Modern non-linear function-on-function regression

Aniruddha Rajendra Rao¹ · Matthew Reimherr¹

Received: 27 November 2022 / Accepted: 18 September 2023 / Published online: 4 October 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
We introduce a new class of non-linear function-on-function regression models for functional data using neural networks. We propose a framework using a hidden layer consisting of continuous neurons, called a continuous hidden layer, for functional response modeling and give two model fitting strategies, function-on-function direct neural networks and function-on-function basis neural networks. Both are designed explicitly to exploit the structure inherent in functional data and capture the complex relations existing between the functional predictors and the functional response. We fit these models by deriving functional gradients and implement regularization techniques for more parsimonious results. We demonstrate the power and flexibility of our proposed method in handling complex functional models through extensive simulation studies as well as real data examples.

Keywords Functional data analysis · Deep learning · Function-on-function regression · Functional neural network · Functional response

1 Introduction

Functional data has gained a lot of attention in the past decades (Ramsay and Silverman 1997, Kokoszka and Remiherr 2018; Ferraty and Vieu 2006; Horváth and Kokoszka 2012) with advancement in data collection and storage technologies. Data that are represented as curves varying over a continuum (time, space, depth, wavelength, etc.) have been encountered widely in several areas such as Computer Vision, Time Series, Natural Language Processing, Genetics, Finance, Meteorology, Longitudinal Studies, and many more. There is a rich amount of literature in functional data analysis (FDA) ranging from topics like Functional Principal Component Analysis, Canonical Correlation Analysis, Functional Regression, etc.

Functional linear models (FLM) are the most well-studied models in FDA. We can distinguish these models as scalar-on-function, function-on-scalar, and function-on-function regression. When discussing such models, the first term indicates the type of the response and the second term indicates the type of the predictor. In contrast to Functional linear models, which have been widely studied and investigated, there has been substantially less work done in the area of non-linear functional regression models. It is natural to begin examining non-linear models after establishing the linear methods but there is also a practical need for such models as the data may contain complicated functional relations and interactions (Fan et al. 2015). Such type of data are common in practice.

In FDA, non-linear regression is an active area of research but most of the work concentrates on the scalar-on-function setting, while the function-on-function setting has picked up more interest recently. Some common ways of dealing with non-linearity in scalar-on-function setting are by using the continuously additive model (CAM) (McLean et al. 2014; Fan et al. 2015; Ma and Zhu 2016; Muller et al. 2013; Wang and Ruppert 2013; Reimherr et al. 2017), Quadratic Model (Yao and Muller 2010), Single Index Model (Stoker 1986; Eilers et al. 2009; Ferraty 2011), Multiple Index Model (James and Silverman 2004; Chen et al. 2012; Ferraty et al. 2013), Functional Neural Network (Rossi et al. 2002; Rossi and Conan-Guez 2006; Conan-Guez and Rossi 2002; Wang et al. 2019a, b) and Functional Direct Neural Network/Functional Basis Neural Network (Rao and Reimherr 2021). The above methods are analogous to their multivariate counterparts and most of statistical methods suffer...
from the drawback of either choosing to capture the complex non-linear relations or the higher order interactions. This is because they assume a very particular structure, which limits the kind of non-linear patterns that can be learned, especially in terms of interactions within the time points for the predictors. Reiss et al. (2016) and Morris (2014) have discussed most of the above approaches in their review papers. The deep learning methods like Functional Neural Network and the architecture proposed in Rao and Reimherr (2021) don’t have such limitations and have been proven to work well when the true nature of the functional data is unknown.

