Sharp Rates in Dependent Learning Theory: Avoiding Sample Size Deflation for the Square Loss

Ingvar Ziemann (Penn), Stephen Tu (USC),
George J. Pappas (Penn), Nikolai Matni (Penn)

SCAN ME

Introduction

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

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

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

Main: establish instance optimal rates for ERM

$$\hat{f} \in \operatorname{argmin}_{f \in \mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} \|f(X_i) - Y_i\|^2$$

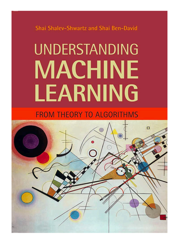
Distinguish between:

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

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

- Can still define $W_{1:n}$ via $f_\star \in \operatorname{argmin}_{f \in \mathscr{F}} \mathbf{E}_{X,Y}\|f(X) - Y\|^2$ and $W_i \triangleq Y_i - f_\star(X_i)$

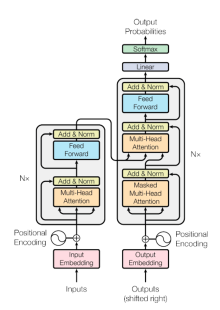
Motivation

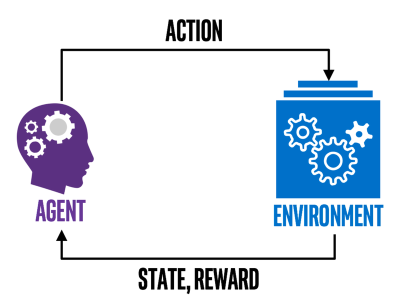


IID learning is very well understood— rich theory:

- Well-known asymptotics, sharp rates for ERM, lots of algorithmic results

Key issue: temporal dependence in data not allowed in IID learning!





Goal: can we build a sharp theory for dependent learning?

Problem: Blocking typically deflates the sample size

Blocking transforms n dep. samples $m = n/t_{\mathrm{mix}}$ independent "blocks"

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

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

Generically employed, deflates rate of converge by a factor t_{mix}...

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

Contribution

Theorem [ZTPM2024]: *Let $\mathscr{F}$ be either a 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:*

- $\|f\|_{\Psi_p} \le L\|f\|_{L^2}^{\eta}$ for some $\eta \in (0,1]$ where $\|f\|_{\Psi_p} \triangleq \sup_{m \ge 1} m^{-1/p}\|f(X)\|_{L^m}$
- n is greater than a polynomial_η *burn-in* in problem constants (t_{mix}, dim, etc.)

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

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

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

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

$(= \mathbf{Var}(W)$ if strictly realizable)

We also show that the weak subG class condition holds

- For finite hypothesis classes
- For smooth (in the input) hypothesis classes
- For parametric hypothesis classes (Loja)

The term $\text{complexity}(\mathscr{F})$:

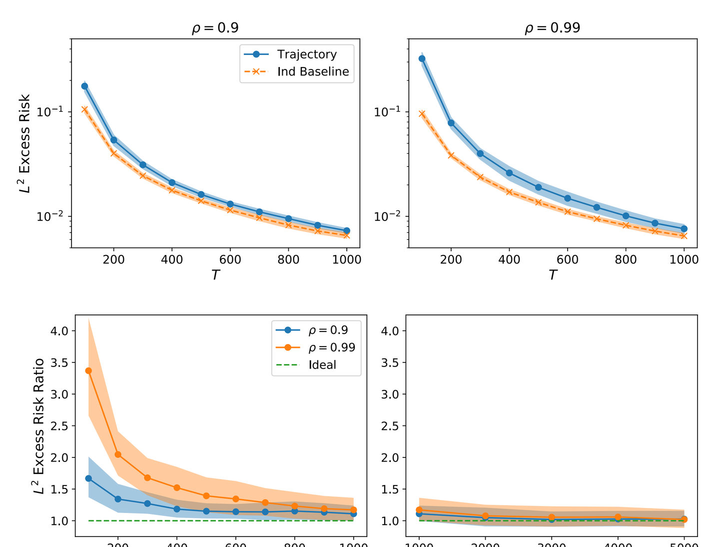
- Is formally a local Talagrand γ_2-functional
- Scales like d_X for d_X-dimensional linear regression
- More generally scales like d for d-dimensional parametric classes (controlled by metric entropy)

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 $T = n$):



