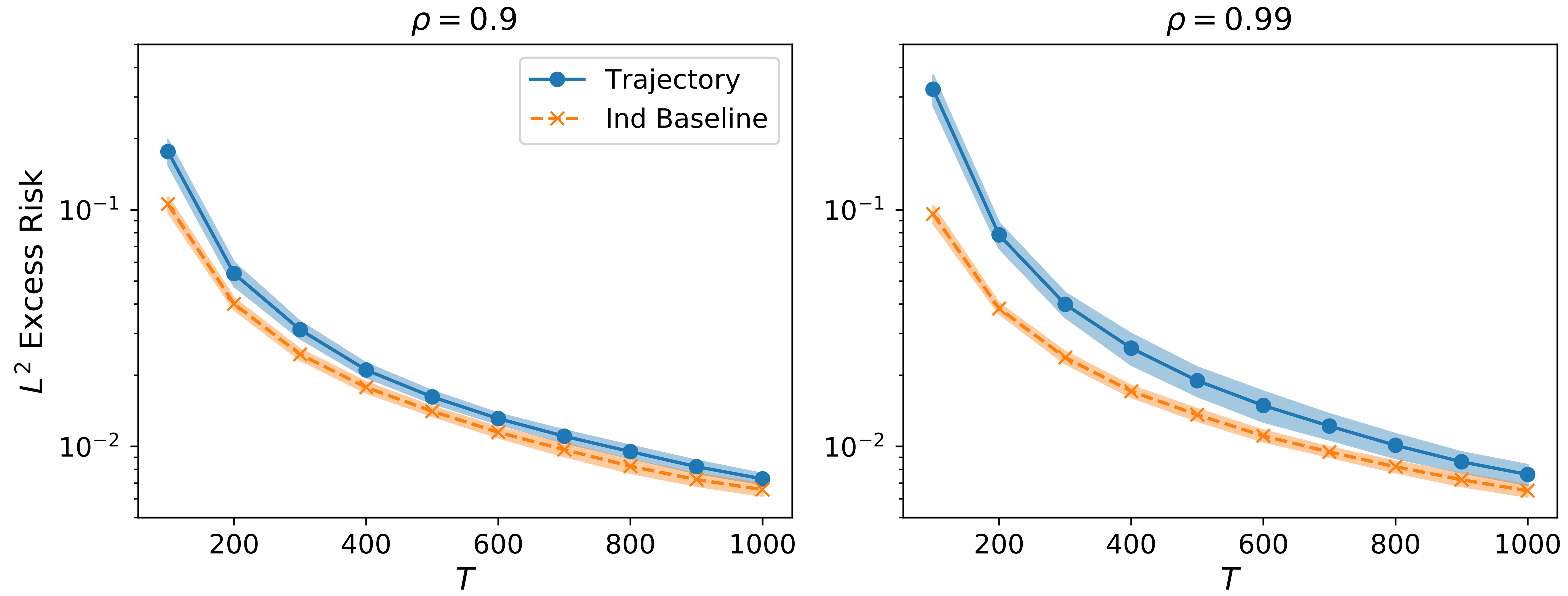
$\Sigma_X = \mathbf{E}[XX^\top]$ $\quad$ $X \sim$ stationary dist of X_t

