Concentration Inequalities for Statistical Learning for Time Dependent Data

Wei Biao Wu

Department of Statistics, University of Chicago

September, 2025, Vienna

Example: Empirical risk minimization

Regularized risk:

$$R(h) = \mathbb{E}\, \underbrace{L(Y, h(X))}_{\text{loss function}} + \underbrace{\rho(h)}_{\text{regularizer}}, \quad h^* = \mathrm{argmin}_{h \in \mathcal{H}} R(h).$$

Given data $(X_i, Y_i)_{i=1}^n$, the regularized empirical risk:

$$R_n(h) = n^{-1} \sum_{i=1}^{n} L(Y_i, h(X_i)) + \rho(h), \quad \hat{h} = \mathrm{argmin}_{h \in \mathcal{H}} R_n(h).$$

- SVM classification: $Y_i = \{0, 1\}$, X_i: feature vectors.
- Support vector regression (SVR): Y_i output variable, X_i input variable.

  L: δ insensitive function, $L_\delta(x) = (|x| - \delta)_+$.

Example: Empirical risk minimization

Regularized risk:

$$R(h) = \mathbb{E}\,\underbrace{L(Y, h(X))}_{\text{loss function}} + \underbrace{\rho(h)}_{\text{regularizer}}, \quad h^* = \operatorname{argmin}_{h \in \mathcal{H}} R(h).$$

Given data $(X_i, Y_i)_{i=1}^n$, the regularized empirical risk:

$$R_n(h) = n^{-1} \sum_{i=1}^{n} L(Y_i, h(X_i)) + \rho(h), \quad \hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} R_n(h).$$

- SVM classification: $Y_i = \{0, 1\}$, X_i: feature vectors.
- Support vector regression (SVR): Y_i output variable, X_i input variable.

  L: δ insensitive function, $L_\delta(x) = (|x| - \delta)_+$.

Example: Empirical risk minimization

Regularized risk:

$$R(h) = \mathbb{E}\, \underbrace{L(Y, h(X))}_{\text{loss function}} + \underbrace{\rho(h)}_{\text{regularizer}}, \quad h^* = \operatorname{argmin}_{h \in \mathcal{H}} R(h).$$

Given data $(X_i, Y_i)_{i=1}^n$, the regularized empirical risk:

$$R_n(h) = n^{-1} \sum_{i=1}^{n} L(Y_i, h(X_i)) + \rho(h), \quad \hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} R_n(h).$$

- LASSO, ridge, support vector regression, ...
- It is well known (cf. Devroye et al. 1996) that

$$0 \le R(\hat{h}) - R(h^*) \le 2\Psi_n, \text{ where } \Psi_n = \sup_{h \in \mathcal{H}} |R_n(h) - R(h)|.$$

- The key issue in statistical learning: a tail probability bound for Ψ_n.

Example: Empirical risk minimization

Regularized risk:

$$R(h) = \mathbb{E}\, \underbrace{L(Y, h(X))}_{\text{loss function}} + \underbrace{\rho(h)}_{\text{regularizer}}, \quad h^* = \mathrm{argmin}_{h \in \mathcal{H}} R(h).$$

Given data $(X_i, Y_i)_{i=1}^n$, the regularized empirical risk:

$$R_n(h) = n^{-1} \sum_{i=1}^{n} L(Y_i, h(X_i)) + \rho(h), \quad \hat{h} = \mathrm{argmin}_{h \in \mathcal{H}} R_n(h).$$

- LASSO, ridge, support vector regression, ...
- It is well known (cf. Devroye et al. 1996) that

$$0 \le R(\hat{h}) - R(h^*) \le 2\Psi_n, \text{ where } \Psi_n = \sup_{h \in \mathcal{H}} |R_n(h) - R(h)|.$$

- The key issue in statistical learning: a tail probability bound for Ψ_n.

Example: Empirical risk minimization

Regularized risk:

$$R(h) = \mathbb{E}\,\underbrace{L(Y, h(X))}_{\text{loss function}} + \underbrace{\rho(h)}_{\text{regularizer}}, \quad h^* = \operatorname{argmin}_{h\in\mathcal{H}} R(h).$$

Given data $(X_i, Y_i)_{i=1}^n$, the regularized empirical risk:

$$R_n(h) = n^{-1}\sum_{i=1}^n L(Y_i, h(X_i)) + \rho(h), \quad \hat{h} = \operatorname{argmin}_{h\in\mathcal{H}} R_n(h).$$

- LASSO, ridge, support vector regression, ...
- It is well known (cf. Devroye et al. 1996) that

$$0 \le R(\hat{h}) - R(h^*) \le 2\Psi_n, \text{ where } \Psi_n = \sup_{h\in\mathcal{H}} |R_n(h) - R(h)|.$$

- The key issue in statistical learning: a tail probability bound for Ψ_n .

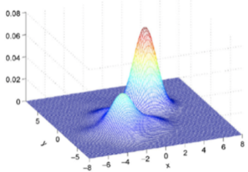
Suprema of the Empirical Processes

For function g, denote $S_n(g) = \sum_{i=1}^{n} g(X_i)$.
We are interested in studying the tail probability

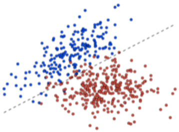
$$T(z) := \mathbb{P}(\Psi_n \geq z), \text{ where } \Psi_n = \sup_{g \in \mathcal{A}} |S_n(g) - \mathbb{E}S_n(g)|.$$

A huge literature when X_i are i.i.d., with various applications.

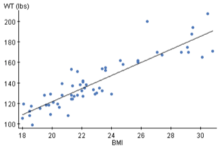
Kernel density estimation



Classification



Regression function estimation



Weak convergence of empirical processes

A huge literature on the weak convergence to Gaussian processes

$$\{n^{-1/2}[S_n(g) - \mathbb{E}S_n(g)], g \in \mathcal{G}\} \Rightarrow \{Z(g), g \in \mathcal{G}\}, \tag{1}$$

where $Z(\cdot)$ is a Gaussian process. For example

- Radulovic, Dragan; Wegkamp, Marten (2018)
- Herold Dehling, Thomas Mikosch, Magda Peligrad, Paul Doukhan,

  ......

Here we primarily focus on the tail probability $T(z) := \mathbb{P}(\Psi_n \geq z)$.

