Nonparametric Quantile Regression with ReLU-Activated Recurrent Neural Networks

Hang Yu

Lyumin Wu

Wen-Xin Zhou

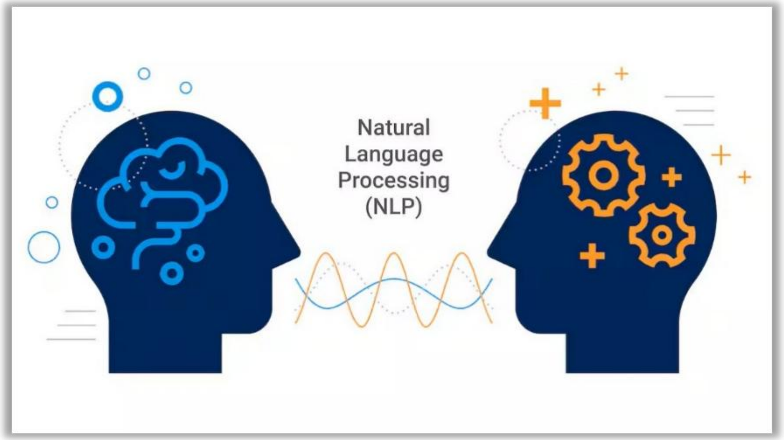
Zhao Ren

Presented by Hang Yu

2025.11.06

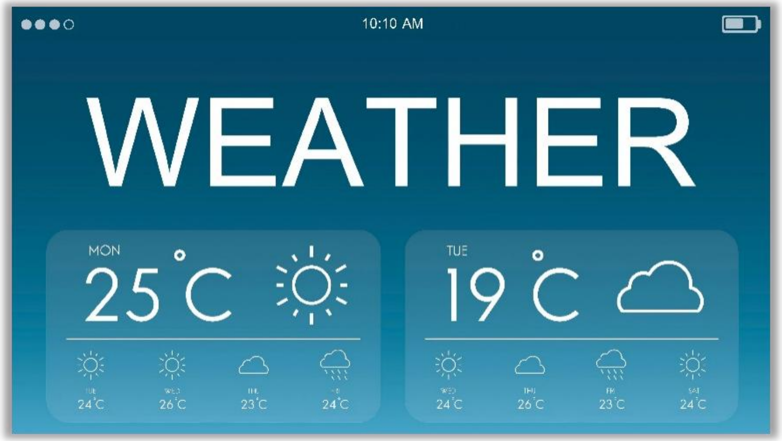
Recurrent Neural Networks (RNNs)

- RNNs have achieved ***remarkable success*** in various applications.



Natural Language Processing

Finance



Weather Forecasting

However, the ***theoretical*** foundations of RNNs remain ***incomplete***.

RNNs and Sparse RNNs (SRNNs)

☐ RNN structure: Many-to-one setting

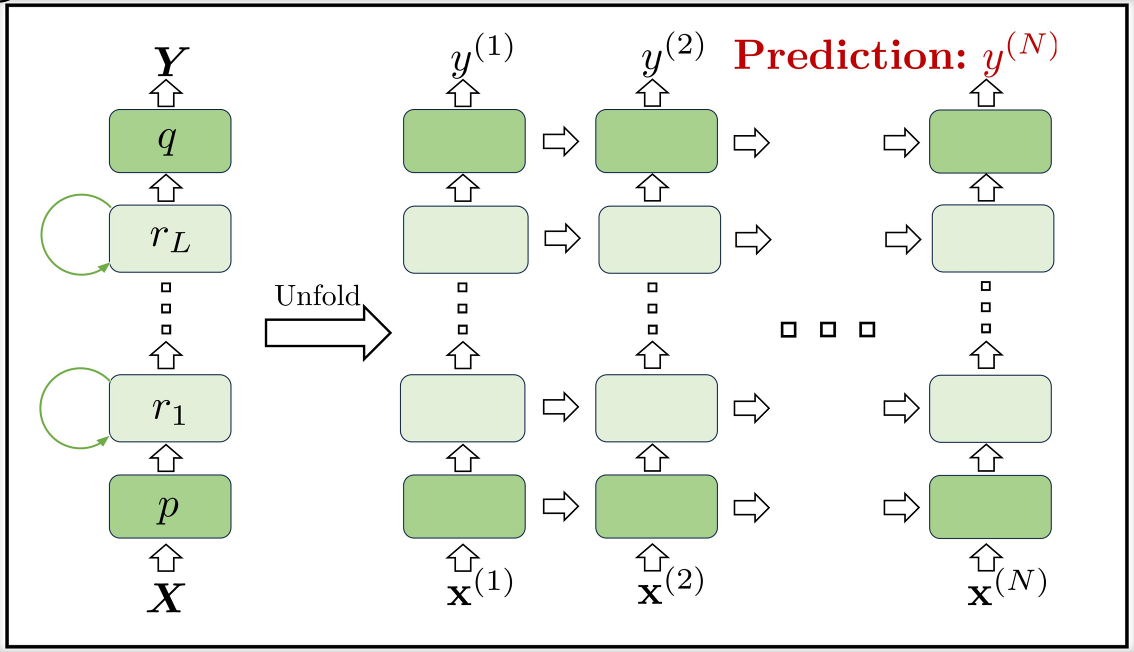
Given the width W, the length L, and the time horizon N, an RNN processes an input sequence $\boldsymbol{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)})$ sequentially through

