# A Subsampling Based Neural Network for Spatial Data

Debjoy Thakur[1*]

[1*]Department of Statistics and Data Science, Washington Univeristy in St. Louis,  St. Louis, 63130, Missouri, USA.

Corresponding author(s). E-mail(s): debjoy@wustl.edu;

### Abstract

The application of deep neural networks in geospatial data has become a trending research problem in the present day. A significant amount of statistical research has already been introduced, such as generalized least square optimization by incorporating spatial variance-covariance matrix, considering basis functions in the input nodes of the neural networks, and so on. However, for lattice data, there is no available literature about the utilization of asymptotic analysis of neural networks in regression for spatial data. This article proposes a consistent localized two-layer deep neural network-based regression for spatial data. We have proved the consistency of this deep neural network for bounded and unbounded spatial domains under a fixed sampling design of mixed-increasing spatial regions. We have proved that its asymptotic convergence rate is faster than that of [1]'s neural network and an improved generalization of [2]'s neural network structure. We empirically observe the rate of convergence of discrepancy measures between the empirical probability distribution of observed and predicted data, which will become faster for a less smooth spatial surface. We have applied our asymptotic analysis of deep neural networks to the estimation of the monthly average temperature of major cities in the USA from its satellite image. This application is an effective showcase of non-linear spatial regression. We demonstrate our methodology with simulated lattice data in various scenarios.

**Keywords:** Spatial Subsampling, Deep Neural Network, Lattice Data, Convergence

# 1 Introduction

In recent years, the application of machine learning and neural network models in spatial statistics has been a highly demanding and challenging research problem. Many researchers have focused on the application of neural networks in geospatial data sets for example pollution data, temperature data, spatial disease mapping, spatial econometrics, and so on. The neural network becomes popular day by day for its flexibility, and prediction accuracy. From a statistical point of view, many researchers focus on its explainability. For example [3] has developed the convergence rate of a multilayer compositional neural network for independent and identically distributed (iid) feature space. [4] has proved the optimum minimax rate of convergence of a fully connected deep neural network. [5] has proved the asymptotic convergence of convolutional neural networks for image classifiers. [6] has proved the asymptotic convergence rate of shallow neural networks. [2] has proved the asymptotic convergence and asymptotic normality of the sieve neural network. These articles critically review the asymptotic convergence of neural networks under some regularity conditions.

Now coming to the application of neural networks in spatial data there are also some significant contributions. [7] has introduced "Deep Kriging" (DK) by considering the basis functions in the input node of deep neural network similar concept of fixed rank kriging (FRK) of [8]. [9] has extended this DK in a spatio-temporal framework. [10] estimate parameter of Matern covariance function. [1] has introduced a generalized least square-based graphical neural network for geospatial data. [11] has proposed a neural network-based spatial autoregressive model. [12] has reviewed almost all recently proposed different types of neural network-based models for spatial and spatio-temporal data.

The literature has a significant gap regarding the asymptotic inference of neural network estimates for spatial data. From the literature, all the researchers mainly focus on solving two major problems, estimating mean function ([7]), and parameter estimation of covariance function ([10]). But [1] has first proposed the spatial dependent graphical neural networking assuming the spatial points as nodes of neural network. But still, some open research questions arise.

Question-1 *How the neighborhood size "M" contributes in asymptotic convergence of [1]'s graphical neural network?*

Question-2 *[7]'s neural network considering [13]'s multi-resolution radial basis in the input nodes and [9]'s long short time memory (LSTM) extension ignore the fact that for every spatial location the basis coefficients are different. As a result, it's not clear how they consider spatial random effects via neural networks.*

Question-3 *[11] only considers the spatial random effect of response and covariate space are linearly additive ignoring the nonlinear strength of the neural network. Here how this network is convergent is unexplained.*

In this research, we have sorted these three problems of the previous research. We have proposed a nonlinear spatial neural network for spatial data which is a generalization of [11]'s spatial regression model. We have generalized [2]'s single-layer sieve neural network into the deep neural network by increasing the width and depth. We explain how convergence is achieved by increasing neighborhood size under some regularity

2

conditions with a faster convergence rate than that of [1]. To overcome the short-comings of considering spatial random effect in "Deep Kriging" of [7] we introduce a localized deep neural network.

In Section (2.1) we have described the spatial sampling design that is required for the inference of localized deep neural network and then we have discussed the details of the algorithm of localized deep neural network in spatial regression. In Section (2.2) we have discussed the asymptotic convergence, convergence rate, and central limit theorem of the weighted sum of localized deep neural network function in the mixed increasing domain under fixed sampling design. In Section (2.3) we explain the confidence interval of localized deep neural network functional, asymptotic behavior of KL divergence. In Section (3) we empirically validate our theoretical results by different simulated data and one monthly average temperature data of major cities of the USA and satellite image.

# 2 Methodology

## 2.1 Localized DNN

Nowadays deep neural networks are famous for their approximation quality. For spatial framework, it's very common to assume the nodes in the location and then perform a neural network modeling. In this section, we will study a localized neural network that will apprehend the spatial dynamicity along with the relationship between response and covariates. This model is a potential generalization of spatial regression. This section is partitioned into two parts. The first part will explain how to import the spatial dependence and the second part will give an overview of how we can execute the algorithm.

### 2.1.1 Spatial Sampling Regions

Let's consider an open connected set of $(-\frac{1}{2}, \frac{1}{2}]^d$ is $\mathcal{R}_0^*$ and a prototype Borel set $\mathcal{R}_0$ satisfying $\mathcal{R}_0^* \subset \mathcal{R}_0 \subset \text{Clo}.\mathcal{R}_0^*$. Let $\{\lambda_n\}_{n \in \mathbb{N}}$ be a sequence of non-decreasing positive real numbers ($\lambda_n = o(n)$) such that $\lambda_n \to \infty$ as $n \to \infty$. The sampling region $\mathcal{R}_n$ is achieved by 'inflating' the prototype set $\mathcal{R}_0$ by a scaling factor $\lambda_n$

$$\mathcal{R}_n = \lambda_n \mathcal{R}_0.$$

As a result, the sampling region $\mathcal{R}_n$ becomes unbounded. Now for a mixed increasing spatial sampling design like [14] we assume a diagonal matrix $\Delta = \text{diag}(e_1, e_2, \cdots, e_d)$ and $\mathcal{Z}^d = \{\Delta \mathbf{i} : \mathbf{i} \in \mathbb{Z}^d\}$ where $e_i \in (0, \infty)$ is an increment in the $i^{th}$ direction. Let $\{\eta_n\}_{n \in \mathbb{N}}$ be a sequence of decreasing positive real numbers ($\eta_n = o(\lambda_n^{-d})$) such that $\eta_n \to 0$ as $n \to \infty$. We consider the sampling sites $\{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{N_n}\}$ under mixed-increasing domain case are given by the points

$$\{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{N_n}\} \equiv \{\mathbf{s} \in \eta_n \mathcal{Z}^d : \mathbf{s} \in \mathcal{R}_n\}. \tag{1}$$

In (1) the sampling points of the scaled lattice $\eta_n \mathcal{Z}^d$ lie in the sampling region $\mathcal{R}_n$ and with increasing $n$ the lattice becomes finer, inflated (unbounded) and makes the

3

region $\mathcal{R}_n$ dense. The maximum separation distance between two adjacent vertices of a lattice rectangle is bounded by $\eta_n \equiv \max\{e_1, e_2, \cdots, e_d\} \to 0$ as $n \to \infty$. As a result, the sample size $n$ will satisfy the growth condition

$$|\Delta^{-1}\mathcal{R}_0|\left(\frac{\lambda_n}{\eta_n}\right)^d \sim N_n$$

of a larger order than the volume of $\mathcal{R}_n$. Next, we consider an increasing sequence of positive real-numbers $\{\delta_n\}_{n\in\mathbb{N}}$ such that $\delta_n = o(\log n)$ and $\delta_n \to \infty$ as $n \to \infty$. Let $\mathcal{N}_n(\mathbf{s}_i, \delta_n, \eta_n)$ is the neighborhood centered at $\mathbf{s}_i$ with radius $\delta_n$ like [15] which will satisfy

$$\mathcal{N}_n(\mathbf{s}_i, \delta_n, \eta_n) \subset \mathcal{N}_{n+1}(\mathbf{s}_i, \delta_{n+1}, \eta_n).$$

Assume $\{\Gamma_n\}_{n\in\mathbb{N}}$ is an increasing sequence of positive integers satisfying $\Gamma_n = \mathcal{O}(\lambda_n)$ that denotes the cardinality of spatial locations included in the neighborhood. As $n \to \infty$ the lattice spacing $\eta_n \to 0$ and, the cardinality, $\Gamma_n$, is going to $\infty$. Under mixed increasing sampling design the sampling region and neighborhood of every location have two properties: the neighborhoods of every location will become denser which incorporates the "infill" property of the spatial domain and the sampling region will be inflated with increasing $N_n$ which incorporates the "increasing domain" property of spatial domain. In the infill spatial neighborhood, we study the asymptotic behavior of a two-layer deep neural network (DNN) estimator.

### 2.1.2 Spatial DNN

We assume that $\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{R}_n\}$ and $\{\mathcal{X}(\mathbf{s}) \subset \mathbb{R}^p : \mathbf{s} \in \mathcal{R}_n\}$ are $(p+1)$ dimensional random fields (rf) on the bounded region $\mathcal{R}_n \subset \mathbb{R}^d$. Let $Y(\mathbf{s})$ and $\mathcal{X}(s)$ are respectively the spatial surface of response and predictor variables. In the spatial autoregression (SAR) model we consider

$$Y(\mathbf{s}) = \rho W Y + \mathcal{X}(\mathbf{s})\beta + \epsilon. \tag{2}$$

In (2) the dependent variable $Y \equiv Y(\mathbf{s}) \in \mathbb{R}^{N_n}$, the independent variables $\mathcal{X} \equiv \mathcal{X}(\mathbf{s}) \in \mathbb{R}^{N_n \times p}$. The spatial weight matrix $W \in \mathbb{R}^{N_n \times N_n}$ and the error $\epsilon(\cdot)$ is a white noise process with variance $\sigma^2$. In this article, we extend the SAR model in a non-linear framework. We assume there is a nonlinear regressive behavior in $Y$ for every spatial location such that

$$\begin{aligned} Y_t(\mathbf{s}_i) &= f_0\left(Y_t(\mathbf{s}_j), \mathcal{X}_t(\mathbf{s}_i), \mathcal{X}_t(\mathbf{s}_j)\right) + \epsilon_t(\mathbf{s}_i), \\ t &= 1, 2, \cdots, n; \ \mathbf{s}_j \in \mathcal{N}_{\Gamma_n}(\mathbf{s}_i, \delta_n, \eta_n) \equiv \mathcal{N}_n(\mathbf{s}_i, \delta_n, \eta_n) \subset \mathcal{R}_n. \end{aligned} \tag{3}$$

Here in (3), we assume that the unknown function $f_0 \in \mathcal{F}$ such that,

$$\mathcal{F} = \left\{ f_2 \circ f_1 : f_i \in \mathcal{M}_i^{\alpha_i, \tilde{\beta}} \right\}$$