Mystical power of independence

Vapnik-Chervonenkis inequality

$$T(z) \leq 8\mathcal{S}(\mathcal{F}, n)e^{-z^2/(32n)}$$

Mystical power of independence

Vapnik-Chervonenkis inequality

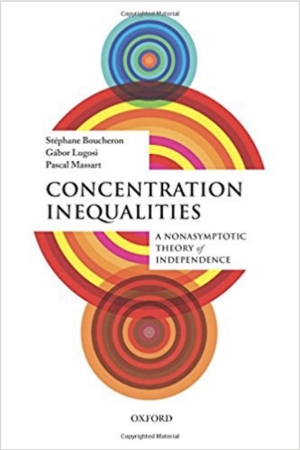
$$T(z) \leq 8\mathcal{S}(\mathcal{F}, n)e^{-z^2/(32n)}$$

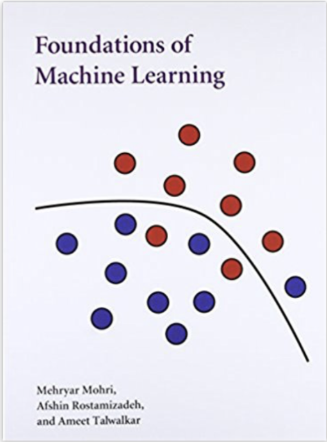
- *Consistency of learning processes*
- *Nonasymptotic theory of the rate of convergence of learning processes*
- *Theory of controlling the generalization ability of learning processes*
- *Theory of constructing learning machines*

– Vladimir N. Vapnik[a]

[a] The Nature of Statistical Learning Theory. 2000.

Mystical power of independence





Mystical power of independence

Example: If (X_i) are i.i.d. random variables and the function class $\mathcal{A} = \{\mathbf{1}_{(-\infty,t]}, t \in \mathbb{R}\}$, the Dvoretzky-Kiefer-Wolfowitz [1] inequality asserts that for all $z \geq 0$,

$$T(z) \leq 2e^{-2z^2/n}.$$

[1] A. Dvoretzky et al. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. 1956.

[2] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. 1990.

Mystical power of independence

Example: If (X_i) are i.i.d. random variables and the function class $\mathcal{A} = \{\mathbf{1}_{(-\infty,t]}, t \in \mathbb{R}\}$, the Dvoretzky-Kiefer-Wolfowitz [1] -Massart[2] inequality asserts that for all $z \geq 0$,

$$T(z) \leq 2e^{-2z^2/n}.$$

[1] A. Dvoretzky et al. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. 1956.

[2] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. 1990.

Time series application

For time series, dependence is the rule rather than the exception!

- Predicting time series with support vector machines. [3] 1402.
- Application of support vector machines in financial time series forecasting [4]. 1637.
- Financial time series forecasting using support vector machines.[5] 2230.
- Time series forecasting using a hybrid ARIMA and neural network model.[6] 5093.
- ...
- Statistical theory being rarely studied! No theoretical guarantee.

[3] K. R. Müller et al. International Conference on Artificial Neural Networks. 1997

[4] FEH. Tay, L. Cao. Omega. 2001

[5] K. Kim. Neurocomputing. 2003.

[6] G. Peter Zhang. Neurocomputing. 2003.

Time series application

For time series, dependence is the rule rather than the exception!

- Predicting time series with support vector machines. [3] 1402.
- Application of support vector machines in financial time series forecasting [4]. 1637.
- Financial time series forecasting using support vector machines.[5] 2230.
- Time series forecasting using a hybrid ARIMA and neural network model.[6] 5093.
- ...
- Statistical theory being rarely studied! No theoretical guarantee.

[3] K. R. Müller et al. International Conference on Artificial Neural Networks. 1997

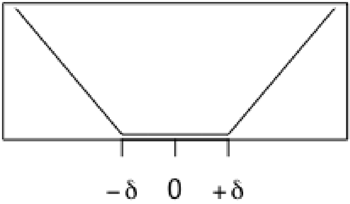
[4] FEH. Tay, L. Cao. Omega. 2001

[5] K. Kim. Neurocomputing. 2003.

[6] G. Peter Zhang. Neurocomputing. 2003.

Dependent vs Independent

- $X_t = \sum_{k \geq 1} a_k \epsilon_{t-k}$, where $a_k = k^{-1.5}$, and $\epsilon_t \sim t_3$.
- $(X'_t)_{t=1}^n$ are i.i.d and $X'_t \sim X_0$.
- Let $L_\delta(x) = (|x| - \delta)_+$, the δ insensitive function



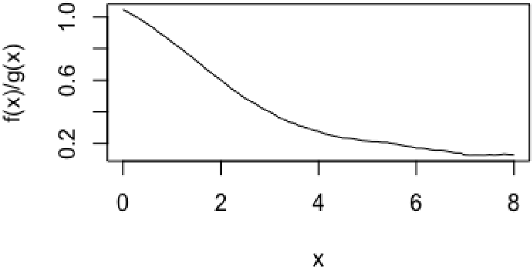
δ insensitive function

Dependent vs Independent

- $X_t = \sum_{k \geq 1} a_k \epsilon_{t-k}$, where $a_k = k^{-1.5}$, and $\epsilon_t \sim t_3$.
- $(X'_t)_{t=1}^n$ are i.i.d and $X'_t \sim X_0$.
- $S_n = \sum_{t=1}^n L_\delta(X_t)$, and $S'_n = \sum_{t=1}^n L_\delta(X'_t)$.
- $g(x) = \mathbb{P}(S_n - \mathbb{E}S_n \geq \sqrt{n}x)$ and $f(x) = \mathbb{P}(S'_n - \mathbb{E}S'_n \geq \sqrt{n}x)$.

Dependent vs Independent

- $X_t = \sum_{k \ge 1} a_k \epsilon_{t-k}$, where $a_k = k^{-1.5}$, and $\epsilon_t \sim t_3$.
- $(X'_t)_{t=1}^n$ are i.i.d and $X'_t \sim X_0$.
- $S_n = \sum_{t=1}^n L_\delta(X_t)$, and $S'_n = \sum_{t=1}^n L_\delta(X'_t)$.
- $g(x) = \mathbb{P}(S_n - \mathbb{E}S_n \ge \sqrt{n}x)$ and $f(x) = \mathbb{P}(S'_n - \mathbb{E}S'_n \ge \sqrt{n}x)$.



Dependence

Our Primary Goal: *to develop sharp or nearly sharp bounds for $T(z)$ under dependence, thus providing theoretical guarantee for statistical learning for time dependent data*.

My talk is based on a series of papers joint with Likai Chen, Yuefeng Han, Danna Zhang and others.

Dependence - Theory

Mixing?

