

Operator Learning for Functional Autoregressive Models

Shuntaro Suzuki Graduate School of Engineering Science, Osaka University
Yoshikazu Terada Graduate School of Engineering Science, Osaka University; RIKEN AIP

1 Introduction

Recent advances in sensing technologies have led to the increasing availability of spatio-temporal data that are observed densely over a spatial domain at successive time points. Operator learning has emerged as a promising machine learning framework for analyzing such data. Operator learning methods approximate mappings between function spaces using deep neural networks (DNNs), and have been applied extensively to data-driven physical simulations.

Although many applications of operator learning can naturally be viewed as functional time series problems, theoretical guarantees for operator learning methods applied to forecasting functional time series remain largely unexplored. In this work, we investigate the statistical properties of operator learning for functional time series prediction. We define the function space $\mathcal{H}$ by $\mathcal{H} := \left\{ f : [0, 1]^d \rightarrow \mathbb{R} \mid \int_{[0, 1]^d} |f(u)|^2 du < \infty \right\}$. We consider the following nonlinear functional autoregressive (NFAR) model on $\mathcal{H}$:

$$X_{t+1}(u) = \Psi_0(X_t)(u) + \xi_t(u), \quad u \in D, t \in \mathbb{N}, \quad (1.1)$$

where $\Psi_0 : \mathcal{H} \rightarrow \mathcal{H}$ is a nonlinear operator.

Since most operator learning architectures are designed to approximate nonlinear operators, the NFAR model provides a natural framework for formulating operator learning for functional time series. Moreover, the nonlinear structure of Ψ_0 allows the model to capture complex spatio-temporal dynamics.

Despite its flexibility, relatively little is known about estimation of the operator Ψ_0 in NFAR models, and a general theoretical foundation has yet to be established. For example, Zhu and Politis (2017) [3] proposed a kernel estimator for Ψ_0 and established its asymptotic normality under an exponential β -mixing condition. However, sufficient conditions under which an NFAR process itself satisfies this mixing condition were not provided.

Mixing conditions quantify the strength of temporal dependence in dependent data and play a fundamental role in establishing statistical guarantees for nonparametric estimators, including deep-learning estimators, under temporal dependence. In infinite dimensional spaces, however, probability measures can easily become mutually singular. Consequently, verifying mixing properties for NFAR processes is substantially more delicate than in finite-dimensional settings.

To address this issue, we identify a set of conditions under which the finite-dimensional Harris-type theory of Hairer and Mattingly (2011) [1] can be adapted to the infinite-dimensional

NFAR setting, and establish sufficient conditions for exponential β -mixing. As an application of this result, we derive an oracle inequality for an operator-learning estimator of Ψ_0 and obtain generalization-error bounds for a particular class of deep neural operator-learning models.

2 Notation and assumption

We first introduce notation used throughout this paper. For an operator $\Phi : \mathcal{H} \rightarrow \mathcal{H}$, let $\text{Im}(\Phi)$ denote its range. Let $\{\xi_t\}_{t \in \mathbb{Z}}$ be a sequence of mean-zero $\mathcal{H}$ -valued Gaussian random elements. For each $t \in \mathbb{N}$, let the covariance kernel of ξ_t be $k(u, v) := \mathbb{E}[\xi_1(u)\xi_1(v)]$, $u, v \in [0, 1]^d$. The corresponding covariance operator $\mathcal{Q} : \mathcal{H} \rightarrow \mathcal{H}$ is defined by

$$(\mathcal{Q}f)(u) := \int_{[0,1]^d} k(u, v)f(v)dv, \quad u \in [0, 1]^d,$$

for $f \in \mathcal{H}$. For the time series $\{X_t\}_{t \in \mathbb{N}}$, define the β -mixing coefficient by

$$\beta(j) := \sup_{n \in \mathbb{N}} \mathbb{E} \left[\sup_{B \in \mathcal{F}_{j+n}^\infty} |\mathbb{P}(B) - \mathbb{P}(B \mid X_n, \dots, X_1)| \right],$$

where $\mathcal{F}_{j+n}^\infty := \sigma(X_{j+n}, X_{j+n+1}, \dots)$.