$$\text{Input layer } p,$$

$$\text{Recurrent layers } \{r_l\}_{l=1}^{L},$$

$$\text{Output layer } q,$$

outputs $\boldsymbol{Y} = (y^{(1)}, \ldots, y^{(N)})$, and obtains the final prediction $y^{(N)}$.



☐ SRNN: The *sparsity* s of an RNN is defined as the number of its nonzero nodes.

We focus on establishing theoretical guarantees for RNNs/SRNNs in a fundamental area — ***quantile regression***.

Quantile Regression

Problem setup

Consider the sequentially ***stationary*** observations $\{(\mathbf{x}_t, y_t)\}_{t=1}^n$, where any consecutive N observations share the same joint distribution as $Z = ((X_1, Y_1), \dots, (X_N, Y_N))$. Given a quantile level $\tau \in (0,1)$ of interest, we define the conditional τ-th quantile of y_t (or Y_N) given $\mathbf{x}_{t-N+1}, \dots, \mathbf{x}_t$ (or $X_1, \dots, X_N$) as

$$q_\tau(y_t|\mathbf{x}_{t-N+1}, \dots, \mathbf{x}_t) = f_0(\mathbf{x}_{t-N+1}, \dots, \mathbf{x}_t), \; N \le t \le n,$$

where $f_0 : \mathbb{R}^{d_{\mathbf{x}} \times N} \to \mathbb{R}$ is the unknown conditional quantile function.

Empirical risk minimization estimator:

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \mathcal{R}_n(f) := \frac{1}{n - N + 1} \sum_{t=N}^{n} \rho_\tau(y_t - f(\mathbf{x}_{t-N+1}, \dots, \mathbf{x}_t)),$$

where $\rho_\tau(u) = (\tau - \mathbb{1}(u < 0))u$ is the check loss.

Quantile Regression with RNNs/SRNNs

NEURAL INFORMATION PROCESSING SYSTEMS

Problem setup

We consider the setting where $\mathcal{F}$ is the class of RNNs/SRNNs.

Consider the sequentially ***stationary*** observations $\{(\mathbf{x}_t, y_t)\}_{t=1}^n$, where any consecutive N observations share the same joint distribution as $Z = ((X_1, Y_1), \ldots, (X_N, Y_N))$. Given a quantile level $\tau \in (0,1)$ of interest, we define the conditional τ-th quantile of y_t (or Y_N) given $\mathbf{x}_{t-N+1}, \ldots, \mathbf{x}_t$ (or $X_1, \ldots, X_N$) as

$$q_\tau(y_t | \mathbf{x}_{t-N+1}, \ldots, \mathbf{x}_t) = f_0(\mathbf{x}_{t-N+1}, \ldots, \mathbf{x}_t), \;\; N \le t \le n,$$

where $f_0 : \mathbb{R}^{d_{\mathbf{x}} \times N} \to \mathbb{R}$ is the unknown conditional quantile function.