- B. Yu (1994), M. Mohri and A. Rostamizadeh (2010), M. Peligrad (1992), Xiaohong Chen and Xiaotong Shen (1998), P. Doukhan (1994), A. Kontorovich and A. Brockwell (2014), I. Steinwart and A. Christmannetc (2009), etc.
- Let $\mathcal{F}_i^j$ be σ-field generated by X_l, $i \leq l \leq j$.
  - Strong mixing: $\alpha(s) = \sup_{t\in\mathbb{Z}, A\in\mathcal{F}_{t+s}^{\infty}, B\in\mathcal{F}_{-\infty}^{t}} |P(A \cap B) - P(A)P(B)|$.
  - β-mixing: $\beta(s) = \sup_{t\in\mathbb{Z}} \mathbb{E}_{B\in\mathcal{F}_{-\infty}^{t}} \left(\sup_{A\in\mathcal{F}_{t+s}^{\infty}} |P(A|B) - P(A)| \right)$.
  - ϕ-mixing : $\phi(s) = \sup_{t\in\mathbb{Z}, A\in\mathcal{F}_{t+s}^{\infty}, B\in\mathcal{F}_{-\infty}^{t}} |P(A|B) - P(A)|$.

Dependence - Theory

Mixing?

- **Limited**. Some simple and widely used AR processes are not strong mixing. [7]

$$X_t = \theta X_{t-1} + \epsilon_t,$$

where $0 < \theta < 1$ and ϵ_t i.i.d Rademacher random variable.

- **Not handy** Generally hard to verify. Difficulty in dealing with high dimensional data set.
- **Less sharp**. Existing results using mixing can be far from being sharp.

[7] Donald W. K. Andrews. Nonstrong mixing autoregressive processes. 1984,
Boris Solomyak (1995) On the Random Series $\sum \pm \lambda^n$ (an Erdos Problem) *Annals of Mathematics* pp. 611-625

Comparison example

$$X_i = \sum_{k \geq 0} a_k \epsilon_{i-k},$$

where $\epsilon_t \in \mathcal{L}^q$ i.i.d. with $q > 2$ and $a_0 = 1$, $a_k \asymp k^{-\alpha}$, $\alpha > 2 + 1/q$.

Certain conditions on H, bounded loss function. For $\delta > n^{-K}$ with $K = 1/4 - (q+1)/(2(\alpha-1)q)$, $z_\delta = n^{1-K}(\log(\delta - n^{-K})^{-1})^{1/2}$,

$$\text{Mohri \& Rostamizadeh (JMLR, 2010): } \mathbb{P}(n|R_n(\hat{h}) - R(\hat{h})| \geq Cz_\delta) \leq \delta, \tag{2}$$

$$\text{Chen \& Wu: linear process (JMLR, 2018) } \mathbb{P}(n|R_n(\hat{h}) - R(\hat{h})| \geq Cz_\delta) \lesssim nz_\delta^{-q\alpha}. \tag{3}$$

- $nz_\delta^{-q\alpha} \ll \delta$.
- Example: let $\alpha = 4$, $q = 4$. Then (2): $O(n^{-1/24})$, (3): $O(n^{-43/3})$.

Comparison example

$$X_i = \sum_{k \geq 0} a_k \epsilon_{i-k},$$

where $\epsilon_t \in \mathcal{L}^q$ i.i.d. with $q > 2$ and $a_0 = 1$, $a_k \asymp k^{-\alpha}$, $\alpha > 2 + 1/q$.

Certain conditions on H, bounded loss function. For $\delta > n^{-K}$ with $K = 1/4 - (q+1)/(2(\alpha - 1)q)$, $z_\delta = n^{1-K}(\log(\delta - n^{-K})^{-1})^{1/2}$,

$$\text{Mohri \& Rostamizadeh (JMLR, 2010): } \mathbb{P}(n|R_n(\hat{h}) - R(\hat{h})| \geq Cz_\delta) \leq \delta, \tag{2}$$

$$\text{Chen \& Wu: linear process (JMLR, 2018) } \mathbb{P}(n|R_n(\hat{h}) - R(\hat{h})| \geq Cz_\delta) \lesssim nz_\delta^{-q\alpha}. \tag{3}$$

- $nz_\delta^{-q\alpha} \ll \delta$.
- Example: let $\alpha = 4$, $q = 4$. Then (2): $O(n^{-1/24})$, (3): $O(n^{-43/3})$.

Dependence - Theory

- A totally different approach: based on **martingale decomposition** and a recent **high-dimensional version Fuk-Nagaev type inequality**.[8]
- Martingale Methods and Inequalities: Lai, Woodroofe, Chow, Freedman's inequality (1975); moment inequality for martingale: Burkholder-Davis-Gundy Inequality; martingale inequality on Banach space: Einmahl and Li (2008), Pinelis (1994).
- Functional dependence measure (Wu, 2005).

[8] V. Chernozhukov et al. Testing many moment inequalities. 2017.

Dependence - Theory

- A totally different approach: based on **martingale decomposition** and a recent **high-dimensional version Fuk-Nagaev type inequality**.[8]
- Martingale Methods and Inequalities: Lai, Woodroofe, Chow, Freedman's inequality (1975); moment inequality for martingale: Burkholder-Davis-Gundy Inequality; martingale inequality on Banach space: Einmahl and Li (2008), Pinelis (1994).
- Functional dependence measure (Wu, 2005).

[8] V. Chernozhukov et al. Testing many moment inequalities. 2017.

Dependence - Theory

- A totally different approach: based on **martingale decomposition** and a recent **high-dimensional version Fuk-Nagaev type inequality**.[8]
- Martingale Methods and Inequalities: Lai, Woodroofe, Chow, Freedman's inequality (1975); moment inequality for martingale: Burkholder-Davis-Gundy Inequality; martingale inequality on Banach space: Einmahl and Li (2008), Pinelis (1994).
- Functional dependence measure (Wu, 2005).

[8] V. Chernozhukov et al. Testing many moment inequalities. 2017.

Time series (X_i) – Examples

- Autoregressive moving average (ARMA)

$$(1 - \sum_{j=1}^{p} \theta_j B^j) X_i = X_i - \sum_{j=1}^{p} \theta_j X_{i-j} = \sum_{k=1}^{q} \phi_k \epsilon_{i-k},$$

where θ_j and ϕ_k are real coefficients such that the root to the equation $1 - \sum_{j=1}^{p} \theta_j u^j = 0$ are all outside the unit disk.

- Fractional autoregressive integrated moving average (FARIMA) . [9]

$$(1 - B)^d (X_i - \sum_{j=1}^{p} \theta_j X_{i-j}) = \sum_{k=1}^{q} \phi_k \epsilon_{i-k},$$

where the index $d \in (0, 1/2)$.