Naive blocking would've given us:

$$\mathbf{ER}(\hat{f}) \lesssim t_{\mathrm{mix}}\sigma_W^2 \times \frac{\text{complexity}(\mathscr{F})}{n}$$

Here: improve this term

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

Overview of Results: dep. data & square loss

Paper	Hypothesis Class	Rate (Hiding logs)	Mixing/ Stability	Guarantee	Realizability
[SMTJR2018]	Linear	$n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Marginal	Param recovery	Required
[SO2020]	Generalized Lin	$t_{\mathrm{mix}} n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Strict	Param recovery	Required
[KNJN2021]	Generalized Lin	$n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Marginal	Param recovery	Required
[RBE2021]	**General**	$t_{\mathrm{mix}} n^{-1+\epsilon}\sigma^2 \times \text{complexity}(\mathscr{F})$	Strict	Excess Risk	**Not Required**
[**ZT**2022]	**General** (4-2 hypcon)	$n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Strict	Excess Risk	Required
[SOO2022]	Bilinear	$n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Marginal	Param Recovery	Required
[**ZTPM**2023]	Linear	$n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Strict	Excess Risk	**Not Required**
[**ZTPM**2024]	**General** (weak subG)	$n^{-1}\sigma^2 \times \text{complexity}(\mathscr{F})$	Strict	Excess Risk	**Not Required**

Sharp Rates* without Realizability. (* up to a logarithm)

Proof Strategy

First since $\mathscr{F}$ is convex or realizable $\mathbf{ER}(\hat{f}) \le \mathbf{E}_X\|\hat{f}(X) - f_\star(X)\|^2$

Define the *quadratic process:*

$$Q_{n,\epsilon}(f) \triangleq \mathbf{E}_X\|f(X) - f_\star(X)\|^2 - \frac{1+\epsilon}{n}\sum_{i=1}^{n}\|f(X_i) - f_\star(X_i)\|^2$$

On the event $\{Q_{n,\epsilon}(f) \le 0 : \forall f \in \mathscr{F} \setminus rS_{L^2}\}$ we have:

$$\mathbf{E}_X\|\hat{f}(X) - f_\star(X)\|^2 \le r + \frac{1+\epsilon}{rn}\sum_{i=1}^{n} 2(1-\mathbf{E}')\left\langle W_i, \frac{r[\hat{f}(X_i) - f_\star(X_i)]}{\|\hat{f}(X_i) - f_\star(X_i)\|_{L^2}}\right\rangle$$

Define the *multiplier process:*

$$M_{n,\epsilon}(f) \triangleq \sum_{i=1}^{n} 2(1-\mathbf{E}')\langle W_i, f(X_i) - f_\star(X_i)]\rangle$$

Hence: the proof boils down to unif. controlling $Q_{n,\epsilon}(f)$ and $M_{n,\epsilon}(f)$ over the localized class $(\mathscr{F} - \mathscr{F}) \cap rS_{L^2}$ and:

$$\text{complexity}(\mathscr{F}): \quad \text{soln. to } r \asymp \sup_{f \in (\mathscr{F}-\mathscr{F}) \cap rS_{L^2}} M_{n,\epsilon}(f)$$

Insight from Mendelson (2014), $Q_{n,\epsilon}(f)$ does not affect the rate, but only the burn-in $\Rightarrow$ can freely block to control $Q_{n,\epsilon}(f)$

The crux is to control $M_{n,\epsilon}(f)$ carefully. We combine:

- A version of Bernstein's inequality (Maurer and Pontil 2021)
  - For $V_i(f) \triangleq (1-\mathbf{E}')\langle W_i, f(X_i) - f_\star(X_i)]\rangle$ with $\|f\|_{L^2} = r$:

$$\frac{1}{n}\sum_{i=1}^{n} V_i \lesssim 2\sqrt{\frac{\mathbf{V}(\bar{V})\ln(1/\delta)}{n}} + \frac{t_{\mathrm{mix}}\|\bar{V}\|_{\Psi_p}\log(1/\delta)}{n} \qquad \bar{V} = \frac{1}{t_{\mathrm{mix}}}\sum_{i=1}^{k} V_i$$

- Localization ($\|f\|_{L^2} = r$)
  - Balances both RHS terms $\Rightarrow$ reintroduces the mixing time in the rate
- The weak subG class condition
  - Breaks the balance and makes the variance term dominant again
- Tail bounds via generic chaining (Dirksen 2015) for local unif. control

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** as in the independent case

In other words: dependence only hurts without realizability and the degradation is graceful