4

where, $\mathcal{M}_i^{\alpha_i,\tilde{\beta}}$ is the isotropic $\tilde{\beta}$ Holder class of function $f_i$[1]. In (3) $Y_t(\mathbf{s}_i), \mathcal{X}_t(\mathbf{s}_i)$ are $n$ independent and identically distributed (iid) copies of $Y(\mathbf{s}_i) \in \mathbb{R}^{N_n}$ and $\mathcal{X}(\mathbf{s}_i) \in \mathbb{R}^{N_n \times p}$. Under infill asymptotic of fixed sampling design the nonlinear unknown function $f_0$ will minimize the population criterion function

$$
\begin{aligned}
\mathcal{L}_n(f) &= \left[ \frac{1}{n} \sum_{t=1}^{n} \{ Y_t(\mathbf{s}_i) - f(Y_t(\mathcal{N}_{\Gamma_n}(\mathbf{s}_i, \delta_n, \eta_n)), \mathcal{X}_t(\mathcal{N}_{\Gamma_n}(\mathbf{s}_i, \delta_n, \eta_n))) \}^2 \right] \\
&= \frac{1}{n} \sum_{t=1}^{n} \{ f(Y_t(\mathcal{N}_{\Gamma_n}(\mathbf{s}_i, \delta_n, \eta_n)), \mathcal{X}_t(\mathcal{N}_n(\mathbf{s}_i, \delta_n))) \\
&\quad - f_0(Y_t(\mathcal{N}_{\Gamma_n}(\mathbf{s}_i, \delta_n, \eta_n)), \mathcal{X}_t(\mathcal{N}_{\Gamma_n}(\mathbf{s}_i, \delta_n, \eta_n))) \}^2 + \sigma^2.
\end{aligned}
\tag{4}
$$

We can obtain a least square estimator of the regression function (4) by minimizing the empirical squared error loss $\mathbb{L}_n(f)$:

$$
\begin{aligned}
\hat{f}_n(\mathbf{s}_i) &= \mathrm{argmin}_{f \in \mathcal{F}} \mathbb{L}_n(f) \\
&= \mathrm{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \{ Y_t(\mathbf{s}_i) - f(Y_t(\mathcal{N}_{\Gamma_n}(\mathbf{s}_i, \delta_n, \eta_n)), \mathcal{X}_t(\mathcal{N}_{\Gamma_n}(\mathbf{s}_i, \delta_n, \eta_n))) \}^2 .
\end{aligned}
\tag{5}
$$

The least-square estimator in (5) over universal $\mathcal{F}$ may face some undesired properties for example consistency. Therefore, optimizing the squared error loss over some less complex functional space $\mathcal{F}_n$ will be better.

**Assumption 2.1.** *The functional space $\mathcal{F}_n$ with VC dimension $\nu$ and the functionals belonging to this space are square-integrable (one choice might be Reproducing Kernel Hilbert Space).*

Here the function space $\mathcal{F}_n$ in Assumption. 2.1 is an approximation of $\mathcal{F}$ and this approximation error tends to 0 with increasing sample size in $\mathcal{N}_{\Gamma_n}(\mathbf{s}_i, \delta_n, \eta_n)$. We consider a sequence of non-decreasing classes of functions

$$
\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \cdots \subset \mathcal{F}
$$

where $\bigcup_{n \geq 1} \mathcal{F}_n$ is dense in $\mathcal{F}$. Moreover for any $f \in \mathcal{F}$ there exists a projection $\pi_n f \in \mathcal{F}_n$ such that some pseudometric defined in $\mathcal{F}$ i.e. $d(f, \pi_n f) \to 0$ where as $n \to \infty$. This type of function is defined as a 2-DNN according to [16]. The DNN estimator $\hat{f}_n$ would satisfy

$$
\mathbb{L}_n(\hat{f}_n) \leq \mathrm{Inf}_{f \in \mathcal{F}_n} \mathbb{L}_n(f) + \mathcal{O}_P(\Xi_n)
$$

---

[1]A function $f_i : \mathbb{R}^d \to \mathbb{R}$ will be from isotropic $\tilde{\beta}$ Holder class $\mathcal{M}_i^{\alpha_i,\tilde{\beta}}$ if there exists a constant $\mathcal{M}_i > 1$; $\alpha_i \in (0,1)$ such that
$$
|D^{\tilde{\beta}} f_i(\mathbf{x}) - D^{\tilde{\beta}} f_i(\mathbf{y})| \leq M_n ||\mathbf{x} - \mathbf{y}||_2^{\alpha_i}
$$
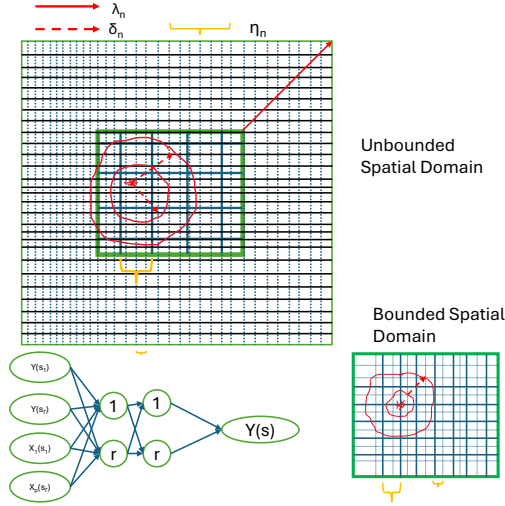where, $D^{\tilde{\beta}} f(\mathbf{x})$ is the $\tilde{\beta}$-th order partial derivative of $f_i(\mathbf{x})$.

**Fig. 1**: Flowchart of Localized DNN Estimator

with $lim_{n\to\infty}\Xi_n = 0$. We will describe the localized DNN with two hidden layers and a `tanh` activation function belonging to the isotropic holder class[1]. We assume a two-layer DNN with $r$ hidden nodes:

$$
\mathcal{F}_{r,n} \equiv \left\{ f_2 \circ f_1 : f_1 \in \mathcal{F}_r^{(1)} \subset \mathcal{M}_1^{\alpha_1,\tilde{\beta}} \ \& \ f_2 \in \mathcal{F}_r^{(2)} \subset \mathcal{M}_2^{\alpha_2,\tilde{\beta}} \right\}
$$

$$
\equiv \left\{ \vartheta_0 + \sum_{i=1}^r \vartheta_i \mathtt{tanh}\left( \sum_{j=1}^r \nu_{ij} \mathtt{tanh}\left( \sum_{k=1}^p \sum_{l=1}^{\Gamma_n} \upsilon_{ijkl} X_k(\mathbf{s}_{\sigma_l}) + \right.\right.\right.
$$

$$
\left.\left. \sum_{l=1}^{\Gamma_n} \upsilon_{ij(p+1)l} Y(\mathbf{s}_{\sigma_l}) + \upsilon_{ij00} \right) + \nu_{i0} \right) : \tag{6}
$$

$$
\sum_{i=0}^r |\vartheta_i| \le V_2; \sum_{j=0}^r |\nu_{ij}| \le V_2 \max_{\substack{1 \le i \le r \\ 1 \le j \le r}} \sum_{k=0}^{p+1} \sum_{l=1}^{\Gamma_n} |\upsilon_{ijkl}| \le V_1 \text{for some } V_1 > 1 \right\};
$$

for some $\min\{V_1, V_2\} > 1; \vartheta_i, \nu_{ij}, \upsilon_{ijkl} \in \mathbb{R}$;

$i, j = 0, \cdots, r; k = 0, \cdots, (p+1); \ l = 1, 2, \cdots, \Gamma_n; \ \sigma_l \in \{1, 2, \cdots, n\}$.

Assume $||Y_i||_\infty \le \min\{V_1, V_2\} \Rightarrow ||f_0||_\infty \le \min\{V_1, V_2\}$. In (6) we assume $\{\mathbf{s}_{\sigma_1}, \mathbf{s}_{\sigma_2}, \cdots, \mathbf{s}_{\sigma_{\Gamma_n}}\} \in \mathcal{N}_{\Gamma_n}(\mathbf{s}_i, \delta_n, \eta_n)$. In Figure 1 we have illustrated the workflow of a two-layer DNN-based spatial regression model. For unbounded regions corresponding to every target location we create a spatial neighborhood of radius $\delta_n$ and consider those locations included in this spatial neighborhood. We feed the response and covariates of those spatial locations in the input node of a DNN having two hidden layers. Here to avoid high-dimensional problems we also give one constraint that

6

$(p + 1)\Gamma_n = o(\max\{N_n, n\})$. This restriction is crucial for bounded and unbounded spatial domains. However, for the bounded domain case, it will be equivalent to the purely "infill" spatial subsampling. In Algorithm (1) we have explained the algorithm 2-DNN based spatial regression model. From this Algorithm (1) it's clearly understandable that this model is capturing spatial dynamicity at the local level by finer grid spacing concept of multi-resolution and overall dynamicity by the sieve class of nonlinear functionals.

---

**Algorithm 1** 2-DNN Spatial Auto Regression

---

1: Spatial sampling region, $R_n = \lambda_n \cdot R_0$.
2: Initial Grid Spacing in Lattice, $\eta_n \leftarrow \frac{1}{\lambda_n + d}$.
3: **for** l in 1:L **do**
4:     Making finer grid $\eta_{nl} \leftarrow \frac{\eta_n}{2^l}$.
5:     **for** each $\delta_n$ **do**
6:         Create Neighborhood $\mathcal{N}_{\Gamma_n}(\mathbf{s}_i, \delta_n, \eta_n)$ and store $Y, \mathcal{X}$ in $\mathcal{Y}_t(\mathbf{s}_i, \mathcal{N}_{\Gamma_n})$.
7:         **for** i in $1 : N_n$ **do**
8:             Define 2-DNN spatial regression model with `tanh` activation:

$$Y_i = f_2(W_2 f_1(W_1 \mathcal{Y}_t(\mathbf{s}_i, \mathcal{N}_{\Gamma_n}) + b_1) + b_2),$$

$$h_i = f(W_i h_{i-1} + b_i), \ \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

9:             **for** epoch = 1 to num_epochs **do**
10:                Shuffle the training data.
11:                **for** b in 1:B **do**
12:                   Compute loss for minibatch.

$$\mathbb{L}_{nb} = \frac{1}{n} \sum_{t=1}^{n} \left\{ Y_t(\mathbf{s}_i) - f(\mathcal{Y}_t(\mathbf{s}_i, \mathcal{N}_{\Gamma_n})) \right\}^2.$$

13:                   Compute gradients for minibatch using backpropagation.

$$\frac{\partial \mathbb{L}_{nb}}{\partial w} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i) \cdot \frac{\partial \hat{Y}_i}{\partial w}$$

14:                   Update model parameters using Adam optimizer.

$$m_p^* = \beta_1^* m_{p-1}^* + (1 - \beta_1^*) g_p^*; v_p^* = \beta_2^* v_{p-1}^* + (1 - \beta_2^*) g_p^{*2}$$

$$\hat{m}_p^* = \frac{m_p^*}{1 - \beta_1^{*p}}; \hat{v}_p^* = \frac{v_p^*}{1 - \beta_2^{*p}}; \theta_{p+1}^* = \theta_p^* - \frac{\alpha^* \hat{m}_p^*}{\sqrt{\hat{v}_p^*} + \epsilon^*}$$

15:                   Update learning rate according to necessity.
16:                **end for**
17:                Evaluate the model's performance on the validation data set.
18:             **end for**
19:         **end for**
20:     **end for**
21: **end for**

---

## 2.2 Asymptotic Analysis

This section is devoted to discussing the asymptotic behavior of a two-layer DNN (2-DNN) estimator. This section is divided into two parts. The first part will study the existence, consistency, and convergence rate of a 2-DNN.