[9] C. Granger and R. Joyeux. An introduction to long-memory time series models and fractional differencing. 1980.

Dependence settings

Moving average (MA) process.

$$X_i = \sum_{k \geq 0} a_k \epsilon_{i-k},$$

where

- ϵ_i i.i.d. with mean 0 and $\mu_q := \|\epsilon_0\|_q < \infty, q \geq 1$.
- $a_k = O(k^{-\beta})$, $\beta > 1/q$.

Properties:

- q: heaviness of the tail; β: dependence strength.
- If $1/2 < \beta < 1$, $q \geq 2$, long-range dependence (LRD); if $\beta > 1$, short-range dependence (SRD).

Dependence settings

- Example: Autoregressive conditional heteroskedasticity (ARCH)

$$X_t = \sigma_t \epsilon_t,\ \sigma_t^2 = a_0 + \sum_{k=1}^{q} a_k X_{t-k}^2, \quad a_k \geq 0$$

- Example: $X_t = \sum_{k\geq 0} a_k \epsilon_{t-k}$. Then $\delta_{t,q} = \|a_t(\epsilon_0 - \epsilon_0')\|_q$.
- Short-range dependent nonlinear process with weaker dependence
- Short-range dependent nonlinear process with stronger dependence
- Short-range dependent linear process
- long-range dependent linear process

Dependence settings

- Non-linear time series:

$$X_t = F(\epsilon_t, \epsilon_{t-1}, \ldots).^{10} \tag{4}$$

- Let (ϵ'_t) be an independent copy of (ϵ_t) and $X_{t,\{0\}} = F(\epsilon_t, \ldots, \epsilon_1, \epsilon'_0, \epsilon_{-1}, \ldots)$. Functional dependence measure [11]

$$\delta_{t,q} = \|X_t - X_{t,\{0\}}\|_q. \tag{5}$$

- Example: $X_t = \sum_{k \geq 0} a_k \epsilon_{t-k}$. Then $\delta_{t,q} = \|a_t(\epsilon_0 - \epsilon'_0)\|_q$.
- Short-range dependent nonlinear process with weaker dependence
- Short-range dependent nonlinear process with stronger dependence
- Short-range dependent linear process
- long-range dependent linear process

[10] M. Priestley. Non-linear and non-stationary time series analysis. 1988.

[11] W. Wu. Nonlinear system theory: Another look at dependence. 2005.

Dependence settings

- Non-linear time series:

$$X_t = F(\epsilon_t, \epsilon_{t-1}, \ldots).^{10} \tag{4}$$

- Let (ϵ'_t) be an independent copy of (ϵ_t) and $X_{t,\{0\}} = F(\epsilon_t, \ldots, \epsilon_1, \epsilon'_0, \epsilon_{-1}, \ldots)$. Functional dependence measure [11]

$$\delta_{t,q} = \|X_t - X_{t,\{0\}}\|_q. \tag{5}$$

- Example: $X_t = \sum_{k\geq 0} a_k \epsilon_{t-k}$. Then $\delta_{t,q} = \|a_t(\epsilon_0 - \epsilon'_0)\|_q$.
- Short-range dependent nonlinear process with weaker dependence
- Short-range dependent nonlinear process with stronger dependence
- Short-range dependent linear process
- long-range dependent linear process

[10] M. Priestley. Non-linear and non-stationary time series analysis. 1988.

[11] W. Wu. Nonlinear system theory: Another look at dependence. 2005.

Definitions and Assumptions

- For a function class $\mathcal{A}$ of bounded functions, define

$$\mathcal{N}_{\mathcal{A}}(\delta) := \min\Big\{m : g_1, \ldots, g_m \in \mathcal{A}, \text{ s.t. } \sup_{g\in\mathcal{A}} \min_{1\le j\le m} |g - g_j|_\infty \le \delta\Big\},$$

where $|g|_\infty = \sup_x |g(x)|$. Let $H_{\mathcal{A}}(\delta) := \log(\mathcal{N}_{\mathcal{A}}(\delta))$.

(A) (Smoothness) For any $g \in \mathcal{A}$, $|g|, |g'|, |g''|$ are uniformly bounded, W.L.O.G. set the bound to be 1.

(A') Assume $\sup_{g\in\mathcal{A}} |g|_\infty < \infty$, $f_\epsilon', f_\epsilon''$ exist and $\int_{-\infty}^{\infty} |f_\epsilon'(x)| \mathrm{d}x, \int_{-\infty}^{\infty} |f_\epsilon''(x)| \mathrm{d}x$ are finite.

(B) (Algebraically Decaying Coefficients) For some $\gamma, \beta > 0$, $|a_k| \le \gamma k^{-\beta}$ holds for all $k \ge 1$.

(B') (Exponentially Decaying Coefficients) For some $\gamma > 0, 0 < \rho < 1$, $|a_k| \le \gamma\rho^k$ holds for all $k \ge 0$.

Functional classes

(D) (Exponential Class) For some constants $N, C, \theta > 0$, the covering number $\mathcal{N}_{\mathcal{A}}(\delta) \leq N \exp(C\delta^{-\theta})$ holds for all $\delta < 1$.(Hölder/Sobolev classes)

(D') (Algebraic Class) For some constants $N, \theta > 0$, the covering number $\mathcal{N}_{\mathcal{A}}(\delta) \leq N\delta^{-\theta}$ holds for all $\delta < 1$. (VC classes, sparse neural networks)

Common settings.

(cf. Kosorok (2006)[12], van der Vaart and Wellner (1996)[13].)

[12] Introduction to Empirical Processes and Semiparametric Inference.

[13] Weak Convergence and Empirical Processes.

Examples

Exponential class:

$\mathcal{F}$: bounded convex functions on a compact convex set C, Lipschitz continuous with coefficient L.
$\mathcal{G}$: $\{g : [0,1] \to [0,1]$ with $\|g^{(m)}\|_{\mathcal{L}_2} \leq 1.\}$

$$\mathcal{N}_{\mathcal{F}}(\delta) \leq \exp\{c_1(1+L)^{d/2}\delta^{-d/2}\}, \quad \mathcal{N}_{\mathcal{G}}(\delta) \leq \exp\{c_2\delta^{-1/m}\}.$$

Polynomial class:

$\mathcal{F}$: $f = \sum_{k=1}^{m} \theta_k \phi_k(\cdot)$, $\theta \in \Theta$ a compact convex subset of $\mathbb{R}^m$, and $(\phi_k)_{k=1}^{m}$ are real-valued basis functions.

$$\mathcal{N}_{\mathcal{F}}(\delta) \leq c_3\delta^{-m}.$$

Sparse Neural Networks (to be discussed later).