When it comes to function-on-function regression, Lian (2007), Kadri et al. (2010) extended on the work by Preda (2007), where they fit a fully non-linear model using Reproducing Kernel Hilbert Spaces (RKHS) to functional outcomes. More recent works are by Scheipl et al. (2015), Kim et al. (2018), Sun et al. (2018), and Sun and Wang (2020), where Scheipl et al. (2015) gave an additive model for function-on-function regression based on Muller et al. (2013) and fitted it using trivariate tensor product basis approach. Kim et al. (2018) introduced another estimation approach to fit the same model using orthogonal B-splines and functional principal component analysis (FPCA). The work by Sun et al. (2018) is an extension of Cai and Yuan (2012) where they use an RKHS framework to estimate a function-on-function linear model. The functional quadratic model by Yao and Muller (2010) was extended to the functional response case by Sun and Wang (2020). Also, functional single index models (Stoker 1986; Eilers et al. 2009; Ferraty 2011) were extended to the function-on-function setting by Jiang and Wang (2011). All these methods help fit different function-on-function models using basis expansion or RKHS but they are still restrictive, either they are linear models or ignore interactions. They all have a particular structure that limits the kind of patterns they can learn.

From a deep learning view, Rossi et al. (2002), Rossi and Conan-Guez (2006), Conan-Guez and Rossi (2002) first introduced the Functional Neural Network (FNN) and proved its validity by theoretical arguments and simulation experiments. FNN was then further investigated by Wang et al. (2019b), Wang et al. (2019a) using FPCA. They also demonstrated the superiority of FNN to other deep learning alternatives (CNN, RNN, LSTM). This expansion of the neural network to functional data enabled efficient non-linear learning when the predictors are functions. In the FNN architecture, only the first layer is a functional hidden layer with functional neurons that give a scalar output, which is then connected to regular hidden layers and finally the output layer. This approach does not embrace functional data completely. Wang et al. (2020) recently extended FNN to model functional response, but again the architecture is limited, as it only embraces the functional data in the first hidden layer and the final output layer. As we can see above, much of the work are extensions of some version of scalar-on-function models. We propose a new architecture inspired from FNN and Rao and Reimherr (2021) for modeling a functional response using functional inputs, where we have a continuous mapping throughout the network.

The major contribution of this work is solving the above problems by giving a novel framework that has no limitation with respect to the patterns it can learn while maintaining the functional nature of the data. Our framework consists of multiple continuous hidden layers consisting of continuous neurons that can learn functional representations from functional inputs to the functional outputs. We solve our framework by using two strategies in the function-on-function setting, function-on-function direct neural networks (FFDNN) and function-on-function basis neural networks (FFBNN). The rest of this paper is organized as follows: In Sect. 2, we first present some background on FDA and function-on-function models. Then we describe our framework and solve it using FFDNN and FFBNN. We end this section with penalization techniques to make our results much smoother and interpretable. We follow this with simulations and several benchmark applications in Sect. 3. We consider multiple simulation settings depending on how the non-linear relation is defined between the functional response and functional predictors, and demonstrate the superiority of our methods over current state-of-the-art approaches. In the final section, we share concluding remarks and future research avenues which pertain to methodological and theoretical directions.

2 Methodology

We first give the necessary notation used in this paper. Let us assume that we have \( N \in \mathcal{N} \) i.i.d subjects, where we observe functions \( \{(X_{i,1}, \ldots, X_{i,R}), (Y_i), i = 1, \ldots, N \} \) over a compact time interval \( T \). These functions could be observed on other intervals, but as long as they are closed and bounded, they can always be rescaled to be on the unit interval \([0,1]\). For the \( i \)th subject, the functional inputs are \( R \) random curves that can be denoted as \( X_{i,r} \in L^2(T) \) and the functional output is \( Y_i \in L^2(T) \) for \( i = 1, \ldots, N \) and \( r = 1, \ldots, R \). We assume that all the functions are completely observed. This is known as dense functional data analysis (Kokoszka and Reimherr 2018). Our main objective is to learn the mapping, \( F : L^2(T) \times \cdots \times L^2(T) \rightarrow L^2(T) \) from the functional inputs \( \{X_{i,r}\} \) to the functional output \( Y_i \):

\[
E[Y_i \mid X_{i,1}, \ldots, X_{i,R}] = F(X_{i,1}, \ldots, X_{i,R}) \tag{1}
\]
2.1 Functional regression

We briefly describe the most common function-on-function regression models, which will be used later in simulations. The integrals without limits are defined over the entire domain.