### 2.2.1 Asymptotic Convergence

The first theorem will prove the existence of this 2-DNN estimator. This theorem is motivated by Theorem 2.2 of [17] and an extension of Corollary 2.1 of [2]. Here the consistency of 2-DNN is different from [2]'s consistency of single-layer sieve neural network (SNN) because [2] has achieved the consistency for single-layer DNN with `sigmoid` activation function. A single-layer DNN is not very effective in dealing with complex spatial problems. But here we propose the result for 2-DNN which is deep and helpful for real data. Here the rate of increase of input nodes and nodes in the hidden layer is slower than the rate of inflation, $\lambda_n$ of the spatial sampling region.

**Theorem 1.** *[Existence] Let $(\Omega, \mathcal{A}, P)$ be a complete probability space and let $(\mathcal{F}, \rho)$ be a pseudo-metric. Let $\{\mathcal{F}_{r,n}\}$ be a sequence of compact subsets of $\mathcal{F}$. Let $\mathbb{L}_n : \Omega \times \mathcal{F}_{r,n} \to \bar{\mathbb{R}}$ be $\mathcal{A} \otimes \mathcal{B}(\mathcal{F}_{r,n})/\mathcal{B}(\bar{\mathbb{R}})$ measurable and suppose for every $\omega \in \Omega$ if $\mathbb{L}_n(\omega, \cdot)$ is lower semicontinuous on $\Theta_n$. Then for every $n = 1, 2, \cdots$ there exists a 2-DNN estimator $\hat{f}_n : \Omega \to \mathcal{F}_{r,n}$ which is $\mathcal{A}/\mathcal{B}(\mathcal{F}_{r,n})$ measurable such that $\mathbb{L}_n(\hat{f}_n(\omega)) = Inf_{f \in \mathcal{F}_{r,n}} \mathbb{L}(f)$.*

*Proof.* Pleas see the proof in Appendix (A) □

Theorem 1 identifies the condition of existence of 2-DNN. The following theorem will justify the convergence of this two-layer DNN.

**Theorem 2.** *[Convergence]*

A2.1 *If $\Gamma_n = o\left(n^\psi\right), \lambda_n = o(n), r = o\left(n^\beta\right), V_2\sqrt{\sigma^2 + 1} << r^2, V_1 V_2 = o(\sqrt{\log n})$ for some $0 < \psi + \beta < 1$.*

A2.2 *The functional class $\mathcal{F}_{r,n}$ with Vapnik-Chervonenkis (VC) dimension $\nu$ with bounded envelope function $\tilde{g}$. This bounded envelop function, $\tilde{g}$ is uniformly bounded by $min\{V_1, V_2\}$.*

A2.3 *The Covering number that grows at a controlled rate such that*

$$\log \mathcal{N}\left(\varphi, \mathcal{F}_{r,n}, ||\cdot||\right) \leq \nu C \varphi^{-d},$$

*for some $C > 0$ then,*

$$||\hat{f}_n - f_0||_n \to^P 0 \ \ as \ \ n \to \infty$$

*Proof.* Please see the proof in Appendix (B) □

Theorem 2 is the first theorem that has established the convergence of two-layer DNN after [3]. But there are two key differences one is we extend DNN in spatial design and the second is our covering number's growth rate is restricted. The next theorem will comment on the convergence rate of this DNN model.

**Theorem 3.** *If (A2.1)-(A2.3) from Theorem 2 hold and,*

$$\zeta_n = \mathcal{O}_P \left( min \left\{ ||\pi_{r,n}f_0 - f_0||_n^2, \right. \right.$$
$$\frac{\left(r^2 + r((p+1)\Gamma_n + 4) + 2\right) \log \left[\left(r^2 + r((p+1)\Gamma_n + 4) + 2\right)(V_1V_2)^2\right]}{n},$$
$$\left. \left. \frac{\left(r^2 + r((p+1)\Gamma_n + 4) + 2\right) \log(n \log n)}{n} \right\} \right) \tag{7}$$

*then,*

$$||\hat{f}_n - f_0||_n = \mathcal{O}_P \left\{ max \left[ ||\pi_{r,n}f_0 - f_0||_n, \left( \frac{n}{n^{\psi+\beta} \log(n^{\psi+\beta} \log n)} \right)^{-1/2}, \right. \right.$$
$$\left. \left. \left( \frac{n}{n^{\psi+\beta} \log(n \log n)} \right)^{-1/2} \right] \right\}.$$

*Proof.* Please see the proof in Appendix (C)                                      □

In Theorem 3 we have proved the convergence rate of 2-DNN for every spatial location with its increasing neighborhood ($\mathcal{N}_{\Gamma_n}(\mathbf{s}_i, \delta_n, \eta_n)$) size, $\Gamma_n$. This convergence rate is faster than the first convergence rate of spatial DNN of [1].

## 2.3 Uncertainity Quantification

This section discusses the subsampling-based confidence interval of the 2-DNN functional of a specified location. We have already observed that for every location 2-DNN becomes a random vector. Therefore here our main goal is to find $\mathcal{C}_\alpha = [L_\alpha, U_\alpha]$. Let's consider $\mathcal{Y}_{\Gamma_j} = y(\mathcal{N}_{\Gamma_j}(\mathbf{s}_i, \delta_n, \eta_n))$ for zero means and for nontrivial mean function $\mathcal{Y}_{\Gamma_j} = \left( y(\mathcal{N}_{\Gamma_j}(\mathbf{s}_i, \delta_j)), \mathcal{X}(\mathcal{N}_{\Gamma_j}(\mathbf{s}_i, \delta_j)) \right)$ such that

$$P(f_0(\mathcal{Y}_{\Gamma_j}) \in \mathcal{C}_\alpha) = 1 - \alpha.$$

To find this $\mathcal{C}_\alpha$ we assume that

$$\text{Var}\left( n^\lambda \hat{f}_n(\mathcal{Y}_{\Gamma_j}) \right) \to 0; \text{ as } n \to \infty, \text{ for some } \lambda > 0.$$

We take a subsample of spatial sampling regions at each iteration of size $\Gamma_j$ with corresponding radius $\delta_j$ where $\Gamma_j < \Gamma_{j+1} << |\mathcal{R}_n| \equiv N_n$. Its corresponding neighborhood size, $\delta_j > \delta_{j-1} > \eta_n \ \forall \ j = 1, 2, \cdots, B$. From each sampling block, we fed the sampling points included in this block and got an estimate of the 2-DNN functional corresponding to the location for each $\Gamma_j$ i.e. $\hat{f}_{n,j}$ (please see Figure 1). In this manner take $B$ sub-samples and got the estimate $\hat{f}_{n1}(\mathbf{s}_i), \hat{f}_{n2}(\mathbf{s}_i), \cdots, \hat{f}_{nB}(\mathbf{s}_i)$. Then the corresponding confidence interval will be

$$\mathcal{C}_\alpha(\mathbf{s}_i) \equiv \bar{\hat{f}}_n(\mathbf{s}_i) \pm z_{1-\alpha/2}\Lambda(\mathbf{s}_i). \tag{8}$$

In (8) the first term of RHS is $\bar{\hat{f}}_n(\mathbf{s}_i) \equiv \frac{1}{B}\sum_{j=1}^{B}\hat{f}_{nj}(\mathbf{s}_i)$ and

$$\Lambda(\mathbf{s}_i) \equiv \sqrt{\frac{1}{B}\sum_{j=1}^{B}\left(y(\mathbf{s}_i) - \hat{f}_{nj}(\mathbf{s}_i)\right)^2}.$$

This confidence interval in (8) briefs the idea of how the 2-DNN varies w.r.t the sample size included in the neighborhood and the inherited spatial dynamicity in the DNN functional estimate. Next, we will discuss the asymptotic behavior of KL divergence of the empirical distribution of observed and predicted responses with increasing subsample sizes. In (3) we have already described that

$$y_t(\mathbf{s}_i) = f_0\left(y_t(\mathbf{s}_{i_1}), \cdots, y_t(\mathbf{s}_{i_m})\right) + \epsilon_t, \ t = 1, 2, \cdots, n.$$

Here $\{y_1(\mathbf{s}_i), y_2(\mathbf{s}_i), \cdots, y_n(\mathbf{s}_i)\}$ are iid copies of $y(\mathbf{s}_i)$. According to (5) we know that

$$\hat{f}_{n,m} = \text{argmin}_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\left(y_t(\mathbf{s}_i) - f\left(y_t(\mathbf{s}_{i_1}), \cdots, y_t(\mathbf{s}_{i_m})\right)\right)^2$$

and $m(n) \equiv m \equiv \Gamma_n \equiv \mathcal{O}\left(n^{\psi}\right)$ for some $0 < \psi < 1$ mentioned in (A2.1) of Theorem 2. As a consequence of Theorem 2, it's clear that $\hat{f}_{nm}$ is consistent. In the next Corollary, we will discuss the asymptotic convergence of KL divergence of the empirical distribution of observed and predicted spatial surfaces. Assume the empirical distribution of $y(\mathbf{s}_i)$ and $\hat{y}(\mathbf{s}_i)$ are respectively

$$\tilde{p}_{n,y(\mathbf{s}_i)}(y) = \frac{1}{n}\sum_{t=1}^{n}I\left(y_t(\mathbf{s}_i) \leq y\right),$$

$$\hat{\tilde{p}}_{n,m,\hat{y}(\mathbf{s}_i)}(y) = \frac{1}{n}\sum_{t=1}^{n}I\left(\hat{y}_{t,m}(\mathbf{s}_i) \leq y\right).$$

**Corollary 2.1.** *If $m(n) = n^{\beta}$ for some $\beta \in (0,1)$ and Theorem 2 is valid then*

$$KL\left(\tilde{p}_{n,y(\mathbf{s}_i)}(y) \,||\, \hat{\tilde{p}}_{n,m,\hat{y}(\mathbf{s}_i)}(y)\right) \to 0, \ \text{whenever } m(n) \to \infty.$$

**Proof 2.1.** *Under the consideration of Theorem 2 and some regularity conditions mentioned in Theorem 2*

$$\tilde{p}_{n,y(\mathbf{s}_i)}(y) \to p_{\infty}\left(y; \theta_{\infty}\right), \ \text{as } n \to \infty.$$

*WLG assume $\exists$ a random set $\mathcal{Q}_m = \{m : |\hat{y}_{t,m}(\mathbf{s}_i) - y(\mathbf{s}_i)| \geq \varepsilon \text{ for some } \varepsilon > 0\}$ with measure $P$ such that*

$$P(\mathcal{Q}_m) \to 0 \ \text{as } m \to \infty.$$

*Then from Theorem 3 we can say*

$$\hat{\tilde{p}}_{n,m,\hat{y}(\mathbf{s}_i)}(y|\mathcal{Q}_m) \rightarrow p_\infty(y;\theta_\infty), \ as \ n \rightarrow \infty.$$

*It suffices the proof.*

In this article, we empirically validate the KL divergence will asymptotically converge to zero with increasing $\delta_n$.