Empirical risk minimization estimator:

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \mathcal{R}_n(f) := \frac{1}{n-N+1} \sum_{t=N}^{n} \rho_\tau(y_t - f(\mathbf{x}_{t-N+1}, \ldots, \mathbf{x}_t)),$$

where $\rho_\tau(u) = (\tau - \mathbb{1}(u < 0))u$ is the check loss.

Hölder Class

- Classically, we assume the true function f_0 belongs to a Hölder class.

Definition 1 (Hölder Class of Functions $\mathcal{C}_d^\beta(\mathcal{X}, K)$)**.** Given a domain $\mathcal{X} \subseteq \mathbb{R}^d$, a positive Hölder smoothness parameter β, and a constant $K > 0$, the β-Hölder function class is defined as

$$\mathcal{C}_d^\beta(\mathcal{X}, K) = \left\{ f : \mathcal{X} \to \mathbb{R} \,\middle|\, \sum_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_1 < \beta} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + \sum_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_1 = r} \sup_{\substack{\mathbf{x},\mathbf{y}\in\mathcal{X} \\ \mathbf{x}\neq\mathbf{y}}} \frac{|\partial^{\boldsymbol{\alpha}} f(\mathbf{x}) - \partial^{\boldsymbol{\alpha}} f(\mathbf{y})|}{\|\mathbf{x}-\mathbf{y}\|_2^s} \le K \right\},$$

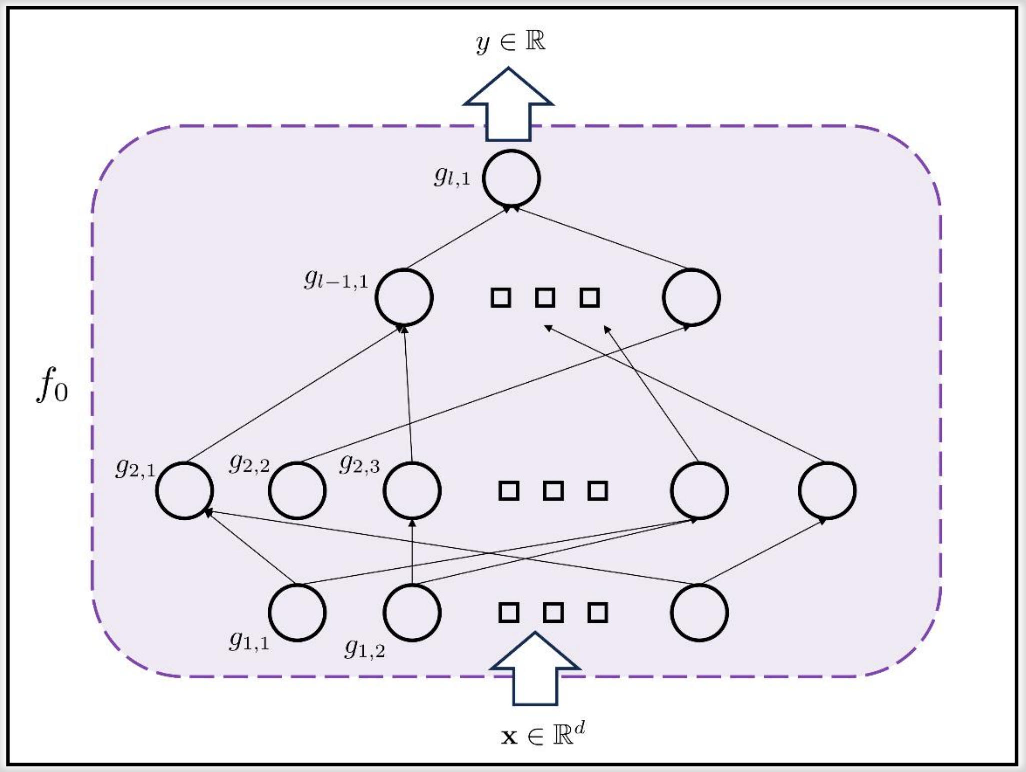
where $r = \lfloor \beta \rfloor$, $s = \beta - r$, $\partial^{\boldsymbol{\alpha}} = \partial^{\alpha_1} \cdots \partial^{\alpha_d}$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ and $\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^d \alpha_i$. Moreover, we refer to $\gamma = \beta/d$ as the dimension-adjusted degree of smoothness of $\mathcal{C}_d^\beta(\mathcal{X}, K)$.