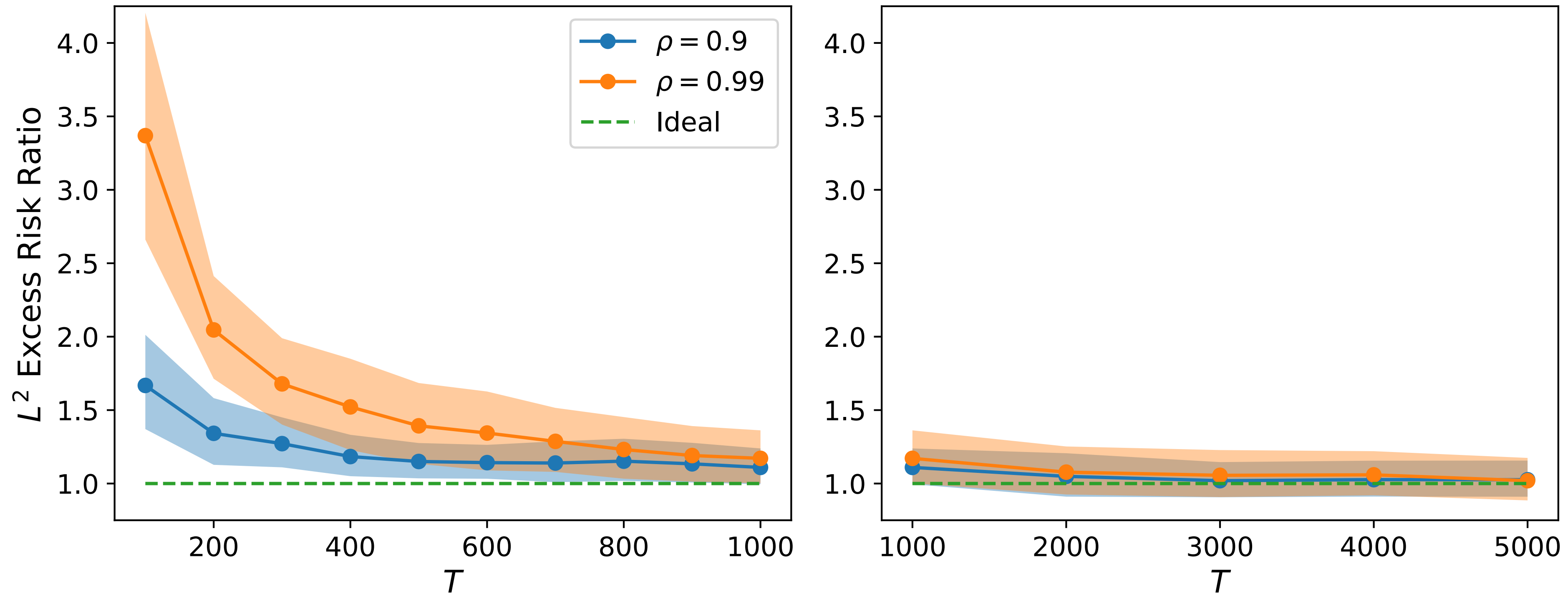
Dependency Deflation?

Stable GLM, numerical ERM experiment

GLM: $Y_t \triangleq X_{t+1} = \phi(A_\star X_t) + W_t$, w/ ϕ a known link function (here leaky relu), $W_t \sim N(0, \sigma_W^2)$ iid

Dependency ($\rho = \rho(A_\star)$) doesn't seem to hurt (at least not for large T):





But blocking yields results of the form:

$$\|\hat{f} - f_\star\|_{L^2}^2 \lesssim t_{\text{mix}} \sigma_W^2 \times \frac{\text{complexity(hyp . class)}}{n}$$

$$t_{\text{mix}} \approx (1-\rho)^{-1}$$

Today: improve this term

Overview of Results: dep. data & square loss

Paper	Hypothesis Class	Rate (Hiding logs)	Mixing/ Stability	Realizability
[SMTJR2018]	Linear	$n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Marginal	Required
[SO2020(2)]	Generalized Lin	$t_{\text{mix}} n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Strict	Required
[KNJN2021]	Generalized Lin	$n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Marginal	Required
[RBE2021]	**General**	$t_{\text{mix}} n^{-1+\epsilon}\sigma^2 \times \text{complexity}(\mathscr{F})$	Strict	**Not Required**
[**Z**T2022]	**General** (4-2 hypcon)	$n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Strict	Required
[SOO2022]	Bilinear	$n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Marginal	Required
[**Z**TPM2023]	Linear	$n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Strict	**Not Required**
[**Z**TPM2024]	**General** (weak subG)	$n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Strict	**Not Required**

A Tutorial on the Non-Asymptotic Theory of System Identification

Ingvar Ziemann[1], Anastasios Tsiamis[2], Bruce Lee[1], Yassir Jedra[3], Nikolai Matni[1], and George J. Pappas[1]

[1]University of Pennsylvania
[2]ETH Zürich
[3]KTH Royal Institute of Technology

Abstract

This tutorial serves as an introduction to recently developed non-asymptotic methods in the theory of—mainly linear—system identification. We emphasize tools we deem particularly useful for a range of problems in this domain, such as the covering technique, the Hanson-Wright Inequality and the method of self-normalized martingales. We then employ these tools to give