# 3 Results and Discussion

## 3.1 Consistency for Unbounded Spatial Domain: Simulation Study

We have simulated lattice data from the Matern covariance function [15] to observe the asymptotic convergence of 2-DNN in the mixed increasing domain setup. The well-known Matern covariance function is

$$\mathcal{C}_\nu(h) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \sqrt{2\nu} \frac{h}{\phi} \right)^\nu \mathcal{K}_\nu \left( \sqrt{2\nu} \frac{h}{\phi} \right).$$

In this Matern covariance function $\sigma^2, \phi, \kappa$ denote sill, range, and smoothing parameters respectively. The parameter space is $\{(\nu, \phi, \sigma^2) : \nu > 0, \sigma^2 > 0, \phi > 0\}$. We are considering $\eta_n = \frac{1}{\lambda_n + d}$ and assume $\Gamma_n = n^\psi$ with $\psi \in (0,1)$. We study our research outcomes for different radius $\delta_n = 0.3, 0.6, 0.8$ of neighborhood $\mathcal{N}_{\Gamma_n}(\mathbf{s}_i; \delta_n, \eta_n)$. Assume $N_n$ is the number of lattice data points and the 10 covariate values such as $X_1, X_2, \cdots, X_{10}$, and $Y = \sum_{i=1}^{10} i X_i$. This neural network will train 70% graphs as train, 20% as validation, and 10% as test set. In Table (1) we have empirically validated the asymptotic convergence of neural network for different parameter combinations of smoothing parameters with increasing neighborhood size in fixed sampling design under mixed increasing domain framework in four scenarios. We have observed the mean square prediction error (MSPE) relatively becomes smaller with increasing $\delta_n$ and decreasing $\eta_n$ from Table (1). In Figure 2 we demonstrate the 95% CI of localized 2-DNN functional for a particular spatial location in the lattice. From Figure 2a it's visible that the CI width becomes narrower with increasing sample size and a similar pattern is detectable for Figure 2b, Figure 2c, and Figure 2d. Likewise, empirical CDF of spatial 2-DNN functional of predicted data converges rapidly with increasing $\delta_n$ since $KL\left(\tilde{p}_{n,y(\mathbf{s}_i)}(y) \,\|\, \hat{\tilde{p}}_{n,m,\hat{y}(\mathbf{s}_i)}(y)\right) \rightarrow 0$ with increasing $m, n$ from Figure 3. In Figure 3c, Figure 3d we detect that the convergence rate is faster than the Figure 3a and Figure 3b with increasing smoothness parameter $\kappa$. From this entire discussion, we can get a brief scenario of the convergence of localized 2-DNN. From Figure 4, we detect that under the assumption of increasing neighborhood size of $o(n^\psi)$ the localized 2-DNN functional converges rapidly with decreasing smoothing parameter $\kappa$ of Matern variogram.

**Table 1**: Simulated Lattice Data from Matern Variogram

| Scenario 1 | $\sigma^2 = 1, \phi = 0.1, \kappa = 0.5$ | |
| --- | --- | --- |
| | $\delta_n$ | MSPE |
| $\lambda_n = 4, n = 20, \eta_n = 0.16$ | 0.3 | 0.046 |
| | 0.6 | 0.0035 |
| | 0.8 | 0.00137 |
| $\lambda_n = 5, n = 35, \eta_n = 0.14$ | 0.3 | 0.042 |
| | 0.6 | 0.0039 |
| | 0.8 | 0.00138 |
| $\lambda_n = 6, n = 50, \eta_n = 0.12$ | 0.3 | 0.04 |
| | 0.6 | 0.0065 |
| | 0.8 | 0.00131 |
| **Scenario 2** | $\sigma^2 = 1, \phi = 0.1, \kappa = 1$ | |
| | $\delta_n$ | MSPE |
| $\lambda_n = 4, n = 20, \eta_n = 0.16$ | 0.3 | 0.05 |
| | 0.6 | 0.0047 |
| | 0.8 | 0.0007 |
| $\lambda_n = 5, n = 35, \eta_n = 0.14$ | 0.3 | 0.047 |
| | 0.6 | 0.0045 |
| | 0.8 | 0.00097 |
| $\lambda_n = 6, n = 50, \eta_n = 0.12$ | 0.3 | 0.04 |
| | 0.6 | 0.0063 |
| | 0.8 | 0.0012 |
| **Scenario:3** | $\sigma^2 = 1, \phi = 0.1, \kappa = 1.5$ | |
| | $\delta_n$ | MSPE |
| $\lambda_n = 4, n = 20, \eta_n = 0.16$ | 0.3 | 0.053 |
| | 0.6 | 0.0059 |
| | 0.8 | 0.00153 |
| $\lambda_n = 5, n = 35, \eta_n = 0.14$ | 0.3 | 0.048 |
| | 0.6 | 0.006 |
| | 0.8 | 0.0024 |
| $\lambda_n = 6, n = 50, \eta_n = 0.12$ | 0.3 | 0.036 |
| | 0.6 | 0.0063 |
| | 0.8 | 0.00125 |
| **Scenario:4** | $\sigma^2 = 1, \phi = 0.1, \kappa = 2$ | |
| | $\delta_n$ | MSPE |
| $\lambda_n = 4, n = 20, \eta_n = 0.16$ | 0.3 | 0.047 |
| | 0.6 | 0.00629 |
| | 0.8 | 0.00113 |
| $\lambda_n = 5, n = 35, \eta_n = 0.14$ | 0.3 | 0.047 |
| | 0.6 | 0.0051 |
| | 0.8 | 0.0022 |
| $\lambda_n = 6, n = 50, \eta_n = 0.12$ | 0.3 | 0.04 |
| | 0.6 | 0.0064 |
| | 0.8 | 0.00134 |
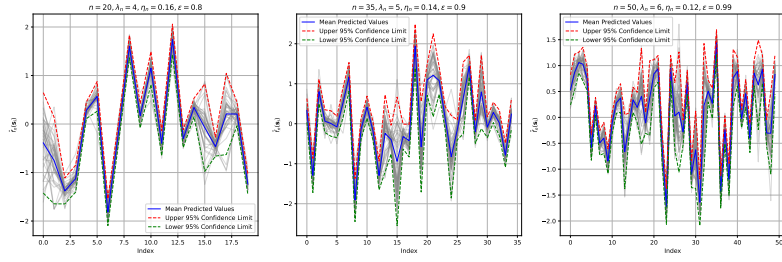
(a) $(\sigma^2, \phi, \kappa) = (1, 0.1, 0.5)$ for n = 20 $(left)$, 35 $(middle)$, 50 $(right)$.
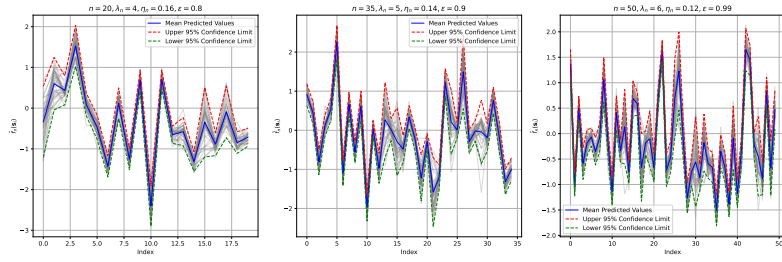


(b) $(\sigma^2, \phi, \kappa) = (1, 0.1, 1)$ for n = 20 $(left)$, 35 $(middle)$, 50 $(right)$.
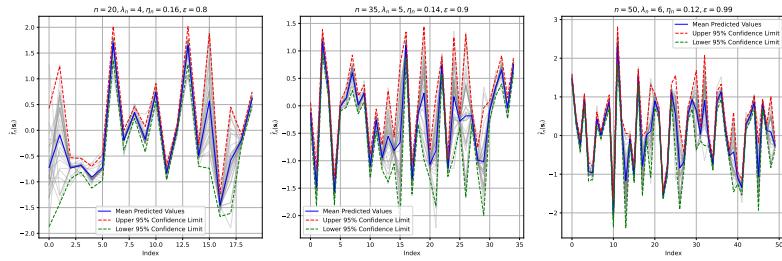


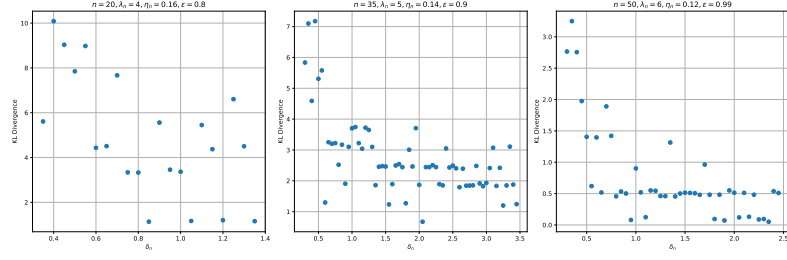(c) $(\sigma^2, \phi, \kappa) = (1, 0.1, 1.5)$ for n = 20 $(left)$, 35 $(middle)$, 50 $(right)$.
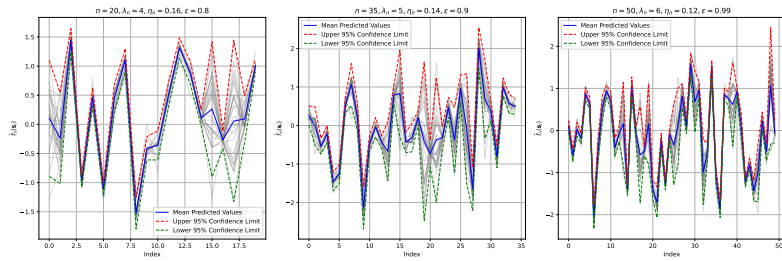


(d) $(\sigma^2, \phi, \kappa) = (1, 0.1, 2)$ for n = 20 $(left)$, 35 $(middle)$, 50 $(right)$.
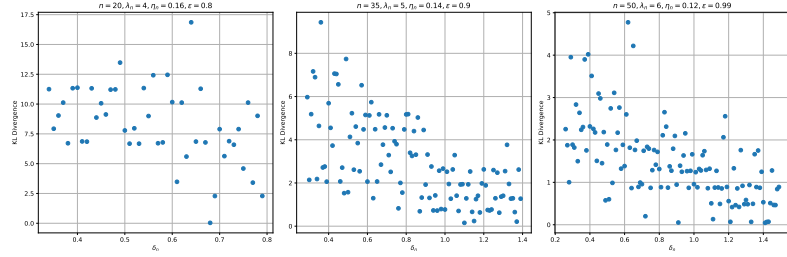
**Fig. 2**: 95% CI of localized 2-DNN functional.
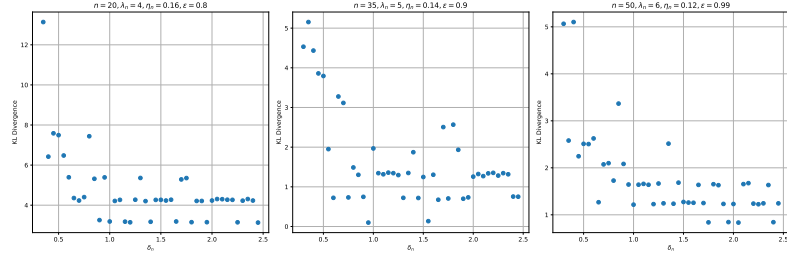
13

(a) $(\sigma^2, \phi, \kappa) = (1, 0.1, 0.5)$ for n $= 20$ $(left), 35$ $(middle), 50$ $(right)$.



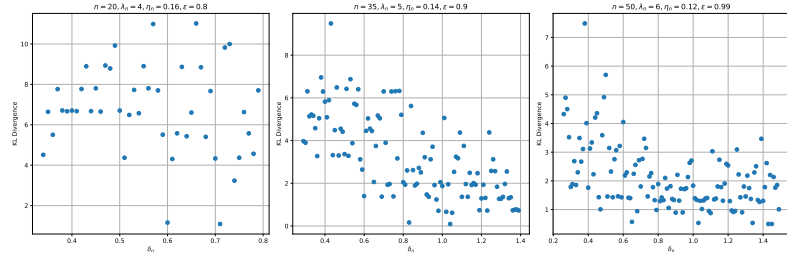(b) $(\sigma^2, \phi, \kappa) = (1, 0.1, 1)$ for n $= 20$ $(left), 35$ $(middle), 50$ $(right)$.



(c) $(\sigma^2, \phi, \kappa) = (1, 0.1, 1.5)$ for n $= 20$ $(left), 35$ $(middle), 50$ $(right)$.