Stone (1982) established that the ***minimax convergence rate*** for estimating a regression function under the L_2 norm over the Hölder class $\mathcal{C}_d^\beta(\mathcal{X}, K)$ is $n^{-\gamma/(2\gamma+1)}$.

Curse of Dimensionality*: In RNN applications, d is often* ***large****, resulting in a* ***small*** *value of the dimension-adjusted smoothness γ, which in turn leads to* ***slow*** *convergence rates.*

Hierarchical Interaction Model

- Assume f_0 belongs to the hierarchical interaction model $\mathcal{H}_d^l(\mathcal{P}, K)$. Then f_0 exhibits a ***compositional structure*** as follows:



- Each function $g_{i,j}$ belongs to a Hölder class.
- We define the intrinsic smoothness of $\mathcal{H}_d^l(\mathcal{P}, K)$ by

$$\gamma^\star = \beta^\star / t^\star, \quad \text{where } (\beta^\star, t^\star) = \operatorname{argmin}_{(\beta,t)\in\mathcal{P}} \beta/t.$$

- $\gamma^\star$ does not depend on the ambient input dimension, thereby ***mitigating*** the curse of dimensionality.

Stationary β-mixing

- Since RNNs naturally handle dependent sequences, we assume the data are stationary and β-mixing instead of i.i.d.

Definition 2 (β-mixing (Bradley, 1983)). *Let $\{\mathbf{z}_t\}_{t=-\infty}^{\infty}$ be a sequence of random vectors. For any $i, j \in \mathbb{Z} \cup \{-\infty, +\infty\}$, define $\sigma_i^j = \sigma(\mathbf{z}_i, \mathbf{z}_{i+1}, \ldots, \mathbf{z}_j)$ as the σ-algebra generated by $\mathbf{z}_k, i \leq k \leq j$. For any $a \in \mathbb{N}$, the $\boldsymbol{\beta}$-mixing coefficient of the stochastic process $\{\mathbf{z}_t\}_{t=-\infty}^{\infty}$ is defined as*

$$\boldsymbol{\beta}(a) = \sup_{k \geqslant 1} \mathop{\mathbb{E}}_{B \in \sigma_{-\infty}^{k}} \left[\sup_{A \in \sigma_{k+a}^{\infty}} |\mathbb{P}(A|B) - \mathbb{P}(A)| \right].$$

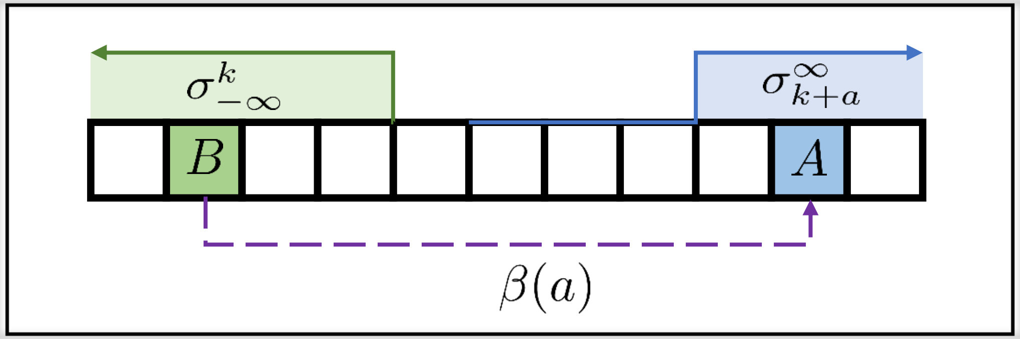
quantifies maximum dependence between past and future observations