Sharp Rates without Realizability

$n^{-1}\sigma^2$: signal to noise ratio

t_{mix} : mixing time

[**Z**TLJNP2023] *A Tutorial on the Non-Asymptotic Theory of System Identification*, CDC 2023

Realizability

Given: Dependent data $(X, Y)_{1:n}$ and a hypothesis class $\mathscr{F}$

Output: a predictor $\hat{f} \in \mathscr{F}$ that "minimizes":

$$\mathbf{ER}(\hat{f}) \triangleq \mathbf{E}_{X,Y} \|\hat{f}(X) - Y\|^2 - \min_{f \in \mathscr{F}} \mathbf{E}_{X,Y} \|f(X) - Y\|^2$$

$$f_\star \in \operatorname{argmin}_{f \in \mathscr{F}} \mathbf{E}_{X,Y} \|f(X) - Y\|^2$$

$$\mathbf{E}_{X,Y} \|f_\star(X) - Y\|^2$$

Realizability: $Y_i = f_\star(X_i) + W_i$, for some MDS "white noise" $W_{1:n}$

Absence of Realizability = Agnostic: no general relation between X and Y $\Rightarrow$ life is significantly harder

Example:

$$Y_t = \sum_{i=1}^{p} A_i^\star Y_{t-i} + \sum_{j=1}^{q} B_i^\star U_{t-j} + W_t$$

realizable iff $\mathscr{F}$ searches over the correct model order (p,q)

Dependent Linear Regression

Theorem [**Z**TMP23]: Suppose that $\{(X_t, Y_t)\}_{t\geq 1}$ is stationary, mixing, and $\sqrt{\mathbf{E}\langle v, X\rangle^4} \leqslant h^2 \cdot \mathbf{E}\langle v, X\rangle^2, v \in \mathbb{S}^{d_X-1}$

Let $\mathscr{F} = \mathbb{R}^{d_X}$, then as long as n is greater than a burn-in, w.p. at least $1-\delta$:

$$\mathbf{ER}(\hat{f}) \lesssim \frac{\mathrm{tr}\Sigma + \|\Sigma\|_{\mathrm{op}} \log(1/\delta)}{n}$$

$$\Sigma := \lim_{n\to\infty} \frac{1}{n}\mathbf{E}\left[\left(\sum_{i=1}^{n} \Sigma_X^{-1/2} X_i W_i\right)\left(\sum_{i=1}^{n} \Sigma_X^{-1/2} X_i W_i\right)^{\mathsf{T}}\right], W_i = f_\star(X_i) - Y_i \quad \Sigma_X = \mathbf{E}[XX^{\mathsf{T}}]$$

Realizable ERM:

$$\mathbf{ER}(\hat{f}) \lesssim \sigma_W^2\left(\frac{d_X + \log(1/n)}{n}\right)$$

Key: The variance Σ interpolates between realizable and non-realizable regimes:

- $Y_i = f_\star(X_i) + W_i$, $W_{1:n}$ martingale a difference sequence $\Rightarrow \mathrm{tr}(\Sigma) = d_X \mathbf{V}(W)$
- If $(X_1, Y_1) = (X_2, Y_2) = \ldots (X_k, Y_k)$ and $(X_{k+1}, Y_{k+1}) = (X_{k+2}, Y_{k+2}) = \ldots$ $\Rightarrow \mathrm{tr}(\Sigma) = k d_X \mathbf{V}(W)$

The latter case is the "worst case" non-realizable distribution

The Noise Level in Dependent Linear Regression: Ziemann, Tu, Pappas and Matni, NeurIPS 2023

Sharp Rates in Dependent Learning Theory

Theorem [**Z**TPM2024]**:** *Let $\mathscr{F}$ be either 1) convex or 2) realizable class and let $\hat{f}$ be the ERM for β-mixing stationary data* $(X, Y)_{1:n}$. Suppose further that:

- $\|f\|_{\Psi_p} \leq L\|f\|_{L^2}^{\eta}$ for some $\eta \in (0,1]$ where $\|f\|_{\Psi_p} \triangleq \sup_{m\geq 1} m^{-1/p}\|f(X)\|_{L^m}$
- n is greater than a polynomial in problem constants