(d) $(\sigma^2, \phi, \kappa) = (1, 0.1, 2)$ for n $= 20$ $(left), 35$ $(middle), 50$ $(right)$.

**Fig. 3**: Asymptotic convergence of KL divergence with increasing $\delta_n$.
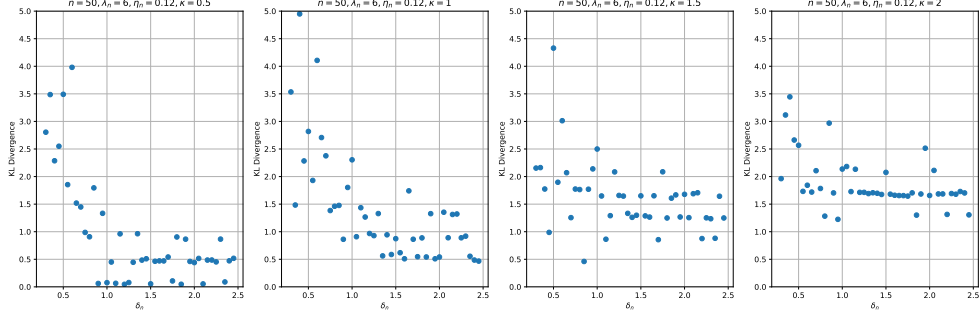
**Fig. 4**: Variation of KL Divergence with $\kappa$ values.

## 3.2 Consistency for Bounded Spatial Domain: Satellite Image

In this section, we discuss the consistency behavior, where we study whether the 2-DNN is consistent. We have already empirically validated that the consistency of localized 2-DNN is achieved under increasing $\delta_n, \lambda_n$ and decreasing $\eta_n$. But if we do $\lambda_n$ is fixed then with increasing density in the neighborhood we achieve the convergence of localized 2-DNN functional. We collect the monthly average surface temperature data for major cities[3] in the United States of America (USA) and satellite images from the website[4] where we collect monthly average surface temperature data (SKT)[4], monthly average of air temperature at 2m above from the surface (t2m)[4], and that of 2m above dew point temperature (d2m)[2].
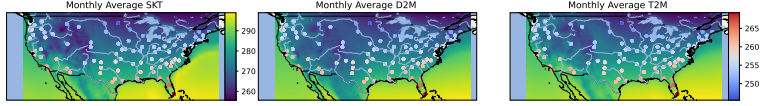


**Fig. 5**: Satellite images with pixel spacing, $\eta_n = 0.25$ and the dotted points are indicating monthly (January) average of 1982 of SKT of cities in USA.

---

[3]https://cds.climate.copernicus.eu
[4]https://power.larc.nasa.gov/data-access-viewer/

Now for each city in Figure 5, we create a neighborhood of $\delta_n$ and consider those spatial pixels included in this neighborhood. We assume monthly average surface temperature as $y_{city}(\mathbf{s}_i)$ and that of neighborhood pixels are $\{y_{im}(\mathbf{s}_{i_1}), y_{im}(\mathbf{s}_{i_2}), \cdots, y_{im}(\mathbf{s}_{i_m})$. Consider t2m as $\mathcal{X}_1$ and d2m as $\mathcal{X}_2$. Also, assume the covariates in those locations included in the neighborhood. Here the initial grid spacing $\eta_0 = 0.25$ and in each resolution to get finer resolution at $l^{th}$ level we consider $\eta_l = \eta_0/2^{l-1}$. Thus the grid spacing becomes a decreasing sequence and sampling region $\mathcal{R}_n$ becomes heavily dense. We consider the neural network as

$$y_{city}(\mathbf{s}_i) = f_{DNN} \{y_{im}(\mathbf{s}_{i_1}), y_{im}(\mathbf{s}_{i_2}), \cdots, y_{im}(\mathbf{s}_{i_m})$$
$$\mathcal{X}_{1,im}(\mathbf{s}_{i_1}), \mathcal{X}_{1,im}(\mathbf{s}_{i_2}), \cdots, \mathcal{X}_{1,im}(\mathbf{s}_{i_m})$$
$$\mathcal{X}_{2,im}(\mathbf{s}_{i_1}), \mathcal{X}_{2,im}(\mathbf{s}_{i_2}), \cdots, \mathcal{X}_{2,im}(\mathbf{s}_{i_m})\} + \epsilon_i.$$

With increasing $\delta_n$ we increase the neighborhood size $m$. To avoid a high-dimensional setup we restrict our number of covariates $(p+1)\Gamma_n = 3m << N_n$. The number of times stands from January 1982 to December 2022. In Figure 5 we depict the monthly average of January 1982 of major cities in US[4] and the corresponding satellite image[3]. We use a two-layer deep neural network with `tanh` activation function in this setup. From Figure 6 we can visualize that root mean square error (RMSE) is decreasing with increasing $\delta_n$. We will discuss one important pattern that is detectable from this Figure 6 that for $\eta_n = 0.25, \delta_n \geq 0.48$ the RMSE $\leq 9.8$, for $\eta_n = 0.125, \delta_n \geq 0.3$ the RMSE $\leq 9.8$, for $\eta_n = 0.06, \delta_n \geq 0.25$ the RMSE $\leq 9.8$, and for $\eta_n = 0.03, \delta_n \geq 0.18$ the RMSE $\leq 9.8$. Thus we state that whenever we decrease $\eta_n$ our $\Gamma_n \equiv m(n)$ becomes large so that for smaller $\delta_n$ we can achieve the desired accuracy.
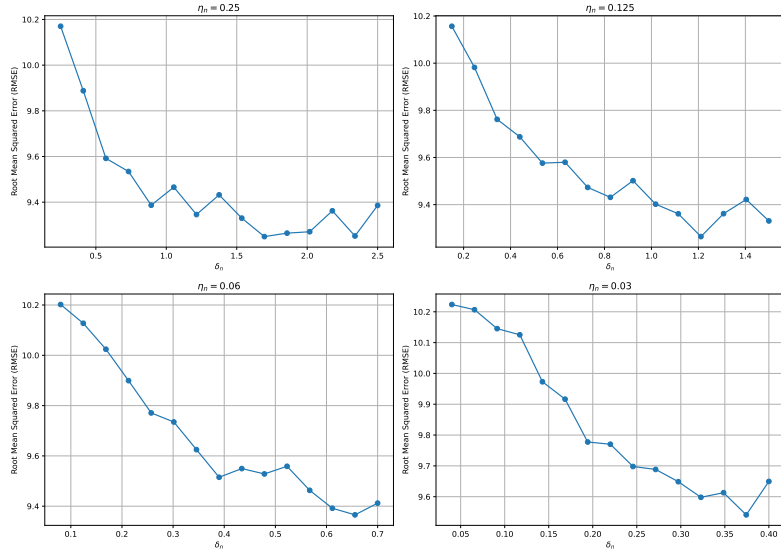


**Fig. 6**: Consistancy of Localized 2-DNN Estimator.

# 4 Conclusion

From this entire discussion, we have empirically and theoretically proved that localized 2-DNN estimator is consistent with decreasing $\eta_n$, increasing $\Gamma_n$ (for higher resolution) for bounded ($\lambda_n$ is fixed) and unbounded ($\lambda_n = o(n)$) for fixed sampling design of lattice data under mixed increasing spatial sampling. This localized 2-DNN function generalizes spatial regression for the first time. The convergence rate of our localized 2-DNN is faster than the convergence rate of spatial GNN of [1], [2]. We empirically observe the convergence of KL divergence of empirical CDF for observed and predicted data towards zero with decreasing smoothness of the Matern variogram. From the application perspective, we illustrate how, if we collect data from different sources, our localized 2-DNN can iteratively achieve its optimum accuracy level. Although our asymptotic analysis is providing an effective outcome for 2-DNN some limitations direct our future research such as we were desperate here to avoid a high-dimensional setup and therefore we always restrict our neighborhood size. In the functional constraint, we ignore the temporal dependence and assume iid input nodes. We do not theoretically justify how variogram parameters are influencing 2-DNN estimates. These problems are still open problems. We will explore these problems in the coming future.

## Declarations

- Conflict of interest/Competing interests: No conflict of interest exists.
- Data availability: Data is open source, and the available URLs are mentioned in the article.
- Code availability: Code is available in https://github.com/debjoythakur/Spatial_subsampling_NN.

## Appendix A    Proof of Theorem 1

*Proof.* We discuss the proof of the existence of 2-DNN in several steps.

Step-1: $\mathbb{L}_n$ is measurable since,

$$\mathbb{L}_n(f) = \frac{1}{n} \sum_{t=1}^{n} \left\{ Y_t(\mathbf{s}_i) - f\left(Y_t(\mathbf{s}_{\sigma_1}), Y_t(\mathbf{s}_{\sigma_2}), \cdots, Y_t(\mathbf{s}_{\sigma_{\Gamma_n}}), \mathcal{X}_t(\mathbf{s}_{\sigma_1}),\right.\right.$$

$$\left.\left.\mathcal{X}_t(\mathbf{s}_{\sigma_2}), \cdots, \mathcal{X}_t(\mathbf{s}_{\sigma_{\Gamma_n}})\right)\right\}^2$$

$$= \frac{1}{n} \sum_{t=1}^{n} \left[ f_0\left(Y_t(\mathbf{s}_{\sigma_1}), Y_t(\mathbf{s}_{\sigma_2}), \cdots, Y_t(\mathbf{s}_{\sigma_{\Gamma_n}}), \mathcal{X}_t(\mathbf{s}_{\sigma_1}), \mathcal{X}_t(\mathbf{s}_{\sigma_2}), \cdots, \mathcal{X}_t(\mathbf{s}_{\sigma_{\Gamma_n}})\right)\right.$$

$$\left. - f\left(Y_t(\mathbf{s}_{\sigma_1}), Y_t(\mathbf{s}_{\sigma_2}), \cdots, Y_t(\mathbf{s}_{\sigma_{\Gamma_n}}), \mathcal{X}_t(\mathbf{s}_{\sigma_1}), \mathcal{X}_t(\mathbf{s}_{\sigma_2}), \cdots, \mathcal{X}_t(\mathbf{s}_{\sigma_{\Gamma_n}})\right) + \epsilon_t\right]^2$$

$$= \frac{2}{n} \sum_{t=1}^{n} \epsilon_t \left(f_{0t} - f_t\right) + \frac{1}{n} \sum_{t=1}^{n} \epsilon_t^2 + \frac{1}{n} \sum_{t=1}^{n} \left(f_{0t} - f_t\right)^2,$$

and $\epsilon$ is measurable.

17

Step-2: For fixed $\omega \in \Omega$ we know that $\mathbb{L}_n(f) \equiv \mathbb{L}_n(f(\omega))$ is continuous in $f$.

Step-3: $\mathcal{F}_{r,n}$ is a compact subset of $\mathcal{F}$

$\Leftrightarrow \exists \quad h : \Theta_n \times \mathcal{X} \to \mathcal{F}_{r,n}$ is continuous since $\Theta_n$ is a compact subset of $\mathbb{R}^{(r^2(d+2)+2r+1)}$ and $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$

$$\Theta_n = \left[ \vartheta_0, \cdots, \vartheta_r, \nu_{00}, \cdots, \nu_{rr}, \upsilon_{1100}, ..., \upsilon_{rr(p+1)\Gamma_n} \right]'$$

$$\in [-V_1, V_1]^{r+1} \times [-V_2, V_2]^{(r+1)^2} \times [-V_1, V_1]^{r^2(p+2)(\Gamma_n+1)}$$

$$\text{where, } h(\theta_n) = \vartheta_0 + \sum_{i=1}^{r} \vartheta_i \tanh\left( \sum_{j=1}^{r} \nu_{ij} \tanh\left( \sum_{k=1}^{p} \sum_{l=1}^{\Gamma_n} \upsilon_{ijkl} X_k(\mathbf{s}_{\sigma_l}) + \right. \right.$$

$$\left. \left. \sum_{l=1}^{\Gamma_n} \upsilon_{ij(p+1)l} Y(\mathbf{s}_{\sigma_l}) + \upsilon_{ij00} \right) + \nu_{i0} \right)$$

$\Leftrightarrow$ It's enough to show for all $\theta_{1n}, \theta_{2n} \in \Theta_n$

$$||h(\theta_{1n}) - h(\theta_{2n})||_n^2 \leq V^* ||\theta_{1n} - \theta_{2n}||_n^2$$

Therefore our main responsibility is to find $V^*$.