- Algebraically $\boldsymbol{\beta}$-mixing:

$$\boldsymbol{\beta}(a) \leq \beta_0 / a^r, \quad \forall a.$$

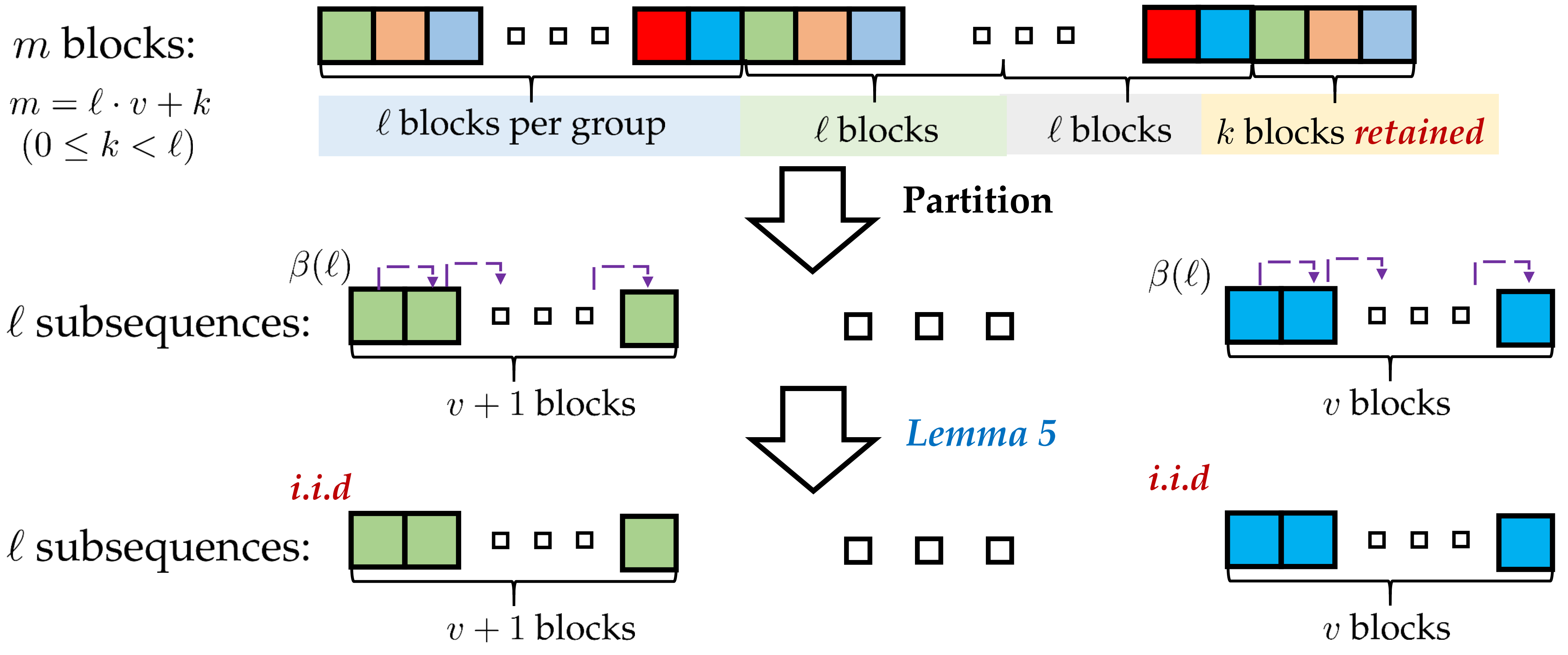
- Exponentially $\boldsymbol{\beta}$-mixing:

$$\boldsymbol{\beta}(a) \leq \beta_0 \exp(-\beta_1 a^r), \quad \forall a.$$



Refined Blocking Technique

- Relate ***mixing*** to ***i.i.d***



Donut-set Decomposition

- In analysis, an important quantity is

$$\mathbb{P}\left(\|\widehat{f}-f_0\|_2 > \delta_\star\right),$$

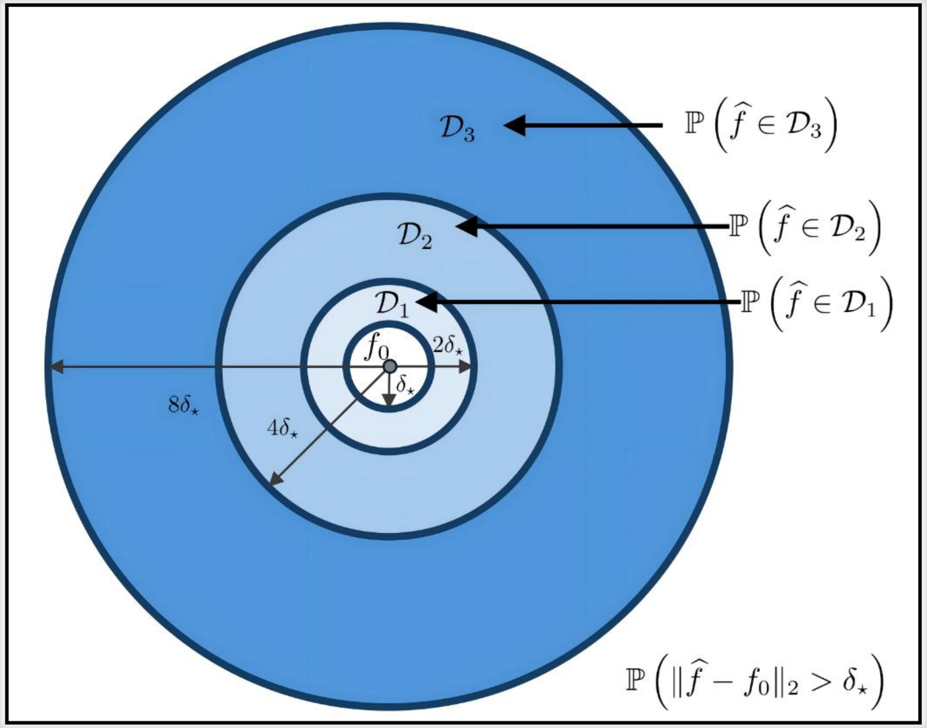
where $\delta_\star = c(\delta_a+\delta_b+\delta_\beta)$, δ_a is the approximation error, δ_b is the stochastic error, and δ_β is the dependence error.

Donut-set Decomposition:

$$\mathbb{P}\left(\|\widehat{f}-f_0\|_2 > \delta_\star\right) \le \sum_{i=1}^{\lfloor \log_2(2K/\delta_\star)\rfloor} \mathbb{P}\left(\widehat{f}\in\mathcal{D}_i\right),$$

where

$$\mathcal{D}_i = \left\{f\in\mathcal{F} : 2^{i-1}\delta_\star < \|f-f_0\|_2 \le 2^i\delta_\star\right\}.$$



Theorem

NEURAL INFORMATION PROCESSING SYSTEMS

☐ Convergence Rate

Theorem 1. Let $\mathcal{RNN}_{d_{\mathbf{x}},1}(W, L, K)$ be the hypothesis class $\mathcal{F}$ and suppose regularity conditions, the hierarchical assumption, and the continuous probability measure assumption.

(i) Under the ***exponentially*** mixing assumption, Let $W_0, L_0 \geq 3$ satisfy $W_0 L_0 \asymp \left(n/(\log n)^{(6+1/r)}\right)^{1/(4\gamma^\star+2)}$, and choose $W = cW_0 \log W_0$ and $L = cL_0 \log L_0$. Then, the ERM estimator $\widehat{f}$ satisfies

$$\|\widehat{f} - f_0\|_2 = \mathcal{O}_p\left(n^{-\gamma^\star/(2\gamma^\star+1)}(\log n)^{(6+1/r)\gamma^\star/(2\gamma^\star+1)}\right).$$

Minimax optimal!

(ii) Under the ***algebraically*** mixing assumption, Let $W_0, L_0 \geq 3$ satisfy $W_0 L_0 \asymp \left(n^{(1-1/r)}/(\log n)^7\right)^{1/(4\gamma^\star+2)}$, and choose $W = cW_0 \log W_0$ and $L = cL_0 \log L_0$. Then, the ERM estimator $\widehat{f}$ satisfies

$$\|\widehat{f} - f_0\|_2 = \mathcal{O}_p\left(n^{-(1-1/r)\gamma^\star/(2\gamma^\star+1)}(\log n)^{(7\gamma^\star/(2\gamma^\star+1))}\right).$$