We have that with probability at least $1-\delta$:

$$\mathbf{ER}(\hat{f}) \lesssim \sigma^2 \times \frac{\text{complexity}(\mathscr{F}) + \log(1/\delta)}{n}$$

The Blue Condition Holds:

- For finite hypothesis classes
- For smooth hypothesis classes
- For parametric hypothesis classes (Loja)

Key takeaway: Under realizability (i.e., $\mathbf{E}[Y_t \mid X_t] = f_\star(X_t)$), the leading variance proxy term tends to the **same variance** in the independent case

dependence only hurts without realizability

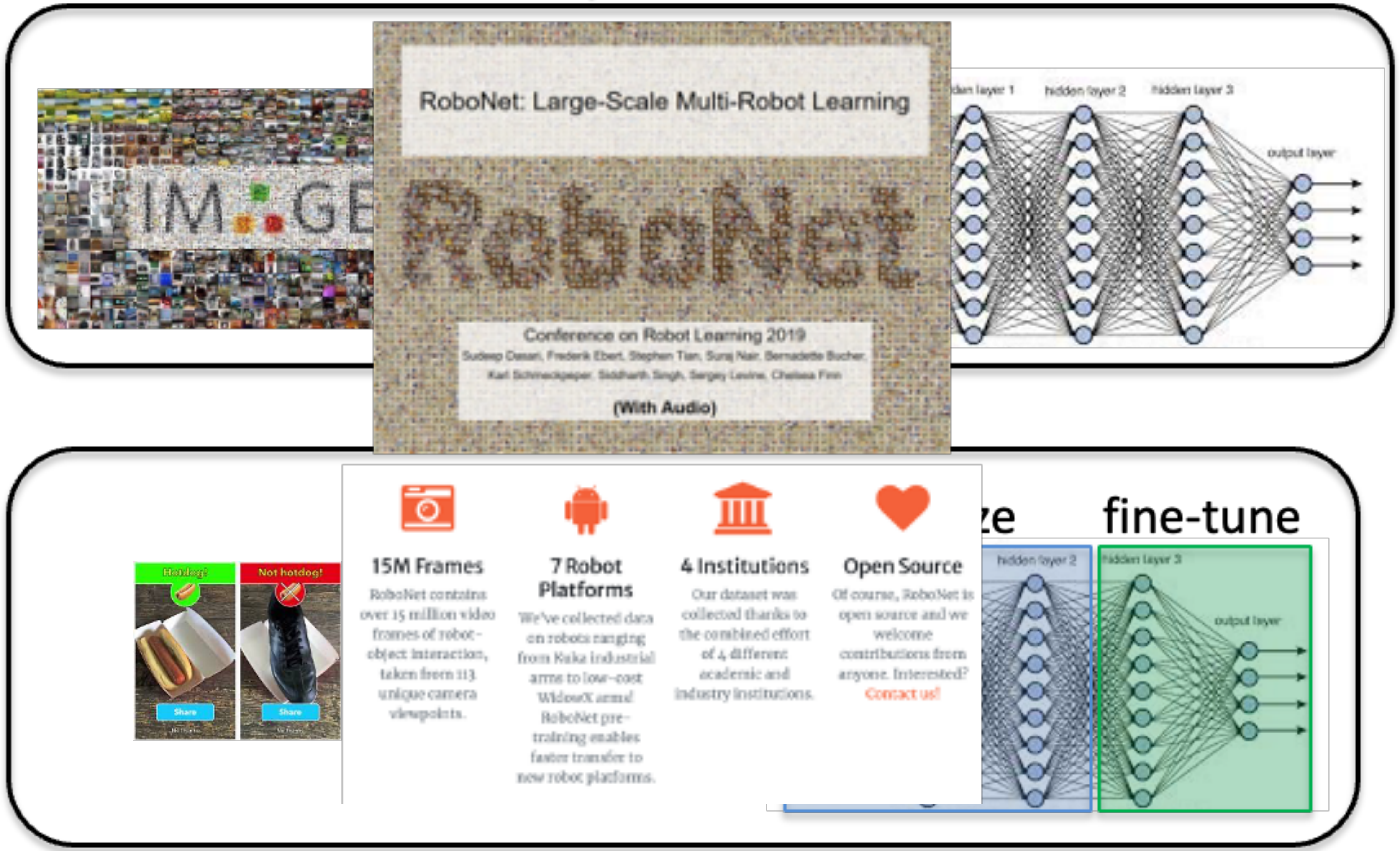
$$\sigma^2 \triangleq \lim_{n\to\infty} \sup_{g\in(\mathscr{F}-\mathscr{F})\cap o(1)S_{L^2}} \mathbf{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\langle W_i, \frac{g(X_i)}{\|g\|_{L^2}}\right\rangle\right) = \mathbf{Var}(W) \text{ if strictly realizable}$$