3 Main results

We first state assumptions that will be used throughout this section.

Assumption 3.1. For every $t \in \mathbb{N}$, ξ_t is independent of $X_{t-1}, \dots, X_1$.

Assumption 3.2. $\overline{\text{Im}(\mathcal{Q})} = \mathcal{H}$.

3.1 Exponential mixing of nonlinear functional autoregressive processes

We first establish exponential mixing of the NFAR model. The following additional assumptions are imposed.

Assumption 3.3. There exist constants $c_1, c_2 > 0$ satisfying $c_1 \|\mathcal{Q}\|_{\text{op}} < 1$ such that $\|m_0(x)\|_{\mathcal{H}} \leq c_1 \|x\|_{\mathcal{H}} + c_2$ for every $x \in \mathcal{H}$.

Assumption 3.4. There exists an operator $m_0 : \mathcal{H} \rightarrow \mathcal{H}$ such that $\Psi_0 = \mathcal{Q} \circ m_0$.

Theorem 3.1 (Exponential mixing of nonlinear functional autoregressive processes). Suppose that Assumption (3.1) and Assumption (3.2) hold. Suppose further that Assumption (3.3) and Assumption (3.4) hold. Then the NFAR process $\{X_t\}$ is exponentially β -mixing; that is, there exist constants $C_1, C_2 > 0$ such that

$$\beta(j) \leq C_1 \exp(-C_2 j),$$

for any $j \in \mathbb{N}$.

Theorem 3.1 provides sufficient conditions under which a nonlinear functional autoregressive process is exponentially β -mixing. We next give a concrete class of operators Ψ_0 for which the conditions above can be verified.

Example 3.1. Consider operators of the form

$$\Psi_0(x)(u) := \int_D f(u, v) \tau(v, x(v)) dv \quad u \in D, x \in \mathcal{H},$$

where $f : D \times D \rightarrow \mathbb{R}$ and $\tau : D \times \mathbb{R} \rightarrow \mathbb{R}$. We provide sufficient conditions on f and τ under which Assumption (3.3) holds. Let $\{\lambda_i\}_{i=1}^\infty$ denote the eigenvalues of $\mathcal{Q}$ and suppose that $\sum_{i=1}^\infty \lambda_i < \infty$. Define $\mathcal{H}_{\tilde{k}} \otimes \mathcal{H}_{\tilde{k}}$ by

$$\mathcal{H}_{\tilde{k}} \otimes \mathcal{H}_{\tilde{k}} := \left\{ f : D \times D \rightarrow \mathbb{R} \left| \sum_{i,j=1}^\infty \frac{f_{ij}^2}{\lambda_i^2 \lambda_j^2} < \infty \right. \right\}.$$

If $f \in \mathcal{H}_{\tilde{k}} \otimes \mathcal{H}_{\tilde{k}}$ and $\sup_{x \in \mathbb{R}, u \in D} |\tau(u, x)| < \infty$, then the resulting operator Ψ_0 satisfies Assumption (3.3). The former condition can be interpreted as a smoothness condition on the kernel f .

3.2 An oracle inequality for operator learning in time-series prediction

We next derive an oracle inequality for an operator-learning estimator used to predict the NFAR model (1.1). We focus on operators of the form

$$\Psi_0(x)(u) := \int_{[0,1]^d} \psi_0(u, v, x(v)) dv, \quad u \in [0, 1]^d,$$

where $\psi_0 : [0, 1]^d \times [0, 1]^d \times \mathbb{R} \rightarrow \mathbb{R}$. We consider an operator-learning procedure that estimates the operator Ψ_0 through estimation of the function ψ_0 . Suppose that observations $\{X_t\}_{t=1}^{T+1}$ generated from (1.1) are available. For a candidate function ψ , define the empirical loss

$$L(\psi) := \sum_{t=1}^{T-1} \int_D \left\{ X_{t+1}(u) - \int_D \psi(u, v, X_t(v)) dv \right\}^2 du.$$

We define the estimator

$$\hat{\psi}_T := \arg \min_{\psi \in \mathcal{F}} L(\psi),$$

where $\mathcal{F}$ is a prescribed function class. For each $\psi \in \mathcal{F}$, let the corresponding operator Ψ_ψ be defined by

$$\Psi_\psi(x)(u) := \int_{[0,1]^d} \psi(u, v, x(v)) dv \mathbf{1}_{\{\|x\|_\infty \leq 1\}}, \quad u \in D, \quad x \in \mathcal{H}.$$

The resulting estimator of Ψ_0 is $\widehat{\Psi}_T := \Psi_{\widehat{\psi}_T}$.