For ease of notation assume,

$$\mathcal{Y}_t = \left[ X_{1t}(\mathbf{s}_{\sigma_1}), \cdots, X_{1t}(\mathbf{s}_{\sigma_{\Gamma_n}}), \cdots, X_{pt}(\mathbf{s}_{\sigma_{\Gamma_n}}), Y_t(\mathbf{s}_{\sigma_1}), \cdots, Y_t(\mathbf{s}_{\sigma_{\Gamma_n}}) \right]$$

$$||h(\theta_{1n}) - h(\theta_{2n})||_n^2$$

$$\leq \frac{1}{n} \sum_{t=1}^{n} \left[ \left| \vartheta_0^{(1)} - \vartheta_0^{(2)} \right| + \sum_{i=1}^{r} \left| \vartheta_i^{(1)} \tanh\left( \sum_{j=1}^{r} \nu_{ij}^{(1)} \tanh\left( \sum_{k=1}^{(p+1)\Gamma_n} \upsilon_{ijk}^{(1)} \mathcal{Y}_{k,t} + \upsilon_{ij0}^{(1)} \right) + \nu_{i0}^{(1)} \right) \right. \right.$$

$$\left. \left. - \vartheta_i^{(2)} \tanh\left( \sum_{j=1}^{r} \nu_{ij}^{(2)} \tanh\left( \sum_{k=1}^{(p+1)\Gamma_n} \upsilon_{ijk}^{(2)} \mathcal{Y}_{k,t} + \upsilon_{ij0}^{(2)} \right) + \nu_{i0}^{(2)} \right) \right| \right]^2$$

$$||h(\theta_{1n}) - h(\theta_{2n})||_n^2$$

$$\leq \frac{1}{n} \sum_{t=1}^{n} \left[ \sum_{i=0}^{r} \left| \vartheta_i^{(1)} - \vartheta_i^{(2)} \right| + V_1 \sum_{i=1}^{r} \left| \nu_{i0}^{(1)} - \nu_{i0}^{(2)} \right| \right.$$

$$\left. + V_1 \sum_{i=1}^{r} \sum_{j=1}^{r} \left| \nu_{ij}^{(1)} \mathtt{tanh} \left( \sum_{k=1}^{(p+1)\Gamma_n} \upsilon_{ijk}^{(1)} \mathcal{Y}_{k,t} + \upsilon_{ij0}^{(1)} \right) - \nu_{ij}^{(2)} \mathtt{tanh} \left( \sum_{k=1}^{(p+1)\Gamma_n} \upsilon_{ijk}^{(2)} \mathcal{Y}_{k,t} + \upsilon_{ij0}^{(2)} \right) \right| \right]^2$$

$$\leq \frac{1}{n} \sum_{t=1}^{n} \left[ \sum_{i=0}^{r} \left| \vartheta_i^{(1)} - \vartheta_i^{(2)} \right| + V_1 \sum_{i=1}^{r} \sum_{j=0}^{r} \left| \nu_{ij}^{(1)} - \nu_{ij}^{(2)} \right| \right.$$

$$\left. + V_1 V_2 \sum_{i=1}^{r} \sum_{j=1}^{r} \left| \upsilon_{ij0}^{(1)} - \upsilon_{ij0}^{(2)} \right| + V_1 V_2 \sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{k=1}^{(p+1)\Gamma_n} \left| \upsilon_{ijk}^{(1)} \mathcal{Y}_{k,t} - \upsilon_{ijk}^{(2)} \mathcal{Y}_{k,t} \right| \right]^2$$

$$\leq \frac{1}{n} \sum_{t=1}^{n} \left[ \sum_{i=0}^{r} \left| \vartheta_i^{(1)} - \vartheta_i^{(2)} \right| + V_1 \sum_{i=0}^{r} \sum_{j=0}^{r} \left| \nu_{ij}^{(1)} - \nu_{ij}^{(2)} \right| + V_1 V_2 \sum_{i=0}^{r} \sum_{j=0}^{r} \sum_{k=0}^{(p+1)\Gamma_n} \left| \upsilon_{ijk}^{(1)} - \upsilon_{ijk}^{(2)} \right| \right]^2$$

$$\leq \left[ \sum_{i=0}^{r} \sum_{j=0}^{r} \sum_{k=0}^{(p+1)\Gamma_n} \left| \vartheta_i^{(1)} - \vartheta_i^{(2)} \right| + V_1 \sum_{i=0}^{r} \sum_{j=0}^{r} \sum_{k=0}^{(p+1)\Gamma_n} \left| \nu_{ij}^{(1)} - \nu_{ij}^{(2)} \right| \right.$$

$$\left. + V_1 V_2 \sum_{i=0}^{r} \sum_{j=0}^{r} \sum_{k=0}^{(p+1)\Gamma_n} \left| \upsilon_{ijk}^{(1)} - \upsilon_{ijk}^{(2)} \right| \right]^2$$

$$\leq \left( V_1 V_2 \left\{ (r+1)(p+1)\Gamma_n + (p+1)\Gamma_n \right\} \right)^2 ||\theta_{1n} - \theta_{2n}||_n^2.$$

As a result, we get

$$V^* = \left( V_1 V_2 \left\{ (r+1)(p+1)\Gamma_n + (p+1)\Gamma_n \right\} \right)^2. \tag{A1}$$

$\square$

# Appendix B Proof of Theorem 2

*Proof.* According to Theorem 3.1 in [18] and Corollary 2.6 in [17] it's enough to show that

$$\sup_{f \in \mathcal{F}_{r,n}} |\mathbb{L}_n(f) - \mathcal{L}_n(f)| \to^{P^*} 0 \quad \text{as} \ \ n \to \infty$$

Now we will give the outline of the proof

Step-1: For any $\varepsilon > 0$ we have

$$
\begin{aligned}
P^* \left( \sup_{f \in \mathcal{F}_{r,n}} |\mathbb{L}_n(f) - \mathcal{L}_n(f)| > \frac{\varepsilon}{2} \right) &\leq P \left( \left| \frac{1}{n} \sum_{t=1}^n \varepsilon_t^2 - \sigma^4 \right| > \frac{\varepsilon}{4} \right) \\
&+ P^* \left( \mathrm{Sup}_{f \in \mathcal{F}_{r,n}} \left| \frac{1}{n} \sum_{t=1}^n \epsilon_i \left\{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} \right| > \frac{\varepsilon}{4} \right)
\end{aligned} \tag{B2}
$$

Step-2: The First part will converge to zero using WLLN but for the second part using Markov's Inequality we are required to prove

$$
E_\epsilon \left( \mathrm{Sup}_{f \in \mathcal{F}_{r,n}} \left| \frac{1}{n} \sum_{t=1}^n \epsilon_t \left\{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} \right| \right) \to 0 \text{ as } n \to \infty
$$

Step-3: According to Symmetrisation for $\alpha-$Mixing spatial surface with grid spacing $\eta_n \to 0$ as $n \to \infty$

$$
\begin{aligned}
E_\epsilon \| \mathbb{P}_n - P \|_{\mathcal{F}_{r,n}} = E_\epsilon \left( \mathrm{Sup}_{f \in \mathcal{F}_{r,n}} \left| \frac{1}{n} \sum_{t=1}^n \epsilon_t \left\{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} \right| \right) \\
\leq 2 E_\epsilon E_\rho \mathrm{Sup}_{f \in \mathcal{F}_{r,n}} \left| \frac{1}{n} \sum_{t=1}^n \rho_t \epsilon_t \left\{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} \right|
\end{aligned} \tag{B3}
$$

where $\rho_t$; $t = 1, 2, ..., n$ are Rademacher random variable independent of $\epsilon_1, \epsilon_2, ..., \epsilon_n$. For fixed values of $\epsilon_1, \epsilon_2, ..., \epsilon_n$ (conditionally) $\sum_{t=1}^n \rho_t \epsilon_t \{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \}$ is a sub-gaussian process.

Step-4: According to the definition of covering number and Corollary 2.2.8 of [19] we will get a universal constant $K$. Then using Dudley's integral entropy bound in VC-dimension $\nu$ with any $f_n^* \in \mathcal{F}_{r,n}$ with $n \geq N_1$ we will get

$$
\begin{aligned}
E_\rho \mathrm{Sup}_{f \in \mathcal{F}_{r,n}} & \left| \frac{1}{n} \sum_{t=1}^n \rho_t \epsilon_i \left\{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} \right| \\
& \leq E_\rho \left| \frac{1}{n} \sum_{t=1}^n \rho_t \epsilon_t \left\{ f_n^*(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} \right| \\
& + K \int_0^\infty \sqrt{\frac{\log D(\varphi, \mathcal{F}_{r,n}, ||.||_\infty)}{n}} d\varphi \\
& \leq E_\rho \left| \frac{1}{n} \sum_{t=1}^n \rho_t \epsilon_t \left\{ f_n^*(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} \right| \\
& + K \sqrt{\frac{\nu}{n}} \int_0^{2 V_1 V_2} \sqrt{\frac{\log \mathcal{N}(\varphi/2, \mathcal{F}_{r,n}, ||.||_\infty)}{n}} d\varphi
\end{aligned}
$$

$$\leq E_\rho \left| \frac{1}{n} \sum_{t=1}^n \rho_t \epsilon_t \left\{ f_n^*(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} \right|$$

$$+ K \sqrt{\frac{\nu}{n}} \int_0^{2V_1 V_2} \sqrt{\frac{\log \mathcal{N}\left( \frac{\varphi}{2V_1 V_2 \sqrt{\sigma^2+1}}, \mathcal{F}_{r,n}, ||.||_\infty \right)}{n}} d\xi \tag{B4}$$

$$= I_3 + I_4$$

Step-5: According to [20] we utilize the concept of universal approximation of $I_3$ in (B4) and then it will be

$$2 \, E_\epsilon E_\rho \left| \frac{1}{n} \sum_{t=1}^n \rho_t \epsilon_t \left\{ f_n^*(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} \right|$$

$$\leq 2\sqrt{\sigma^2+1} \times \operatorname{Sup}_{\mathcal{Y}} |f_n^*(\mathcal{Y}_t) - f_0(\mathcal{Y}_t)| \tag{B5}$$

$$\leq 2\sqrt{\sigma^2+1} \times \operatorname{Sup}_{\mathcal{Y}} |f_n^*(\mathcal{Y}_t) - f_0(\mathcal{Y}_t)| \to 0 \text{ as } n \to \infty$$