- Function-on-function Linear Model (FFLM):
  \[ E[Y_i(t) | X_{i,1}, \ldots, X_{i,R}] = \alpha(t) + \sum_{r=1}^{R} \int \beta_r(s, t) X_{i,r}(s) ds, \]
  \hspace{1cm} (2)

  where \( \alpha(t) \) is an intercept function and \( \beta_r(s, t) \) are regression coefficient functions.

- Continuously additive model (CAM):
  \[ E[Y_i(t) | X_{i,1}, \ldots, X_{i,R}] = \sum_{r=1}^{R} \int f(X_{i,r}(s), s, t) ds \]
  \hspace{1cm} (3)

  where the trivariate parameter function \( (x, s, t) \rightarrow f(x, s, t) \) is smooth (Ma and Zhu 2016; Reimherr et al. 2017; Scheipl et al. 2015).

- Single Index Model:
  \[ E[Y_i(t) | X_{i,1}, \ldots, X_{i,R}] = g \left( t, \sum_{r=1}^{R} \langle \beta_r, X_{i,r} \rangle \right) \]
  \hspace{1cm} (4)

  where the function \( g() \) is any unknown smooth function defined on the real line (Jiang and Wang 2011).

- Multiple Index Model:
  \[ E[Y_i(t) | X_{i,1}, \ldots, X_{i,R}] = g \left( t, \sum_{r=1}^{R} \langle \beta_{1r}, X_{i,r} \rangle, \ldots, \sum_{r=1}^{R} \langle \beta_{Pr}, X_{i,r} \rangle \right) \]
  \hspace{1cm} (5)

  We define the above multiple Index models for simulation purposes, which is extension of the single index model with \( P \) indexes. The scalar-on-function version of this model is well studied in Chen et al. (2012), Ferraty et al. (2013), Muller and Yao (2008), Zhu et al. (2013).

- Quadratic Model:
  \[ E[Y_i(t) | X_{i,1}, \ldots, X_{i,R}] = \alpha(t) + \sum_{r=1}^{R} \int \beta_r(s, t) X_{i,r}(s) ds \]
  \hspace{1cm} + \sum_{r=1}^{R} \int \int \beta_r(q, s, t) X_{i,r}(q) X_{i,r}(s) dq ds, \]
  \hspace{1cm} (6)

  \[ E[Y_i(t) | X_{i,1}, \ldots, X_{i,R}] = \sum_{r=1}^{R} \int f(X_{i,r}(s), s, t) ds + \sum_{r=1}^{R} \int \int f(X_{i,r}(q), X_{i,r}(t), q, s, t) dq ds \]
  \hspace{1cm} (7)

  Equation (6) represents a Quadratic Model given by Sun and Wang (2020) where \( \alpha(t), \beta(s, t) \) and \( \beta(q, s, t) \) are unknown function coefficients. The Quadratic term is regarded as an interaction effect of the functional predictors on the functional response. For simulation purposes we define Eq. (7) (Complex Quadratic Model).

These non-linear models are analogous to their multivariate and scalar-on-function counterparts, which help to overcome the curse of dimensionality by using higher order terms or through indexes. But these methods come with their own drawbacks like ignoring interactions, an inability to capture complex non-linear relationships, and restrictive structures. The true form of the relation between a functional response and the functional predictors in practice is difficult to identify. Hence, developing an approach that can deal with more general relationships between these functions without knowing the underlying ground truth is beneficial.