We now introduce the risk used to evaluate the estimator. For an operator $\Psi : \mathcal{H} \rightarrow \mathcal{H}$, define its generalization error relative to Ψ_0^* by

$$R(\Psi, \Psi_0^*) := \mathbb{E} [|\Psi(X^*) - \Psi_0^*(X^*)|_{\mathcal{H}}^2],$$

where $\Psi_0^*(x) := \Psi_0(x)\mathbf{1}_{\{\|x\|_{\infty} \leq 1\}}$, and X^* is an independent copy having the same marginal distribution as $\{X_t\}_{t=1}^{T+1}$.

To quantify the complexity of the operator-learning class, we use the covering number $\mathcal{N}(\delta, (\Psi_{\psi})_{\psi \in \mathcal{F}}, \|\cdot\|_{\infty})$, where, for an operator Ψ , $\|\Psi\|_{\infty} := \sup_{x \in \mathcal{H}} \|\Psi(x)\|_{\mathcal{H}}$.

We impose the following additional assumptions.

Assumption 3.5. For every t , the sample path $u \mapsto X_t(u)$ is continuous on D .

Assumption 3.6. The process $\{X_t\}_{t \in \mathbb{N}}$ satisfying (1.1) is exponentially β -mixing.

Theorem 3.2 (Oracle inequality). Suppose that Assumption (3.1), Assumption (3.5), and Assumption (3.6) hold. Under suitable regularity conditions on the operator class $(\Psi_{\psi})_{\psi \in \mathcal{F}}$, we have

$$R(\Psi_{\widehat{\psi}_T}, \Psi_0^*) \lesssim \inf_{\psi \in \mathcal{F}} R(\Psi_{\psi}, \Psi_0^*) + \frac{(\log \mathcal{N}_T)(\log T)}{T},$$

where $\mathcal{N}_T := \mathcal{N}\left(1/T, (\Psi_{\psi})_{\psi \in \mathcal{F}}, \|\cdot\|_{\infty}\right)$.

This result extends the finite-dimensional oracle inequality of Kurisu, Fukami, and Koike (2025) [2] to an infinite-dimensional functional setting. The oracle inequality can subsequently be combined with approximation and complexity bounds to derive generalization-error rates for a broad class of operator-learning architectures.

3.3 Convergence rates for the generalization error

We next specialize the preceding result to ReLU-based deep neural networks. The network takes $u = [u_1, \dots, u_d]^{\top}, v = [v_1, \dots, v_d]^{\top}, x(v)$ as inputs and returns the pointwise value $\psi(u, v, x(v)) \in \mathbb{R}$. A schematic illustration of the architecture is shown below.

We next specify the function class containing the true function ψ_0 . Let $\mathcal{C}^{\beta}(Q)$ denote a Hölder class with smoothness parameter β and radius Q . Define the compositional function class $\mathcal{G}_{\text{comp}}(q, d, t, \beta, 1)$ by

$$\{g_q \circ \dots \circ g_0 \mid g_i = (g_{ij})_j, g_i : [a_i, b_i]^{d_i} \rightarrow [a_{i+1}, b_{i+1}]^{d_{i+1}}, g_{ij} \in \mathcal{C}^{\beta_i}([a_i, b_i]^{t_i}, Q), \text{ for some } |a_i|, |b_i| \leq Q\}.$$

Corollary 3.1. Suppose that Assumption (3.1), Assumption (3.5), and Assumption (3.6) hold. Assume that $\psi_0 \in \mathcal{G}_{\text{comp}}(q, d, t, \beta, 1)$, and choose the neural-network hyperparameters appropriately. Then

$$R(\Psi_{\widehat{\psi}_T}, \Psi_0^*) \lesssim \phi_T (\log T)^4,$$

where $\phi_T := \min_{i=0, \dots, d} T^{-\frac{\beta_i}{t_i + \beta_i}}$.