Short-range dependent nonlinear processes

Given $X_t = F(\epsilon_t, \epsilon_{t-1}, \dots)$ and the function dependence measure $\delta_{t,q} = \|X_t - X_{t,\{0\}}\|_q$, define the dependence adjusted norm (d.a.n.)

$$\|X_\cdot\|_{q,\alpha} = \sup_{i \geq 0}(i+1)^\alpha \sum_{j=i}^{\infty} \delta_{j,q}, \ \alpha \geq 0, \tag{6}$$

which plays a key role for asymptotics under dependence.

- Assume $\mathbb{E}X_t = 0$. Let $q \geq 1$. The d.a.n. $\|X_\cdot\|_{q,\alpha} \geq \|X_t\|_q$
- The d.a.n. $\|X_\cdot\|_{q,\alpha}$ is non-decreasing in α, q.
- If $\sum_{j=m}^{\infty} \delta_{j,q} \asymp m^{-\beta}$, $\beta > 0$, then the d.a.n. $\|X_\cdot\|_{q,\alpha} = \infty$ for all $\alpha > \beta$, and $\|X_\cdot\|_{q,\beta} < \infty$

More properties of d.a.n.

- It can happen $\|X_{\cdot}\|_{q,0} = \infty$, which leads to long-range dependence
- weak dependence, short-range dependence or short-memory: $\|X_{\cdot}\|_{q,0} = \sum_{j=0}^{\infty} \delta_{j,q} < \infty$
- larger α with $\|X_{\cdot}\|_{q,\alpha} < \infty$ means weaker dependence
- The long run variance $\sigma_{\infty}^2 = \sum_{t=-\infty}^{\infty} cov(X_0, X_t) \leq \|X_{\cdot}\|_{2,0}^2$

Short-range dependent nonlinear processes

For function g, denote $S_n(g) = \sum_{i=1}^n g(X_i)$. Recall the tail probability

$$T(z) := \mathbb{P}(\Psi_n \geq z), \text{ where } \Psi_n = \sup_{g \in \mathcal{A}} |S_n(g) - \mathbb{E}S_n(g)|.$$

Theorem

(weak dependence case with weaker dependence) Assume that all $g \in \mathcal{A}$ satisfies $|g'|_\infty \leq 1$. (i) If $\alpha > 1/2 - 1/q$, then

$$\mathbb{P}(\Psi_n \geq x) \lesssim \frac{n\ell^{q/2}}{x^q} \|X_\cdot\|_{q,\alpha}^q + \exp(-c_{q,\alpha} \frac{x^2}{n\|X_\cdot\|_{2,\alpha}^2}) \tag{7}$$

for all $x \geq \sqrt{n\ell}\|X_\cdot\|_{2,\alpha} + n^{1/q}\ell^{3/2}\|X_\cdot\|_{q,\alpha}$, where $\ell = \log(\mathcal{N}_{\mathcal{A}}(x/n))$.

Short-range dependent nonlinear processes

Theorem

(Continued, weak dependence case with stronger dependence) (ii) If $0 < \alpha < 1/2 - 1/q$, *then*

$$\mathbb{P}(\Psi_n \geq x) \lesssim \frac{n^{q/2-\alpha q}\ell^{q/2}}{x^q}\|X_\cdot\|_{q,\alpha}^q + \exp(-c_{q,\alpha}x^2/(n\|X_\cdot\|_{2,\alpha}^2)) \quad (8)$$

for all $x \geq \sqrt{n\ell}\|X_\cdot\|_{2,\alpha} + n^{1/2-\alpha}\ell^{3/2}\|X_\cdot\|_{q,\alpha}$, *where* $\ell = \log(\mathcal{N}_{\mathcal{A}}(x/n))$.

Short-range dependent linear processes

Let $q' := q \wedge 2$, $c(n,q) = n^{1/q'}$ if $q \neq 2$ and $n^{1/2}\log^{1/2}(n)$ if $q = 2$.

Theorem 1

Assume $(A)(or(A'))$ and (B), $\beta, q > 1$ and $q\beta \geq 2$. Then we have

$$\mathbb{P}\Big(\Psi_n \geq C_1 c(n,q) + z\Big)$$

$$\leq \underbrace{C_2 \frac{n}{z^{q\beta}}}_{\text{Polynomial term}} + \underbrace{3\exp\Big\{-\frac{z^2}{C_3 n} + H_{\mathcal{A}}\Big(\frac{z}{4n}\Big)\Big\} + 2\exp\Big\{-\frac{z^v}{C_4} + H_{\mathcal{A}}\Big(\frac{z}{4n}\Big)\Big\}}_{\text{Exponential term}},$$

where $v = v_{q,\beta} = (q'\beta - 1)(3q'\beta - 1)^{-1}$.

Short-range dependent linear processes

Let $q' := q \wedge 2$, $c(n,q) = n^{1/q'}$ if $q \neq 2$ and $n^{1/2}\log^{1/2}(n)$ if $q = 2$.

Theorem 1

Assume $(A)(or(A'))$ and (B), $\beta, q > 1$ and $q\beta \geq 2$. Then we have

$$\mathbb{P}\Big(\Psi_n \geq C_1 c(n,q) + z\Big)$$

$$\leq \underbrace{C_2 \frac{n}{z^{q\beta}}}_{\text{Polynomial term}} + \underbrace{3\exp\Big\{-\frac{z^2}{C_3 n}\Big\} + 2\exp\Big\{-\frac{z^v}{C_4}\Big\}}_{\text{Exponential term}},$$

where $v = v_{q,\beta} = (q'\beta - 1)(3q'\beta - 1)^{-1}$.

Short-range dependent linear processes

Corollary 1

Assume (A) (or (A')) and (B). Let $\beta > 1$ and $q > 2$. If either:

(i) assumption (D), and $z \geq cn^{1/2+\alpha}$;

(ii) assumption (D') and $z \geq cn^{1/2}\log^{1/2}(n)$, then

$$\mathbb{P}\Big(\Psi_n \geq z\Big) \leq C\frac{n}{z^{q\beta}},$$

where $\alpha = \max\{\theta/(\theta+2), (\theta-v)/(\theta+v)\}/2$ and C is a constant that does not rely on n and z.

Recall the Dvoretzky-Kiefer-Wolfowitz inequality,

$$T(z) \leq 2e^{-2z^2/n}.$$

Short-range dependent linear processes

Coefficients decay **exponentially**: $|a_k| \le \gamma\rho^k$.