2.2 Our network framework

We propose a framework below, which develops a continuous mapping from layer to layer using multiple continuous hidden layers consisting of multiple continuous neurons. In our framework, as seen in Fig. 1, we have three types of layers that pass functions through the network: an input layer, which is the first layer that takes in the functional predictors, the continuous hidden layers containing continuous neurons, and a continuous output layer that provides the final fitted functions using continuous neurons. We learn using the functional (univariate and bivariate) weights that are continuous over time (or other continuums). The main motivation for this approach is to preserve the functional nature of the data throughout the whole modeling process. This provides a richer structure to model functional data and also exploits the domain information and continuity of the predictor functions and response function. The \( l \)th continuous hidden layer and its \( k \)th continuous neuron is defined as:

\[ H_{(k)}^{(l)}(s) = \sigma \left( b_{(k)}^{(l)}(s) + \sum_{j=1}^{J} \int w_{(j,k)}^{(l)}(s, t) H_{(j)}^{(l-1)}(t) dt \right) \]

\hspace{1cm} (8)
where \( l = 1, 2, 3, \ldots, L, \ H_{(j)}^{(0)}(s) = X_j(s), \ H_{(j)}^{(l)}(t) = \hat{Y}(t) \), \( b_{(k)}^{(l)} \in \mathcal{L}^2(\mathcal{T}) \) is the unknown intercept function, \( w_{(j,k)}^{(l)} \in \mathcal{L}^2(\mathcal{T} \times \mathcal{T}) \) is the bivariate parameter function for the \( k \)th continuous neuron in the \( l \)th hidden layer coming from the \( j \)th continuous neuron of the \((l-1)\)th hidden layer and \( \sigma(\cdot) \) is a non-linear activation function. For simplicity and applicability of other methods, we have assumed a single output function but our approach can easily accommodate for multiple response functions.

The proposed approach can be implemented either using FFDNN or FFBNN. FFDNN is a direct solution when the data is dense and is straightforward to implement. We learn through the basis weights, which allows the data to be changed the grid length (s) between the input and output layer. FFBNN uses basis expansions that allow the data to be irregular and learns through the functional weights and also have the flexibility for greater parsimony.

Functional Direct Neural Network (FFDNN) Let us assume that Fig. 1 has \( R \) inputs (functional predictors), \( L \) continuous hidden layers, each with \( J \) incoming and \( K \) outgoing connections. FFDNN is similar to a neural network in nature for functional data, where we can define the number of continuous hidden layers and the number of continuous neurons in each continuous hidden layer. We use either ReLU, tanh, sigmoid/logistic, or linear (output layer) as the activation function in the continuous neurons. We reach from the functional predictors to the functional response via equation (8).

While, using backpropagation (i.e. gradient decent), we learn the weight functions and the intercept functions.

We use the functional gradients given below to optimize for the network parameters. The gradient of the weight function measures the change in a functional to a change in a function on which the functional depends. Rossi et al. (2002), Wang et al. (2019b), Olver (2019) give the necessary assumptions and mathematical tools (Fréchet derivatives) for getting the functional gradients. During the backpropagation step, we pass through the network backwards and we calculate the partial derivatives for the weight functions and the intercept functions. We define the loss function as \( \mathcal{L}(\theta) \), where \( \theta \) is the collection of all functional parameters. We define a quadratic loss function,

\[
\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left( Y_i(t) - \hat{Y}_i(t) \right)^2 dt.
\]

The partial derivatives needed for the parameters in our proposed architecture (as shown in Fig. 1) are as follows:

\[
\frac{\partial H_{(j)}^{(l)}}{\partial b_{(k)}^{(l)}}(s) = \sigma' \left( b_{(k)}^{(l)} \right) + \sum_{j=1}^{J} \int w_{(j,k)}^{(l)}(s,t) \ H_{(j)}^{(l-1)}(t) dt
\]

\[
\frac{\partial H_{(j)}^{(l)}}{\partial w_{(j,k)}^{(l)}}(s,t) = \sigma' \left( b_{(k)}^{(l)} \right) + \sum_{j=1}^{J} \int w_{(j,k)}^{(l)}(s,t') \ H_{(j)}^{(l-1)}(t') dt'
\]

\[
\left( \frac{\partial}{\partial w_{(j,k)}^{(l)}} \right) \int \frac{\partial}{\partial w_{(j,k)}^{(l)}} w_{(j,k)}^{(l)}(s,t) \ H_{(j)}^{(l-1)}(t) dt
\]