rS_{L^2}: rad r sphere in L^2

Multi-task Representation Learning

Pretrain for the representation $g_\star$

Robonet [Berkley, CMU, Penn, Stanford]



Math Vignette

T tasks n datapoints per task

$$Y_{i,t} = h^t_\star g_\star(X_{i,t}) + W_{i,t} \quad i \in [n], t \in [T]$$

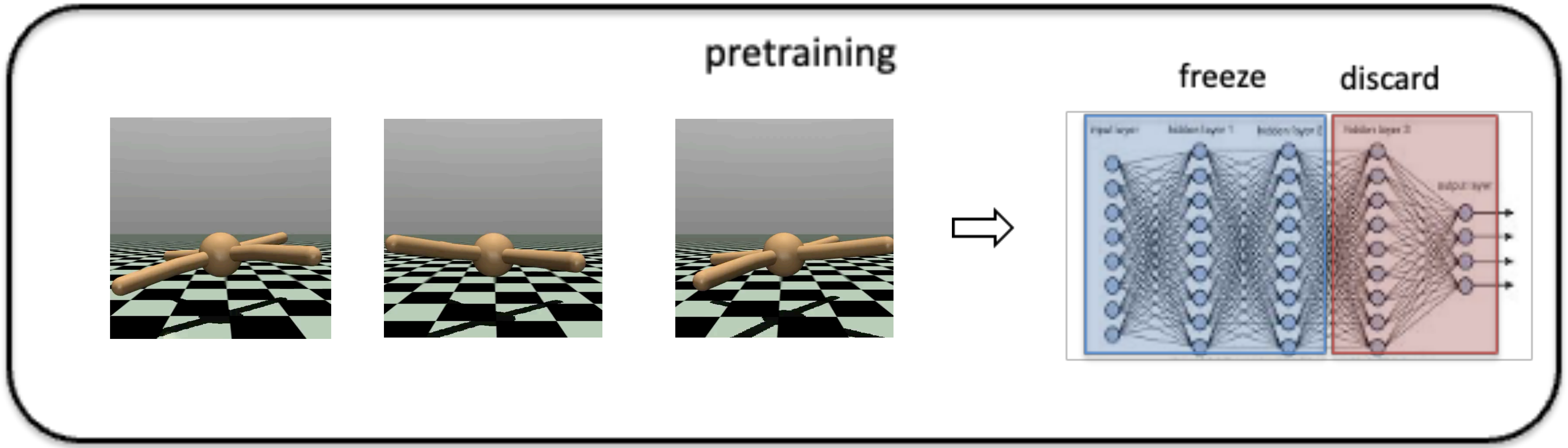
Target task "0", also with n datapoints

$$Y_{i,0} = h^t_\star g_\star(X_{i,0}) + W_{i,0} \qquad i \in [n]$$

Want to perform well on task "0" using **all**

$(T+1) \times n$ datapoints

Representation Learning for Control

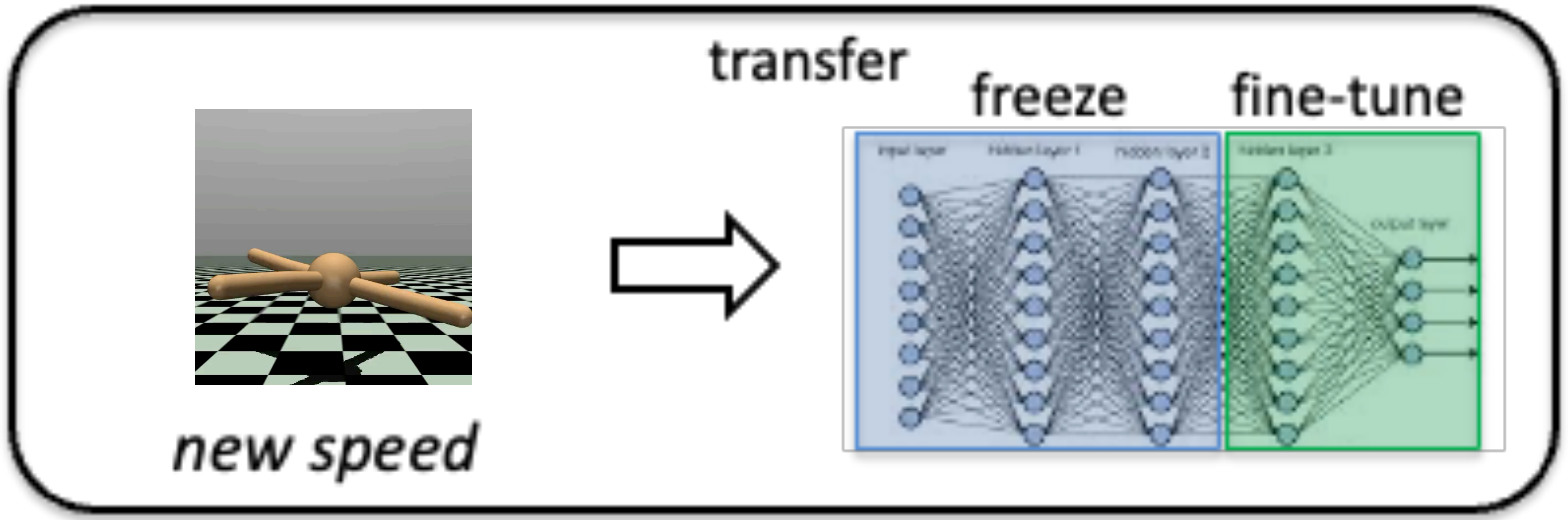


Learn a controllers for

different speeds

Adapt to a new

speed with less data



Nonlinear Representation Learning "ERM"