Theorem 2

Let $\mathcal{A} = \{g : \mathbb{R} \mapsto \mathbb{R}, |g|_\infty \le 1, |g'|_\infty \le 1\}$. Assume that the coefficients of (X_i) satisfy (B'). Then for $q' = \min\{q, 2\}$,

$$\mathbb{P}(\Psi_n \ge C_1\sqrt{n}/(1-\rho) + z) \le C_2\Big(\frac{e^{-C_3(1-\rho)n}}{z^q(1-\rho)^{q+q/q'}} + e^{-C_4 z^2(1-\rho)^2/n}\Big).$$

- ARMA: $(1 - \sum_{j=1}^p \theta_j B^j)X_i = \sum_{k=1}^q \phi_k \epsilon_{i-k}$, some constants $p, q \in \mathbb{N}$. $\rho = \max\{|u| : 1 - \sum_{j=1}^p \theta_j u^{-j} = 0\}$

Short-range dependent linear processes

Coefficients decay **exponentially**: $|a_k| \le \gamma \rho^k$.

Theorem 2

Let $\mathcal{A} = \{g : \mathbb{R} \mapsto \mathbb{R}, |g|_\infty \le 1, |g'|_\infty \le 1\}$. Assume that the coefficients of (X_i) satisfy (B'). Then for $q' = \min\{q, 2\}$,

$$\mathbb{P}(\Psi_n \ge C_1\sqrt{n}/(1-\rho) + z) \le C_2\left(\frac{e^{-C_3(1-\rho)n}}{z^q(1-\rho)^{q+q/q'}} + e^{-C_4 z^2 (1-\rho)^2/n}\right).$$

- ARMA: $(1 - \sum_{j=1}^p \theta_j B^j)X_i = \sum_{k=1}^q \phi_k \epsilon_{i-k}$, some constants $p, q \in \mathbb{N}$. $\rho = \max\{|u| : 1 - \sum_{j=1}^p \theta_j u^{-j} = 0\}$

Long-range dependent linear processes

Corollary 2

Assume (A) (or (A')) and (B). Let $q > 2$, $1/2 < \beta < 1$. If either
(i) condition (D) with $0 < \alpha \le \beta - 1/2$, $\theta < 2\alpha/(\beta - 1/2 - \alpha)$ and $z \ge cn^{3/2-\beta+\alpha}$
(ii) condition (D') with $\alpha > 1/2$, $z \ge cn^{3/2-\beta}\log^{\alpha}(n)$. Then

$$\mathbb{P}\Big(\Psi_n \ge z\Big) \le C\frac{n^{1+(1-\beta)q}}{z^q},$$

where C is some constant that does not rely on n.

Long-range dependent linear processes

Theorem 3

Assume $(A)(or(A'))$ and (B), $q > 2$, $1/2 < \beta < 1$. Then for all $z > 0$,

$$\mathbb{P}\Big(\Psi_n \geq C_1 n^{3/2-\beta} + z\Big) \leq \underbrace{C_2 \frac{n^{1+(1-\beta)q}}{z^q}\Big(1 + \frac{[H_{\mathcal{A}}(z/4n) + \log(n)]^q}{\tilde{c}^q(n,\beta)}\Big)}_{\text{Polynomial term}}$$

$$+ \underbrace{3\exp\Big(-\frac{z^2}{C_3 n^{3-2\beta}} + H_{\mathcal{A}}(\frac{z}{4n})\Big)}_{\text{Exponential term}},$$

where

$$\tilde{c}(n,\beta) = \begin{cases} n^{1/4-|3/4-\beta|} & \text{if } \beta \neq 3/4, \\ n^{1/4}/\log(n) & \text{if } \beta = 3/4. \end{cases}$$

Sub-exponential innovations

Sub-exponential: $\mathbb{E}(e^{c_0|\epsilon_0|}) < \infty$.

Theorem 4

Let $\mathcal{G} = \{g : |g|_\infty \le 1, |g'|_\infty \le 1\}$. Assume (B) and $|f_\epsilon|_\infty \le f_*$, $f_* > 0$.

(a) for SRD case ($\beta > 1$), we have for all $z > 0$,

$$\mathbb{P}(\Psi_n \ge C_1\sqrt{n} + z) \le 2e^{-C_2 z^2/n},$$

(b) for LRD case ($1/2 < \beta < 1$), we have for all $z > 0$,

$$\mathbb{P}(\Psi_n \ge C_3 n^{3/2-\beta} + z) \le 2e^{-C_4 z^2/n^{3-2\beta}}.$$

Tail behavior

Innovation\Coefficient	Poly	Exp
Finite moment	Exp+Poly	Exp
Sub-exp	Exp	Exp

Kernel density estimation

$X_i \sim \mathrm{MA}(\infty)$ with a marginal density f. Kernel density estimator of f:

$$\hat{f}_n(x) = \frac{1}{n}\sum_{j=1}^{n} K_b(x - X_j), \quad K_b(\cdot) = b^{-1}K(\cdot/b),$$

where the bandwidth $b = b_n$ with $b_n \to 0$ and $nb_n \to \infty$.

$$\mathbb{P}\Big(\sup_{x\in\mathbb{R}} n|\hat{f}_n(x) - \mathbb{E}\hat{f}_n(x)| \geq z\Big).$$

- non-asymptotic confidence bounds (Giné and Nickl (2010) AOS)
- clustering problem (Rinaldo et al. (2012) JMLR)
- forest density estimation (Liu et al. (2011) JMLR)

Kernel density estimation

Assume (B), the kernel K is symmetric with support $[-1,1]$ and $|K|, |K'|$, $|f_\epsilon|, |f_\epsilon'|, |f_\epsilon''|$ are all bounded by some finite constant L.

(a) In the SRD case with $\beta > 1, q > 1, q\beta \geq 2$, if $nb_n \geq \log(n)$ and $z \geq c(n/b_n)^{1/2}\log^{1/2}(n)$ for a sufficiently large c, then

$$\mathbb{P}\Big(\sup_{x\in\mathbb{R}} n|\hat{f}_n(x) - \mathbb{E}\hat{f}_n(x)| \geq z\Big) \leq C\mu_q^q \frac{n}{z^{q\beta}},$$

(b) In the LRD case with $1/2 < \beta < 1, q > 2$, if $z \geq c\max\{n^{3/2-\beta}, (n/b_n)^{1/2}\}\log^{\alpha}(n)$ holds for some $\alpha > 1/2$, $c > 0$, then

$$\mathbb{P}\Big(\sup_{x\in\mathbb{R}} n|\hat{f}_n(x) - \mathbb{E}\hat{f}_n(x)| \geq z\Big) \leq C\mu_q^q \frac{n^{1+(1-\beta)q}}{z^q},$$

where C is a constant that does not rely on n, z.