\[
= \sigma' \left( b_{(k)}^{(l)} \right) + \sum_{j=1}^{J} \int w_{(j,k)}^{(l)}(s,t') \ H_{(j)}^{(l-1)}(t') dt'
\]

\[
\times H_{(j)}^{(l-1)}(t)
\]

where \( l = 1, 2, 3, \ldots, L, \ H_{(j)}^{(0)}(s) = X_j(s), \ H_{(j)}^{(l)}(t) = \hat{Y}(t) \) and \( \sigma'(\cdot) \) represents the first derivative of \( \sigma(\cdot) \).

In our architecture, the grid/domain (s), the number of continuous hidden layers, and the number of continuous neurons in each of the continuous hidden layer can be considered as hyperparameters. We optimize for the parameters in the same way for FFDNN as in classic neural networks, using backpropagation with the help of the gradients given above.

**Fig. 1** General architecture for our proposed approach with multiple continuous hidden layers
Functional Basis Neural Network (FFBNN) In FFBNN, the continuous neurons are expanded using a basis expansion. The network learns the weights of the basis functions rather than the weight function itself. Basis functions can also help us to deal with irregular data more efficiently (Kokoszka and Reimherr 2018). In the forward propagation phase, FBBNN uses basis expansions where we replace the weight functions \( w_{(j,k)}^{(l)}(s,t) \) in Eq. (8) by a linear combination of basis functions as shown below. Assume subscripts \( j,k \) represent the continuous neurons of the neural network and superscript \( l \) denote the \( l \)th layer for the weight parameter \( W_{(j,k)}^{(l)} \). The notation \( v \) represents a standard collection of basis functions (splines, wavelets, or sine and cosine functions). Let us assume for simplicity, that we use the same basis functions and the same number of basis in each of the continuous neurons in the continuous hidden layers across the network. This gives us the following Equation:

\[
H_{(k)}^{(l)}(s) = \sigma \left( b_{(k)}^{(l)}(s) + \sum_{j=1}^{J} w_{(j,k)}^{(l)}(s,t) H_{(j)}^{(l-1)}(t) dt \right) \\
= \sigma \left( \sum_{b=1}^{B} b_{(k)}^{(l),b} v_{b}^{s}(s) \right) \\
+ \sum_{j=1}^{J} \sum_{c=1}^{C} \sum_{d=1}^{D} w_{(j,k),c,d}^{(l)} v_{c}(s) \int v_{d}(t) H_{(j)}^{(l-1)}(t) dt \\
= \sigma \left( \sum_{b=1}^{B} b_{(k)}^{(l),b} v_{b}^{s}(s) \right) \\
+ \sum_{j=1}^{J} \sum_{c=1}^{C} \sum_{d=1}^{D} w_{(j,k),c,d}^{(l)} v_{c}(s) A_{(j),d}^{(l)} \\
= \sigma \left( b_{(k)}^{(l)} v^{s}(s) + \sum_{j=1}^{J} W_{(j,k)}^{(l)}(s) A_{(j)}^{(l)} \right) \\
\tag{12}
\]

where \( l = 1, 2, 3, \ldots, L, H_{(j)}^{(0)}(s) = X_{j}(s) \) and \( H_{(L)}^{(l)}(t) = \tilde{Y}(t) \). \( B_{(k)}^{(l)}, W_{(j,k)}^{(l)} \) and \( W_{(j,k)}^{(L+1)} \) are the unknown basis weights corresponding to their respective basis functions \( v^{s}(s), v(s) \), and \( v(t) \). \( B, C \) and \( D \) specify the number of basis functions used in each of the continuous neuron of the continuous hidden layers. \( \sigma(\cdot) \) is a non-linear activation function.