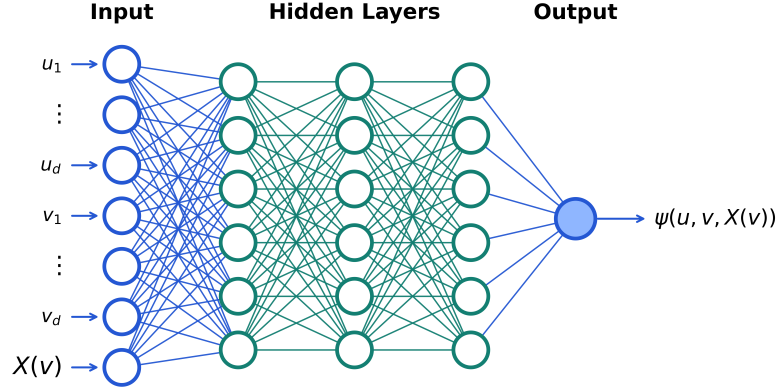


Figure 1: Deep neural network architecture for approximating ψ_0 .

Thus, up to logarithmic factors, the proposed estimator attains the minimax-optimal convergence rate associated with the corresponding finite-dimensional nonparametric regression problem.

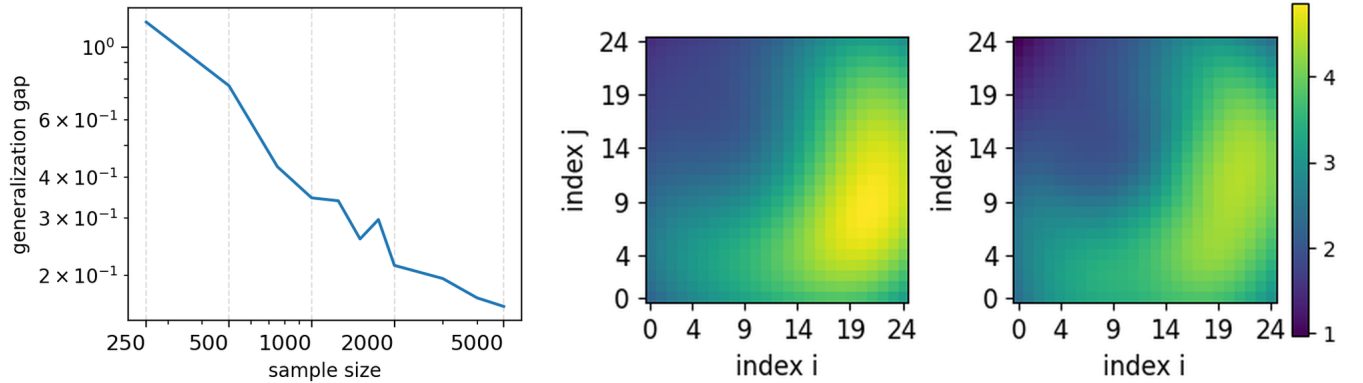


Figure 2: Numerical illustration of the convergence of the generalization error. The left panel shows the generalization error as a function of the training sample size. The middle and right panels compare the outputs of the true operator Ψ_0 and the estimated operator $\hat{\Psi}_T$, respectively.

References

- [1] M. Hairer and J. C. Mattingly. “Yet another look at Harris’ ergodic theorem for Markov chains”. In: *Seminar on Stochastic Analysis, Random Fields and Applications VI*. Basel: Springer, 2011, pp. 109–117.
- [2] D. Kurisu, R. Fukami, and Y. Koike. “Adaptive deep learning for nonlinear time series models”. In: *Bernoulli* 31.1 (2025), pp. 240–270.

- [3] T. Zhu and D. Politis. “Kernel estimates of nonparametric functional autoregression models and their bootstrap approximation”. In: *Electron. J. Stat.* 11 (2017), pp. 2876–2906.