Empirical risk minimization

Consider (X_i) satisfy the MA(∞) process, $(\eta_i)_{i\in\mathbb{Z}}$ are i.i.d. random errors independent of (ϵ_i) in (X_i), and

$$Y_i = H_0(X_i, \eta_i),$$

where H_0 is an unknown measurable function. Assume the loss function $0 \le L \le 1$. Denote $\mathcal{A} = \{L(x, y, h(x)) : h \in \mathcal{H}\}$.

Assume (B), the density $f_\epsilon \in \mathcal{C}^2(\mathbb{R})$, $\int_{-\infty}^{\infty} |f_\epsilon'(x)| + |f_\epsilon''(x)| \mathrm{d}x < \infty$. Under SRD conditions in Corollary 1, we have

$$\mathbb{P}\Big(\Psi_n \ge C_q a_* \mu_q c(n, q) + z\Big) \le C \frac{n}{z^{q\beta}}.$$

Under LRD conditions in Corollary 2, we have

$$\mathbb{P}\Big(\Psi_n \ge z\Big) \le C \frac{n^{1+(1-\beta)q}}{z^q}.$$

Gaussian approximation

- Define the long-run covariance function

$$\sigma_{gh} := \sum_{k\in\mathbb{Z}} \mathrm{Cov}[g(X_i), h(X_{i+k})] \text{ and } \sigma_g := \sigma_{gg}.$$

Let $(Z_g)_{g\in\mathcal{A}}$ be a gaussian field such that for any finite subset $g_1, ..., g_v \in \mathcal{A}$, $v \geq 1$, the gaussian vector $(Z_{g_1}, \ldots, Z_{g_v})^T$ has mean zero and covariance matrix $(\sigma_{g_i g_j})_{i,j=1}^v$. Recall $S_n(g) = \sum_{i=1}^n g(X_i)$.

Theorem 4

Assume (A), (B) with $q > 4$, $\beta > 1$. Let $\mathcal{A}$ be a class of functions g with $\mathbb{E}g(X_i) = 0$ and $\mathcal{A}_0 = \{\sigma_g^{-1/2} g | g \in \mathcal{A}\}$. Assume $\mathcal{N}_{\mathcal{A}_0}(\delta) \leq L\delta^{-\theta}$, where $L, \theta > 0$ and there exists a constant $c > 0$ such that $\inf_{g\in\mathcal{A}} \sigma_g \geq c$. Then we have GA

$$\sup_{u\geq 0}\left|\mathbb{P}\left(\sup_{g\in\mathcal{A}} \left|(n\sigma_g)^{-1/2} S_n(g)\right| \geq u\right) - \mathbb{P}\left(\sup_{g\in\mathcal{A}} |\sigma_g^{-1/2} Z_g| \geq u\right)\right| \to 0.$$

Sharpness

Empirical distribution functions:

$$S_n(t) = n[\hat{F}_n(t) - F(t)] = \sum_{i=1}^{n}[\mathbf{1}_{X_i \le t} - F(t)].$$

(A_1) For $F_\epsilon(u) = \mathbb{P}(\epsilon_0 \le u)$, the cumulative distribution function of ϵ_0, assume that $f_\epsilon = F'_\epsilon$ and F''_ϵ are both bounded, W.L.O.G. set the bound to be 1.

Sharpness

Corollary 3

Assume (A_1) and (B).

(i) (SRD) If $\beta, q > 1$, $q\beta \geq 2$, and if for some $\alpha > 1/2$, $c > 0$, we have $z \geq cn^{1/2}\log^{\alpha}(n)$, then

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}}|S_n(t)| > C_1 c(n,q) + z\right) \leq C_2 \frac{n}{z^{q\beta}},$$

(ii) (LRD) if $1/2 < \beta < 1$ and $q > 2$, and if for some $\alpha > 1/2$, $c > 0$, we have $z \geq cn^{3/2-\beta}\log^{\alpha}(n)$, then

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}}|S_n(t)| > C_1 n^{3/2-\beta} + z\right) \leq C_2 \frac{n^{1+(1-\beta)q}}{z^{q}}.$$

Precise rate

Under certain forms of tail probability of the innovations, we can have a more refined result.

Corollary 4

Assume $(A_1)(B)$ and $\beta, q > 1$, $q\beta \geq 2$. Assume for any $x > 1$,

$$\mathbb{P}(|\epsilon_0| > x) \leq C\log^{-r_0}(x)x^{-q},$$

some $r_0 > 1, C > 0$. If $z \geq \sqrt{n}\log^{\alpha}(n), \alpha > 1/2$, then

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}}|S_n(t)| > z\right) \lesssim \frac{n}{z^{q\beta}\log^{r_0}(z)},$$

where constant in $\lesssim$ only depends on β, q, γ, r_0, C.

Precise rate

What is the best decay rate we can possibly expected?

Theorem 6

Assume $(A_1)(B)$ with coefficients $a_k = (k \vee 1)^{-\beta},\ k \geq 0$, and ϵ_0 is symmetric with the tail distribution

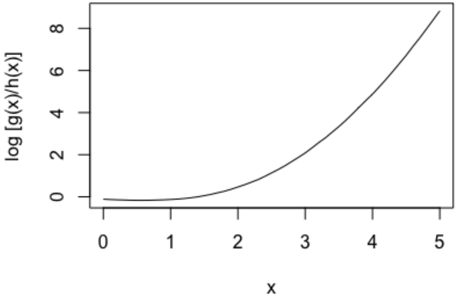
$$\mathbb{P}(|\epsilon_0| \geq x) \sim \log^{-r_0}(x) x^{-q}, \text{ as } x \to \infty,$$

some $r_0 > 1$. Let $q\beta > 2$ and $\beta > 1$. If for some $\alpha > 1/2$, there exists a constant $\Gamma > 0$ such that for all z with $\sqrt{n}\log^{\alpha}(n) \leq z \leq n/\log^{\Gamma}(n)$,

$$\mathbb{P}\left(S_n(t) > z\right) = (1 + o(1)) C_{t,\beta,F} \frac{n}{\log^{r_0}(z) z^{q\beta}}, \quad n \to \infty.$$

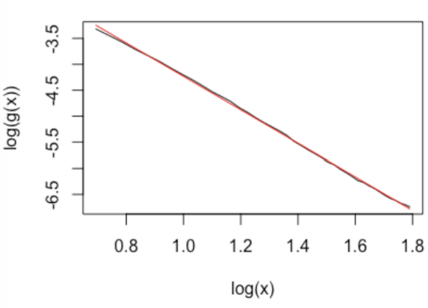
Tail behavior

- $X_t = \sum_{k\geq 1} a_k \epsilon_{t-k}$, where $a_k = k^{-1.5}$, and $\epsilon_t \sim t_3$ i.i.d.
- $S_n = \sum_{t=1}^{n}(L_\delta(X_t) - \mathbb{E}L_\delta(X_t))$ where $L_\delta(x) = (|x| - \delta)_+$.
- $g(x) = \mathbb{P}(S_n/(\sqrt{n}\sigma) \geq x)$ and $h(x) = 1 - \Phi(x)$, $\Phi(\cdot)$ is distribution for standard normal.