Step-6: According to [21] from Lemma 14.3, Lemma 14.7

$$\mathcal{N}\left( \frac{\varphi}{2V_1 V_2 \sqrt{\sigma^2+1}}, \mathcal{F}_{r,n}, ||.||_\infty| \right)$$

$$\leq \max_{\mathcal{Y}} \mathcal{N}\left( \frac{\varphi}{V_1 \sqrt{\sigma^2+1}}, \mathcal{F}_{r,n}^{(1)}, ||.||_\infty^\xi \right)$$

$$\times \mathcal{N}\left( \frac{\varphi}{V_1 V_2 \sqrt{\sigma^2+1}}, \mathcal{F}_{r,n}^{(2)}, ||.||_\infty \right)$$

$$\leq \left( \frac{16eV_2\sqrt{\sigma^2+1}}{\varphi} \right)^{r^2} \times \left( \frac{(V_1 V_2)^2 \left\{ r((p+1)\Gamma_n + 4) + 2 \right\}}{2(V_1 V_2 - 1)\varphi} \right)^{r((p+1)\Gamma_n+4)+2}$$

$$\leq \tilde{C}_{r,p,\Gamma_n,V_1,V_2} \, \varphi^{-(r^2+r((p+1)\Gamma_n+4)+2)}$$

$$\Leftrightarrow \log \mathcal{N}\left( \frac{\varphi}{2V_1 V_2 \sqrt{\sigma^2+1}}, \mathcal{F}_{r,n}, ||.||_\infty| \right) \leq r^2 \log \left( \frac{16eV_2\sqrt{\sigma^2+1}}{\varphi} \right)$$

$$+ (r((p+1)\Gamma_n + 4) + 2) \log \left( \frac{(V_1 V_2)^2 (r((p+1)\Gamma_n + 4) + 2)}{2(V_1 V_2 - 1)\varphi} \right)^{r((p+1)\Gamma_n+4)+2}$$

According to Assumption (A2.1) we deduce,

$$\left( 8V_2\sqrt{\sigma^2+1} \right) \left( r(((p+1)\Gamma_n) + 4) + 2 \right) \leq \left( r^2 + r(((p+1)\Gamma_n) + 4) + 2 \right)$$

then,

$$
\log \left( \frac{\left( 8V_2\sqrt{(\sigma^2+1)} \right) (r(((p+1)\Gamma_n)+4)+2)(V_1V_2)^2}{V_1V_2-1} \right)^{\left( r^2+r(((p+1)\Gamma_n)+4)+2 \right)}
$$

$$
+ \left( r^2 + r(((p+1)\Gamma_n)+4)+2 \right) \log(1/\varphi)
$$

$$
\leq \left( r^2 + r(((p+1)\Gamma_n)+4)+2 \right) \log \left( \frac{(r^2+r(((p+1)\Gamma_n)+4)+2)(V_1V_2)^2}{V_1V_2-1} \right)
$$

$$
+ \left( r^2 + r(((p+1)\Gamma_n)+4)+2 \right) \log(1/\varphi)
$$

$$
\leq \left( r^2 + r(((p+1)\Gamma_n)+4)+2 \right) \log \left( \frac{(r^2+r(((p+1)\Gamma_n)+4)+2)(V_1V_2)^2}{V_1V_2-1} \right)
$$

$$
\times \left( 1 + \frac{1}{\varphi} \right).
$$

Consider $C_{r,d,V_1,V_2} = \left( r^2 + r(((p+1)\Gamma_n)+4)+2 \right) \log \left( \frac{(r^2+r(((p+1)\Gamma_n)+4)+2)(V_1V_2)^2}{V_1V_2-1} \right)$.
Hence we have got

$$
\int_0^{2V_1V_2} \sqrt{log\,\mathcal{N}\left( \frac{\varphi}{2V_1V_2\sqrt{\sigma^2+1}}, \mathcal{F}_{r,n}, ||.||_\infty| \right)}
$$

$$
\leq \int_0^{2V_1V_2} C_{r,p,\Gamma_n,V_1,V_2}^{1/2} \left( 1 + \frac{1}{\varphi} \right)^{1/2} d\varphi
$$

$$
\leq 4\sqrt{2} C_{r,p,\Gamma_n,V_1,V_2}^{1/2}
$$

Step-7: Combining the previous all the steps, Step-3, Step-4, Step-5, and Step-6 in Step-2 we will get

$$
E_\epsilon \left( \mathrm{Sup}_{f\in\mathcal{F}_{r,n}} \left| \frac{1}{n}\sum_{t=1}^n \epsilon_t \{f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t)\} \right| \right)
$$

$$
\leq 2E_\epsilon E_\rho \mathrm{Sup}_{f\in\mathcal{F}_{r,n}} \left| \frac{1}{n}\sum_{t=1}^n \rho_t\epsilon_t \{f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t)\} \right|
$$

$$
\leq 2\sqrt{\sigma^2+1} \times \mathrm{Sup}_{\mathcal{Y}} |f_n^*(\mathcal{Y}_t) - f_0(\mathcal{Y}_t)| \tag{B6}
$$

$$
+ 2K\sqrt{\frac{\nu}{n}} \int_0^{2V_1V_2} \sqrt{\frac{log\,\mathcal{N}\left( \frac{\varphi}{2V_1V_2\sqrt{\sigma^2+1}}, \mathcal{F}_{r,n}, ||.||_\infty \right)}{n}} d\varphi
$$

$$
\approx 2\sqrt{\sigma^2+1} \times \mathrm{Sup}_{\mathcal{Y}} |f_n^*(\mathcal{Y}_t) - f_0(\mathcal{Y}_t)| + 8\sqrt{2}K\sqrt{\frac{\nu C_{r,p,\Gamma_n,V_1,V_2}}{n^2}}
$$

$$= 2\sqrt{\sigma^2 + 1} \times \mathrm{Sup}_{\mathcal{Y}} |f_n^*(\mathcal{Y}_t) - f_0(\mathcal{Y}_t)|$$

$$+ 8\sqrt{2\nu}K \sqrt{\frac{\left(r^2 + r(((p+1)\Gamma_n) + 4) + 2\right)(V_1 V_2)^2 \log\left(\frac{(r^2 + r(((p+1)\Gamma_n)+4)+2)(V_1 V_2)^2}{V_1 V_2 - 1}\right)}{n^2}}$$

$$\approx 2\sqrt{\sigma^2 + 1} \times \mathrm{Sup}_{\mathcal{Y}} |f_n^*(\mathcal{Y}_t) - f_0(\mathcal{Y}_t)| + 8\sqrt{2}K\sqrt{\frac{\nu}{n}}$$

$$\sqrt{\frac{\left(r^2 + r(((p+1)\Gamma_n) + 4) + 2\right)(V_1 V_2)^2 \log\left((r^2 + r(((p+1)\Gamma_n) + 4) + 2)(V_1 V_2)\right)}{n}}$$

To prove the convergence it's necessary to satisfy (B6) should converge to 0 as $n \to \infty$. The first part of the addition converges to zero because of the universal approximation theorem. For the second part since we have assumed $\Gamma_n = o(n^\beta)$, $r = o(n^\psi)$ and $\psi + \beta < 1$ then the second part will converge to 0 as $n \to \infty$. $\qquad\square$

# Appendix C    Proof of Theorem 3

*Proof.* This proof will discuss the convergence rate of spatial 2-DNN in detail.

Step-I: According to [21] from Lemma 14.3, Lemma 14.7 of covering number inequality utilizing exponential decay of $\alpha-$Mixing condition will satisfy the following modified inequality

$$\mathcal{N}_\infty\left(\varphi, \mathcal{F}_{r,n}, ||\cdot||_\infty\right) \leq \left(\frac{2e}{\varphi}\right)^{r^2} \times \left(\frac{(V_1 V_2)^2 \left(r((p+1)\Gamma_n + 4) + 2\right)}{\varphi(V_1 V_2 - 1)}\right)^{r((p+1)\Gamma_n + 4) + 2}$$

$$\Rightarrow log\mathcal{N}_\infty\left(\varphi, \mathcal{F}_{r,n}, ||\cdot||_\infty\right) \leq r^2 log\left(\frac{2e}{\varphi}\right) + \left(r((p+1)\Gamma_n + 4) + 2\right)$$

$$\times log\left(\frac{(V_1 V_2)^2 \left(r((p+1)\Gamma_n + 4) + 2\right)}{\varphi(V_1 V_2 - 1)}\right)$$

$$\leq \left(r^2 + r((p+1)\Gamma_n + 4) + 2\right)$$

$$\times log\left(\frac{2e(V_1 V_2)^2 \left(r^2 + r((p+1)\Gamma_n + 4) + 2\right)}{\varphi(V_1 V_2 - 1)}\right)$$

**Step - II** From Lemma 3.8 of [22] for $\xi < 1$ along with the we have

$$\int_0^\xi \sqrt{log\mathcal{N}_\infty \left(\varphi, \mathcal{F}_{r,n}, ||\cdot||_\infty\right)}d\varphi$$

$$\leq \sqrt{(r^2 + r((p+1)\Gamma_n + 4) + 2)}\xi\sqrt{log\left(\frac{2e\left(V_1V_2\right)^2\left(r^2 + r((p+1)\Gamma_n + 4) + 2\right)}{\xi\left(V_1V_2 - 1\right)}\right)}$$

$$= \Upsilon_n(\xi)$$

**Step-III** Define a function $h : \xi \to \frac{\Upsilon_n(\xi)}{\xi^a}$ in $(0, \infty)$ for some $a > 0$ such that

$$h(\xi) = \sqrt{(r^2 + r((p+1)\Gamma_n + 4) + 2)}\xi^{(1-a)}$$

$$\times \sqrt{log\left(\frac{2e\left(V_1V_2\right)^2\left(r^2 + r((p+1)\Gamma_n + 4) + 2\right)}{\xi\left(V_1V_2 - 1\right)}\right)}$$

its derivative will become,

$$h'(\xi) = \frac{(1-a)}{\xi^a}\sqrt{log\left(\frac{2e\left(V_1V_2\right)^2\left(r^2 + r((p+1)\Gamma_n + 4) + 2\right)}{\xi\left(V_1V_2 - 1\right)}\right)}$$

$$- \frac{\xi^{-a}}{\sqrt{log\left(\frac{2e(V_1V_2)^2(r^2+r((p+1)\Gamma_n+4)+2)}{\xi(V_1V_2-1)}\right)}}$$

Thus the function $h$ is decreasing whenever, $a > 1$ and $0 < \xi < 1$.