Nonlinear rep class $\mathscr{G}$ and linear heads $\mathscr{H} \equiv \mathbb{R}^{d_Y \times r}$

E.g. fine-tuning last linear layer of neural network

Rep fit on training tasks:

$$(\{\hat{h}^t\}_{t=1}^T, \hat{g}) \in \underset{h^t, g}{\operatorname{argmin}} \frac{1}{nT} \sum_{t=1}^{T} \sum_{i=1}^{n} \|Y_{i,t} - h^t g(X_{i,t})\|^2$$

Target head fit on target data *passed through* $\hat{g}$

$$\hat{h}^0 = \underset{h^0}{\operatorname{argmin}} \sum_{i=1}^{n} \|Y_{i,0} - h^0 \hat{g}(X_{i,0})\|^2$$

Goal: bound excess risk of the two-stage *empirical risk minimizers* $(\hat{h}^0, \hat{g})$

State of Affairs for the two-stage ERM

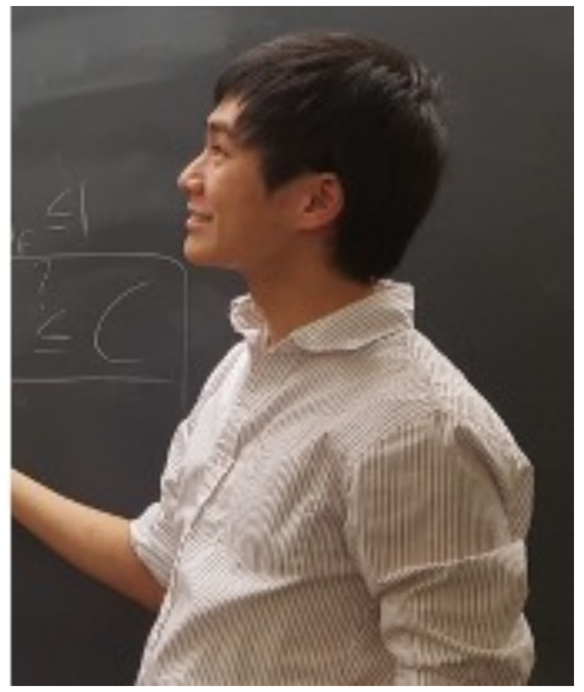
	iid Covariates	Trajectory Data	Order r Samples per Task
Linear Representation	✓	✓	✗
Nonlinear Representation	✓	✗	✗