Tail behavior

- $X_t = \sum_{k \geq 1} a_k \epsilon_{t-k}$, where $a_k = k^{-1.5}$, and $\epsilon_t \sim t_3$ i.i.d.
- $S_n = \sum_{t=1}^{n} (L_\delta(X_t) - \mathbb{E} L_\delta(X_t))$ where $L_\delta(x) = (|x| - \delta)_+$.
- $g(x) = \mathbb{P}(S_n / (\sqrt{n}\sigma) \geq x)$ and $h(x) = 1 - \Phi(x)$, $\Phi(\cdot)$ is standard normal c.d.f.



Nonlinear Auto-regressive Processes

Consider the nonlinear auto-regressive process

$$X_t = H(X_{t-1}, \ldots, X_{t-\ell}) + \epsilon_t, \tag{9}$$

where ϵ_t are i.i.d., and $H(\cdot)$ satisfies the Lipschitz condition

$$|H(u_1, \ldots, u_\ell) - H(u'_1, \ldots, u'_\ell)| \le \sum_{i=1}^{\ell} h_i |u_i - u'_i|, \tag{10}$$

where Lipschitz constants $h_1, \ldots, h_\ell \ge 0$ are real coefficients. Assume

- $\mu_p := \mathbb{E}(|\epsilon_t|^p) < \infty$ for some $p > 0$
- $\sum_{i=1}^{\ell} h_i < 1$.

Then (9) is geometric moment contracting (GMC) with a stationary solution with $\mathbb{E}(|X_t|^p) < \infty$.

Nonlinear Auto-regressive Processes

Let $1 > \rho \geq \sum_{i=1}^{\ell} h_i$ be the root to the equation $\sum_{i=1}^{\ell} h_i \rho^{-i} = 1$.

Theorem

Assume that function g satisfies $|g|_\infty \leq m$ and is Lipschitz continuous $|g(x) - g(x')| \leq L|x - x'|$ for all x, x'. Then there exists a constant K, only depending on p and μ_p, such that

$$\mathbb{P}(|S_n(g) - n\mathbb{E}g(X_1)| \geq z) \leq 2\exp(-\frac{z^2 A^2 K}{nm^2}), A = \frac{\log \min(e, \rho^{-1})}{\log \max(e, L/m)}$$

- Sharp Azuma-Hoeffding inequality for nonlinear AR processes
- Convenient to use, with explicit dependence on m, L, ρ

Nonlinear Auto-regressive Processes

We have the following Bernstein inequality.

Theorem

Assume that function g is Lipschitz $|g(x) - g(x')| \leq L|x - x'|$ for all x, x', and ϵ_t is sub-exponential: $K := \mathbb{E}\exp(c_0|\epsilon_t|) < \infty$ for some $c_0 > 0$. Then there exists constants c_1, c_2, only depending on c_0 and K, such that

$$\mathbb{P}(|S_n(g) - n\mathbb{E}g(X_1)| \geq z) \leq 2\exp(-\frac{z^2(1-\rho)^2}{c_1 L^2 n + c_2 L z(1-\rho)}).$$

- Sharp Bernstein inequality for nonlinear AR processes
- Convenient to use, with explicit dependence on L, ρ

Non-parametric Estimation of Nonlinear AR Processes

We want to estimate H based on data $(X_t)_{t=1}^n$ from the nonlinear AR

$$X_t = H(X_{t-1}, \ldots, X_{t-\ell}) + \epsilon_t.$$

Let d be a lag and $Y_t = (X_t, X_{t-1}, \ldots, X_{t-d+1})$. We estimate H by

$$\mathrm{argmin}_{g \in \mathcal{G}} \sum_{t=d+1}^{n} (X_t - g(Y_{t-1}))^2$$

For practical implementation, consider those Y_{t-1} in a compact interval C:

$$\mathrm{argmin}_{g \in \mathcal{G}} Q_n(g), \text{ where } Q_n(g) = \sum_{t=d+1}^{n} (X_t - g(Y_{t-1}))^2 \omega(Y_{t-1})$$

and the weight function $\omega(y) = 1$ if $y \in C$ and $\omega(y) = 0$ if $dist(y, C) \geq a$ for some $a > 0$, for example $\omega(y) = (1 - dist(y, C))^+$.

The function class $\mathcal{G}$

A neural network with L layers, N_l nodes at the lth layer, $1 \leq l \leq L$, input dim $N_0 = d$, output dim $N_{L+1} = 1$, and rectifier linear unit (ReLU) activation function $\sigma(x) = x^+$, and

$$f(x) = W_L \sigma_{v_L} \dots W_1 \sigma_{v_1} W_0 x$$

Let $\mathcal{F}_M(L, N, s)$ be the sparse networks of such f with at most s non-zero weights contained in $[-M, M]$. Then the entropy of the covering number

$$\log_2 \mathcal{N}(\delta, \mathcal{F}_1(L, N, s), |\cdot|_\infty) \leq 4sL \log_2(8\delta^{-1} L \max_{l \leq L} N_l) \tag{11}$$

when $C = [0, 1]^d$; see Schmidt-Hieber (2020), Ohn and Kim (2022), Beknazaryan and Sang (2022) among others.

Non-parametric Estimation of Nonlinear AR Processes

Let $\mathcal{G} \subset \mathcal{F}_1(L, N, s)$ with $|f|_\infty \le c_1$ and Lipschitz constant $\le c_2$.

Theorem

Assume that $\mu_p = \mathbb{E}|\varepsilon_t|^p < \infty$, $p > 2$, and

$$z/\sqrt{n} \ge K_1 sL \log(8L \max_l N_l) \tag{12}$$

where constant K_1 depends on c_1, c_2, p, μ_p and ρ. Then

$$\mathbb{P}(\max_{g \in \mathcal{G}} |Q_n(g) - \mathbb{E}Q_n(g)| \ge z) \le K_2 \frac{n^{1+p/2}}{z^p} (\log n)^{p/2} \tag{13}$$

Thank you!