**Step-IV** Assume $\gamma_n \lesssim ||\pi_{r,n}f_0 - f_0||_n^{-1}$ where $\gamma_n = \sqrt{\zeta_n^{-1}}$

$$\gamma_n^2\Upsilon_n\left(\frac{1}{\gamma_n}\right) = \gamma_n\sqrt{(r^2 + r((p+1)\Gamma_n + 4) + 2)}$$

$$\sqrt{log(\gamma_n) + log\left(\frac{2e\left(V_1V_2\right)^2\left(r^2 + r((p+1)\Gamma_n + 4) + 2\right)}{\left(V_1V_2 - 1\right)}\right)}$$

$$\lesssim \gamma_n\sqrt{(r^2 + r((p+1)\Gamma_n + 4) + 2)}$$

$$\sqrt{log(\gamma_n) + 2 + log\left(\left(V_1V_2\right)^2\left(r^2 + r((p+1)\Gamma_n + 4) + 2\right)\right)}$$

Step-V  Now $\gamma_n^2 \Upsilon_n \left( \frac{1}{\gamma_n} \right) \lesssim \sqrt{n}$ if and only if

$$\gamma_n^2 \left( r^2 + r((p+1)\Gamma_n + 4) + 2 \right)$$
$$\left( \log(\gamma_n) + \log \left( (V_1 V_2)^2 \left( r^2 + r((p+1)\Gamma_n + 4) + 2 \right) \right) \log n \right) \lesssim n$$

Therefore, $\gamma_n \lesssim$

$$\min \left\{ \left( \frac{n}{(r^2 + r((p+1)\Gamma_n + 4) + 2) \left( \log \left( (V_1 V_2)^2 \left( r^2 + r((p+1)\Gamma_n + 4) + 2 \right) \right) \right)} \right)^{1/2}, \right.$$
$$\left. \left( \frac{n}{(r^2 + r((p+1)\Gamma_n + 4) + 2) \log(n^{\psi+\beta} \log n)} \right)^{1/2} \right\}.$$

If we replace $\Gamma_n = o(n^\psi)$, $V_1 V_2 = o(\sqrt{\log n})$ and $r = o\left( n^\beta \right)$ such that $\psi + \beta < 1$ then,

$$\gamma_n \lesssim \min \left\{ \left( \frac{n}{n^{\psi+\beta} \log(n^{\psi+\beta} \log n)} \right)^{1/2}, \left( \frac{n}{n^{\psi+\beta} \log(n \log n)} \right)^{1/2} \right\}. \tag{C7}$$

From (C7) using Lemma 1, Lemma 2 of [2], Theorem 3.4.1 of [23] and triangle inequality we get

$$||\hat{f}_n - f_0||_n \leq ||\hat{f}_n - \pi_{r,n} f_0||_n + ||\pi_{r,n} f_0 - f_0||_n$$
$$= \mathcal{O}_P \left( max \left\{ ||\pi_{r,n} f_0 - f_0||_n, \right. \right.$$
$$\sqrt{ \frac{(r^2 + r((p+1)\Gamma_n + 4) + 2) \log \left( r^2 + r((p+1)\Gamma_n + 4) + 2 \right) (V_1 V_2)^2}{n} },$$
$$\left. \left. \sqrt{ \frac{(r^2 + r((p+1)\Gamma_n + 4) + 2) \log(n \log n)}{n} } \right\} \right)$$
$$\approx \mathcal{O}_P \left\{ max \left[ ||\pi_{r,n} f_0 - f_0||_n, \left( \frac{n}{n^{\psi+\beta} \log(n^{\psi+\beta} \log n)} \right)^{-1/2}, \right. \right.$$
$$\left. \left. \left( \frac{n}{n^{\psi+\beta} \log(n \log n)} \right)^{-1/2} \right] \right\}.$$

$\square$

Step-1: For any $\varepsilon > 0$, we have

$$P^* \left( \sup_{f \in \mathcal{F}_{r,n}} |\mathbb{L}_{n,\lambda}(f) - \mathcal{L}_{n,\lambda}(f)| > \frac{\varepsilon}{2} \right) \leq P \left( \left| \frac{1}{n} \sum_{t=1}^n \varepsilon_t^2 - \sigma^4 \right| > \frac{\varepsilon}{4} \right)$$
$$+ P^* \left( \sup_{f \in \mathcal{F}_{r,n}} \left| \frac{1}{n} \sum_{t=1}^n \epsilon_t \left\{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} + \lambda \|f\|_1 - \lambda \|f_0\|_1 \right| > \frac{\varepsilon}{4} \right). \tag{C8}$$

25

Step-2: The first part will converge to zero using the WLLN, but for the second part, using Markov's Inequality, we need to prove

$$E_\epsilon \left( \sup_{f \in \mathcal{F}_{r,n}} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \left\{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} + \lambda \|f\|_1 - \lambda \|f_0\|_1 \right| \right) \to 0 \text{ as } n \to \infty.$$

Step-3: According to symmetrization for $\alpha$-Mixing spatial surfaces with grid spacing $\eta_n \to 0$ as $n \to \infty$,

$$E_\epsilon \|\mathbb{P}_{n,\lambda} - P_\lambda\|_{\mathcal{F}_{r,n}} = E_\epsilon \left( \sup_{f \in \mathcal{F}_{r,n}} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \left\{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} + \lambda \|f\|_1 - \lambda \|f_0\|_1 \right| \right)$$

$$\leq 2E_\epsilon E_\rho \sup_{f \in \mathcal{F}_{r,n}} \left| \frac{1}{n} \sum_{t=1}^{n} \rho_t \epsilon_t \left\{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} + \lambda \|f\|_1 - \lambda \|f_0\|_1 \right|,$$

(C9)

where $\rho_t$; $t = 1, 2, ..., n$ are Rademacher random variables independent of $\epsilon_1, \epsilon_2, ..., \epsilon_n$. For fixed values of $\epsilon_1, \epsilon_2, ..., \epsilon_n$ (conditionally), $\sum_{t=1}^{n} \rho_t \epsilon_t \left\{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\}$ is a sub-Gaussian process.

Step-4: Using the definition of the covering number and Corollary 2.2.8 of [19], we obtain a universal constant $K$. Then, using Dudley's integral entropy bound in VC-dimension $\nu$ with any $f_n^* \in \mathcal{F}_{r,n}$ with $n \geq N_1$, we get

$$E_\rho \sup_{f \in \mathcal{F}_{r,n}} \left| \frac{1}{n} \sum_{t=1}^{n} \rho_t \epsilon_t \left\{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} + \lambda \|f\|_1 - \lambda \|f_0\|_1 \right|$$

$$\leq E_\rho \left| \frac{1}{n} \sum_{t=1}^{n} \rho_t \epsilon_t \left\{ f_n^*(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} \right|$$

$$+ K \sqrt{\frac{\nu}{n}} \int_0^{2V_1 V_2} \sqrt{\frac{\log \mathcal{N}(\varphi, \mathcal{F}_{r,n}, \| \cdot \|_\infty)}{n}} d\varphi.$$

(C10)

Step-5: The first term in the inequality can be handled similarly as in the original proof, using the concept of universal approximation. The second term will now include the effect of the $\ell_1$ regularization, leading to

$$E_\rho \sup_{f \in \mathcal{F}_{r,n}} \left| \frac{1}{n} \sum_{t=1}^{n} \rho_t \epsilon_t \left\{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \right\} \right|$$

$$+ K \sqrt{\frac{\nu}{n}} \int_0^{2V_1 V_2} \sqrt{\frac{\log \mathcal{N}\left( \frac{\varphi}{2V_1 V_2 \sqrt{\sigma^2 + 1}}, \mathcal{F}_{r,n}, \| \cdot \|_\infty \right)}{n}} d\varphi$$

$$+ \lambda \sup_{f \in \mathcal{F}_{r,n}} \|f\|_1 - \lambda \|f_0\|_1.$$

(C11)

26

Step-6: We proceed similarly to the original proof, but now also controlling the effect of the $\ell_1$ penalty. Assuming $\lambda = o(n^{-\alpha})$ for some $\alpha > 0$, the regularization term will diminish as $n \to \infty$.

Step-7: Combining all steps, including the regularization term, we get

$$
E_\epsilon \left( \sup_{f \in \mathcal{F}_{r,n}} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_t \{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \} + \lambda \| f \|_1 - \lambda \| f_0 \|_1 \right| \right)
$$

$$
\leq 2 E_\epsilon E_\rho \sup_{f \in \mathcal{F}_{r,n}} \left| \frac{1}{n} \sum_{t=1}^{n} \rho_t \epsilon_t \{ f(\mathcal{Y}_t) - f_0(\mathcal{Y}_t) \} \right| \qquad \text{(C12)}
$$

$$
+ 8 \sqrt{2} K \sqrt{\frac{\nu C_{r,p,\Gamma_n,V_1,V_2}}{n^2}} + \lambda \sup_{f \in \mathcal{F}_{r,n}} \| f \|_1 - \lambda \| f_0 \|_1.
$$

Step-8: To prove convergence, the expression should tend to zero as $n \to \infty$. The first part converges to zero by the universal approximation theorem. The second part also converges to zero under the assumptions $\Gamma_n = o(n^\beta), r = o(n^\psi)$, and $\psi + \beta < 1$. The third term involving $\lambda$ converges to zero by the choice of $\lambda$ as $o(n^{-\alpha})$.

# References

[1] Zhan, W., Datta, A.: Neural networks for geospatial data. Journal of the American Statistical Association (just-accepted), 1–21 (2024)

[2] Shen, X., Jiang, C., Sakhanenko, L., Lu, Q.: Asymptotic properties of neural network sieve estimators. Journal of nonparametric statistics **35**(4), 839–868 (2023)

[3] Schmidt-Hieber, J.: Nonparametric regression using deep neural networks with relu activation function (2020)

[4] Kohler, M., Langer, S.: On the rate of convergence of fully connected deep neural network regression estimates. The Annals of Statistics **49**(4), 2231–2249 (2021)

[5] Kohler, M., Krzyżak, A.: On the rate of convergence of a deep recurrent neural network estimate in a regression problem with dependent data. Bernoulli **29**(2), 1663–1685 (2023)

[6] Braun, A., Kohler, M., Langer, S., Walk, H.: Convergence rates for shallow neural networks learned by gradient descent. Bernoulli **30**(1), 475–502 (2024)

[7] Chen, W., Li, Y., Reich, B.J., Sun, Y.: Deepkriging: Spatially dependent deep neural networks for spatial prediction. arXiv preprint arXiv:2007.11972 (2020)

[8] Cressie, N., Johannesson, G.: Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society Series B: Statistical Methodology **70**(1), 209–226 (2008)

[9] Nag, P., Sun, Y., Reich, B.J.: Spatio-temporal deepkriging for interpolation and probabilistic forecasting. Spatial Statistics **57**, 100773 (2023)

[10] Gerber, F., Nychka, D.: Fast covariance parameter estimation of spatial gaussian process models using neural networks. Stat **10**(1), 382 (2021)

[11] Xiao, S., Song, Y., Wang, Z.: Nonparametric spatial autoregressive model using deep neural networks. Spatial Statistics **57**, 100766 (2023)

[12] Wikle, C.K., Zammit-Mangion, A.: Statistical deep learning for spatial and spatiotemporal data. Annual Review of Statistics and Its Application **10**(1), 247–270 (2023)

[13] Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., Sain, S.: A multiresolution gaussian process model for the analysis of large spatial datasets. Journal of Computational and Graphical Statistics **24**(2), 579–599 (2015)

[14] Lahiri, S.N.: Central limit theorems for weighted sums of a spatial process under a class of stochastic and fixed designs. Sankhyā: The Indian Journal of Statistics, 356–388 (2003)

[15] Cressie, N.: Statistics for Spatial Data. John Wiley & Sons, 111 River Street, Hoboken, NJ 07030, USA (2015)

[16] Grenander, U.: Abstract inference. (No Title) (1981)

[17] White, H.: Some results for sieve estimation with dependent observa-tions. Nonparametric and semiparametric methods in econometrics and statistics (1991)

[18] Chen, X.: Large sample sieve estimation of semi-nonparametric models. Handbook of econometrics **6**, 5549–5632 (2007)

[19] Vaart, A.v.d., Wellner, J.A.: Weak convergence and empirical processes with applications to statistics. Journal of the Royal Statistical Society-Series A Statistics in Society **160**(3), 596–608 (1997)

[20] Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural networks **2**(5), 359–366 (1989)

[21] Anthony, M., Bartlett, P.L., Bartlett, P.L., *et al.*: Neural Network Learning: Theoretical Foundations. Cambridge University press, Cambridge (1999)

[22] Mendelson, S., Smola, A.J.: Advanced Lectures on Machine Learning: Machine Learning Summer School 2002, Canberra, Australia, February 11-22, 2002, Revised Lectures vol. 2600. Springer, Canberra, Australia (2003)

[23] Van Der Vaart, A.W., Wellner, J.A., Vaart, A.W., Wellner, J.A.: Weak Convergence. Springer, New York (1996)