Similar to FFDNN, the values of the grid/domain size \( (s) \), the number of continuous hidden layers, the number of continuous neurons in each continuous hidden layer, the selection of the basis functions, and the number of basis functions can be considered as hyperparameters. In FFBNN, we learn using the basis weights of the basis functions, which are required to reach from the functional predictors to the functional response via a chain of multiple CAMs in the forward propagation phase. The partial derivatives required for the parameters in our proposed architecture (as seen in Fig. 1) are as follows:

\[
\frac{\partial H_{(j)}^{(l)}(s)}{\partial B_{(k)}^{(l)}} = \sigma' \left( B_{(k)}^{(l)} v^{s}(s) + \sum_{j=1}^{J} W_{(j,k)}^{(l)}(s) A_{(j)}^{(l)} \right) \times v^{s}(s) \\
\frac{\partial H_{(j)}^{(l)}(s)}{\partial W_{(j,k)}^{(l)}} = \sigma' \left( B_{(k)}^{(l)} v^{s}(s) + \sum_{j=1}^{J} W_{(j,k)}^{(l)}(s) A_{(j)}^{(l)} \right) \\
\times v(s) A_{(j)}^{(l)} \\
\tag{13}
\]

where \( l = 1,2,3,\ldots,L, H_{(j)}^{(0)}(s) = X_{j}(s), H_{(L)}^{(l)}(t) = \tilde{Y}(t) \), and \( \sigma'(\cdot) \) represents the first derivative of \( \sigma(\cdot) \).

2.3 Regularization

We run into the problem of over-fitting with our approach like any other Machine Learning and Deep Learning method due to the overparameterization of the model. We first explore early stopping, which requires a validation set to stop the model fitting process early, before it starts to over-fit. Early stopping is straightforward to implement but splitting the training set further into validation set, can lead to a loss in performance, especially when the sample size is small. We instead consider a cross-validated early stopping scheme to overcome this problem. We use k-fold cross-validation and note the early stopping iteration values in each fold along with the estimated parameter functions. We check the result of different early stopping schemes using central tendency of the iterations and the functional parameter values in the k-folds as shown in “Appendix” Tables 5 and 6.

Along with early stopping to deal with the over-fitting, we also consider a regularization approach by penalizing the roughness of the parameter functions in our framework. This regularization is called Roughness Penalization and is a more common regularization approach in FDA as well as other branches of nonparametric statistics. Interpreting black box models like our approach is a significant challenge in machine learning and deep learning, and can significantly reduce barriers to the adoption of the technology. We want to make use of the Roughness Penalty to move towards a more interpretable model. In this case, our loss function from Eq. (9) changes to

\[
L_{\lambda}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \int \left( Y_{i}(t) - \hat{Y}_{i}(t) \right)^{2} dt \\
+ \lambda_{b} J_{b}(\theta_{b}) + \lambda_{w} J_{w}(\theta_{w}) \\
\tag{14}
\]

where \( \lambda_{b} \) and \( \lambda_{w} \) are the smoothing parameters. The terms \( J_{b}(\theta_{b}) \) and \( J_{w}(\theta_{w}) \) are used to penalize the intercept functions and bivariate functions respectively. We define \( J_{b}(\theta_{b}) \) and \( J_{w}(\theta_{w}) \) as the second order derivative and Laplacian operator of the parameter functions (Ramsay and Silverman 1997; Kokoszka and Reimherr 2018) as follows,
We consider a single predictor function (i.e. $R=1$) which leads to a smoother result.

In this section, we compare different approaches under multiple simulation setting to show the effectiveness of FFDNN and FFBNN, as well as bench marking with multiple real-world data sets. We demonstrate the following objectives through our results: (1) FFDNN and FFBNN are effectively trained by the derived optimization method (2) FFDNN and FFBNN enable learning of complex relation between the functional predictors and the functional response (3) FFDNN and FFBNN outperform other methods (4) Regularization methods to our method is effective at improving the performance of FFDNN and FFBNN, as seen in the next section.