Experiments

Simulations

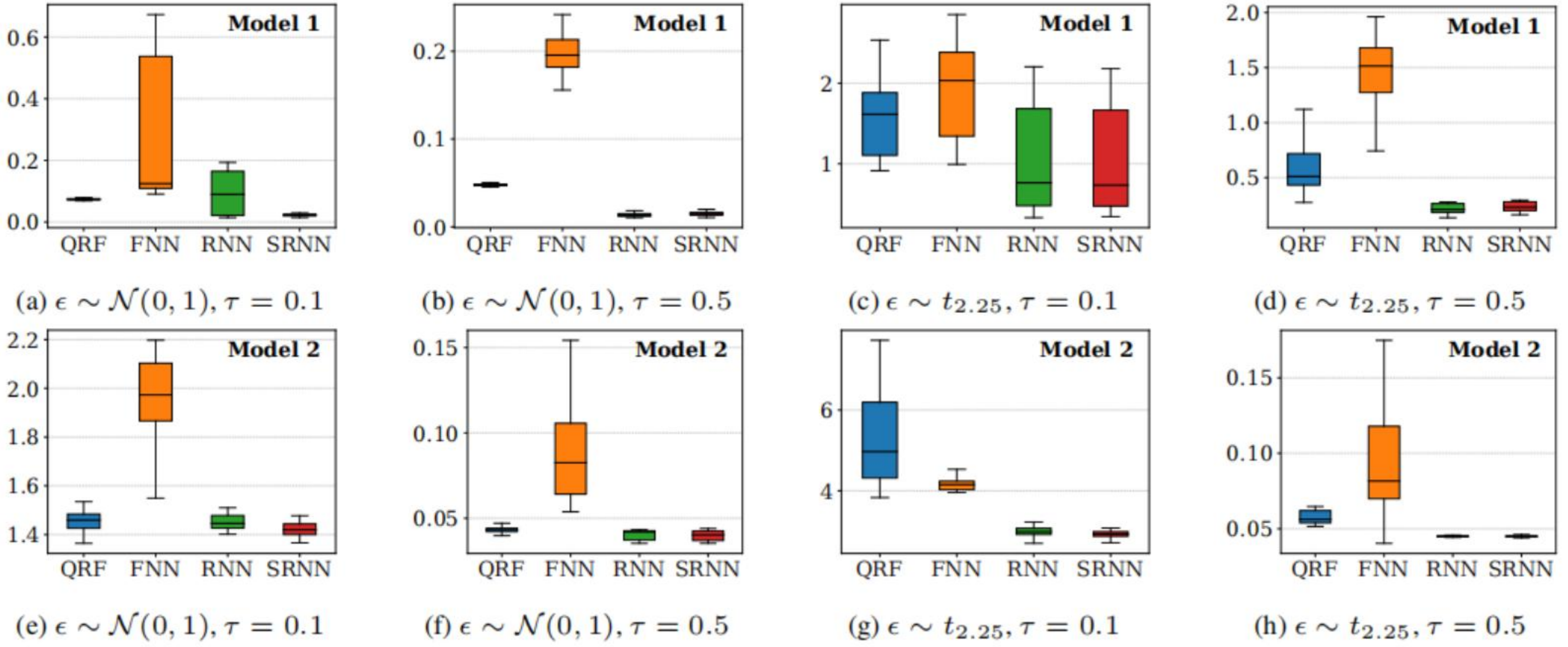


Figure 1: MSE comparison for Model 1 and Model 2 with different noise distributions and quantile levels.

Real data analysis

- Dow Jones Industrial Average (DJIA) analysis: ***stationary***

Model	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
QRF	0.456	0.698	0.817	0.616	0.365
FNN	0.538	0.735	0.810	0.662	0.404
RNN	0.410	**0.640**	0.760	0.562	0.306
SRNN	**0.406**	0.647	**0.759**	**0.561**	**0.305**

Table 1: Out-of-sample prediction errors at different quantiles for DJIA growth analysis.

- GDP analysis: ***non-stationary***

Model	$\tau = 0.1$	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 0.9$
QRF	0.849	1.225	1.410	1.246	0.911
FNN	0.867	1.180	1.773	2.505	2.657
RNN	**0.835**	**1.113**	1.349	**1.154**	0.904
SRNN	0.837	1.700	**1.211**	1.200	**0.898**

Table 2: Out-of-sample prediction errors at different quantiles for GDP growth analysis.

Summary

- Problem: ***Quantile regression*** *with* ***RNNs/SRNNs***
- Underlying function class: **Hierarchical Interaction Model**
- Data assumption: ***Stationary and*** β***-mixing***
- Two technique: ***Blocking technique and donut-set decomposition***
- Experiments: Simulations, DJIA analysis, and GDP analysis

Thanks!