[DHKLL2021]

require d_X samples per task

[ZTAM2024]

algorithmic soln' to this

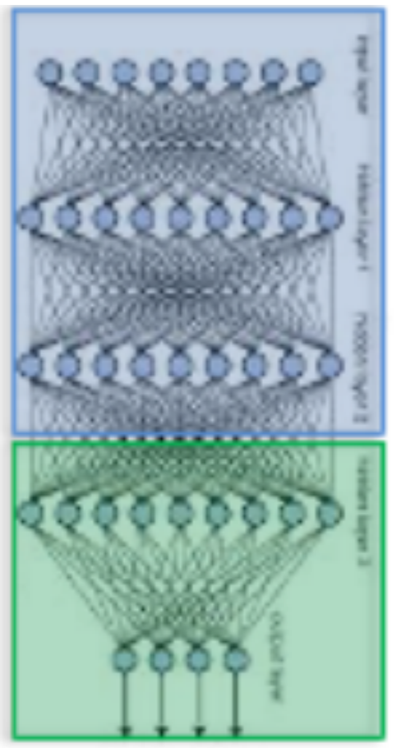


Guarantees for Nonlinear Representation Learning

Theorem [ZL**Z**PM2024]: As long as $n \gtrsim d_Y r + \mathrm{Comp}(\mathscr{G})/T$ we have that:

$$\mathbf{ER}(\hat{h}^{(0)}, \hat{g}) \lesssim \sigma_W^2 (1 + C_{\mathbf{X}} C_{\mathbf{h}}) \left[\frac{d_Y r}{n} + \left(\frac{\mathrm{Comp}(\mathscr{G})}{nT} \right) \right]$$

Rep Learning: many tasks ⇒ small



fine-tuning: only the last layer ⇒ small

Capture **two sources of task relatedness**:

$C_{\mathbf{X}}$: for all h, h', g, g'

$$\mathbf{E}_{X_0} \| h \circ g(X_0) - h' \circ g'(X_0) \|^2 \leq \frac{C_{\mathbf{X}}}{T} \sum_{t=1}^{T} \mathbf{E}_{X_t} \| h \circ g(X_t) - h' \circ g'(X_t) \|^2$$

- "Avg overlap of covariate distributions $\mathbf{P}_{X_0}$ vs $\mathbf{P}_{X_t}, t \in [T]$"

$C_{\mathbf{h}}$: $\mathbf{h}^{(t)} \triangleq h_\star^{(t)\top} h_\star^{(t)}, \forall t$. $C_{\mathbf{h}} = \left\| \mathbf{h}^{(0)\top} \left(T^{-1} \sum_{t=1}^{T} \mathbf{h}^{(t)} \right)^{-1} \right\|$

- "Avg overlap of task-specific heads $h_\star^0$ vs $h_\star^t, t \in [T]$"

If $\mathscr{H}, \mathscr{G}$ linear:

$$C_{\mathbf{X}} \mathbf{E}_{X_{1:T}} X X^\top \succeq \mathbf{E}_{X_0} X X^\top$$

Conclusion

References

[**Z**T2022] Learning with little mixing, NeurIPS 2022
https://arxiv.org/abs/2206.08269

[**Z**TPM2023] The Noise Level in Dependent Linear Regression, NeurIPS 2023
https://arxiv.org/abs/2305.11165

[**Z**TPM2024] Sharp Rates in Dependent Learning Theory, ICML 2024
https://arxiv.org/abs/2402.05928

[ZL**Z**PM2024] Guarantees for Nonlinear Representation Learning, ICML 2024

- Now have a sharp theory for learning with dependent data for the square loss
  - Other loss functions? Strongly convex should not be too hard
- Used this to understand representation learning in dynamical systems
- Follow up work/applications
  - Streaming Algorithms meet dependence
  - Learning without mixing
    - Can we find a unified proof approach for β-mixing and martingale setups?

Thanks for Listening!

ingvarz@seas.upenn.edu