### 3 Empirical results

In this section, we compare different approaches under multiple simulation setting to show the effectiveness of FFDNN and FFBNN, as well as bench marking with multiple real-world data sets. We demonstrate the following objectives through our results: (1) FFDNN and FFBNN are effectively trained by the derived optimization method (2) FFDNN and FFBNN enable learning of complex relation between the functional predictors and the functional response (3) FFDNN and FFBNN outperform other methods (4) Regularization methods to our method is effective at improving the performance of FFDNN and FFBNN, as seen in the next section.

#### 3.1 Simulations

We consider a single predictor function (i.e. $R=1$) which is observed on a dense and regular grid. We generate $n = 1100$ iid random curves $\{X_1(t), \ldots, X_n(t)\}$ from a Gaussian process with mean 0 and covariance given as $C_X(t,s) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-\frac{1}{2}}} \left(\frac{\sqrt{2\nu|t-s|}}{\rho}\right)^\nu K_\nu \left(\frac{\sqrt{2\nu|t-s|}}{\rho}\right)$. This is the Matérn covariance function and $K_\nu$ is the modified Bessel function of the second kind. We set $\rho = 0.5$, $\nu = 5/2$ and $\sigma^2 = 1$. These curves are observed in an equally-spaced interval from $[0, 1]$ at $m = 100$ time points.

Table 1 gives the different data generating models for the function-on-function setting. The response curves are evaluated at $m_y = 75$ equally-spaced time points from $[0, 1]$. Random noise ($\epsilon_i(t)$) is added to our functional response model to adjust the signal to noise ratio. Each simulation setting is repeated 100 times and we divide the 1100 samples into 600 for training and 500 for testing. We use 100 samples from the training set as a validation set, which is required for Early Stopping in all the deep learning models (models that do not require early stopping use the 600 observations for training). In case of early stopping using Laplace operator and basis weights, we use 5-fold cross-validation to tune the smoothing parameter $\lambda$. We report the root mean square error (RMSE) of the prediction using Eq. (9). We apply traditional NN directly to the curves and FNN based on Wang et al. (2019a). For our approach, we consider multiple combinations of continuous hidden layers ($l = 1, 2$) and continuous neurons ($j = 1, 2, 4$). The grid ($s$) value in the continuous neurons is in the range $(30, 300)$. The basis functions used in FFBNN are B-splines and the number of basis used is in the range $(5, 100)$.

We compare the performance of each method under different simulation settings with a functional response in Table 2, where the best result is shown in bold. We observe that only when the true model is Linear do all methods perform roughly the same, but as the relations become more complex and interactions are considered, many of the methods fail to capture the true relation between the functional predictor and the functional response. We see a substantial gain in such situations in our approach, which dominates all the other methods by giving the lowest RMSE (square root value of Eq. 9) values. The Complex Quadratic setting is especially interesting as the difference in performance in other methods to our method is substantial. The best possible RMSE is one (since that is the variance of the error function) and our method is performing well irrespective of the true nature between the functional predictor and the functional response.

Table 3 compares the result of our method under early stopping and regularization using the roughness penalty (denoted by FFDNN$_R$ and FFBNN$_R$). We perform 5-fold cross-validation on the training set (600 samples) to tune the smoothing parameter ($\lambda$) in the regularization approach. The roughness penalty helps in preventing overfitting. We observe that the RMSE for the roughness penalty is slightly higher compared to the early stopping results but still competitive, and is better in comparison to other methods in Table 2. This marginal decrease in performance for roughness penalty helps us to gain smoother results. Figure 2 represents the intercept function for an arbitrary continuous neuron in FFDNN and FFBNN with and without the roughness penalty. We can see that the roughness penalty intercept function captures the general shape very well while still maintaining reasonable performance in terms of RMSE. The bivariate
**Table 1** Different simulation scenarios with functional response and a single predictor function

Simulation scenarios

1. Linear: \( Y_i(t) = \alpha(t) + \int \beta(s, t) X_i(s) ds + \epsilon_i(t) \)
\[ \text{where } \alpha = 0, \beta(s, t) = \beta_1(s) \ast \beta_2(t), \beta_1(s) = 5 \sin(2\pi t), \beta_2(t) = 3 \sin(3\pi t) \]

2. CAM: \( Y_i(t) = \int f(X_i(s), s, t) ds + \epsilon_i(t) \)
\[ \text{where } f(X_i(s), s, t) = X_i(s)^2st \]

3. Single index: \( Y_i(t) = g(t, (\beta_1, X_i))) + \epsilon_i(t) \)
\[ \text{where } g(a, b) = a^0b^2, \beta(s, t) = \beta_1(s) \ast \beta_2(t), \beta_1(s) = 5 \sin(2\pi t), \beta_2(t) = 3 \sin(3\pi t) \]

4. Multiple index: \( Y_i(t) = g(t, (\beta_1, X_i), (\beta_2, X_i))) + \epsilon_i(t) \)
\[ \text{where } g(a, b, c) = a^0b^2c^2, \beta_1(s, t) = \beta_1(s) \ast \beta_2(t), \beta_1(s) = 5 \sin(2\pi t), \beta_2(t) = 3 \sin(3\pi t), \beta_3(t) = 4 \sin(5\pi t), \beta_4(t) = 2 \sin(3\pi t) \]

5. Quadratic: \( Y_i(t) = \int \beta(s, t) X_i(s) ds + \int \int f(q, s, t) X_i(q) X_i(s) dq ds + \epsilon_i(t) \)
\[ \text{where } \alpha = 0, \beta(s, t) = \beta_4(s) \ast \beta_5(t), \beta_4(s) = 5 \sin(2\pi s), \beta_5(t) = 3 \sin(3\pi s), \beta(q, s, t) = \beta_6(q) \ast \beta_7(s) \ast \beta_8(t), \beta_6(q) = 5 \sin(3\pi q), \beta_7(s) = 5 \sin(\pi s), \beta_8(t) = 5 \sin(\pi t) \]

6. Complex quadratic: \( Y_i(t) = \int f(X_i(s), s, t) ds + \int \int f(X_i(q), X_i(s), q, s, t) dq ds + \epsilon_i(t) \)
\[ \text{where } f(X_i(s), s, t) = X_i(t)^2st, f(X_i(q), X_i(s), q, s, t) = a^3, a \text{ is the quadratic term from above.} \]

**Table 2** Comparison of the RMSE on the test set for the methods under different setting between the functional predictor and the functional response

| Method            | FFLM | CAM | NN  | FNN  | FFDNN | FFBN | FBNN |
|-------------------|------|-----|-----|------|-------|------|------|
| Linear            | 1.014| 1.014| 1.011| 1.007| 1.009 | **1.003** |
| CAM               | 1.732| 1.409| 1.651| 1.578| 1.225 | **1.213** |
| Single index      | 1.572| 1.505| 1.089| 1.056| **1.019** | 1.047 |
| Multiple index    | 2.106| 1.875| 2.123| 1.854| **1.422** | 1.552 |
| Quadratic         | 1.672| 1.221| 1.072| 1.059| **1.025** | 1.052 |
| Complex quadratic | 6.777| 4.624| 4.098| 3.575| **2.293** | 2.633 |

**Fig. 2** Smoothing for the intercept function \( b(s) \) under regularization method
function for FFDNN with and without the roughness penalty is shown in Fig. 3, where we can see a huge difference in the shape of the function, but the performance in terms of RMSE is similar. In the case of FFBNN, the bivariate function is relatively smoother compared to FFDNN without the roughness penalty and this is because we are using basis expansion. By adding the roughness penalty to FFBNN, we make the bivariate function much smoother and reduce its range significantly as seen in Fig. 4. We again see that the bivariate function with roughness penalty captures the general shape very well in FFBNN and maintains reasonable performance in terms of RMSE as well. This is expected as we are penalizing the second derivative of the function. Regularization does not need a validation set, so there is no loss of information. Unlike early stopping, where we can stop at any intermediate iteration depending on the performance on the validation set, with roughness penalty, we can let the model stabilize to get the best result. The roughness penalty helps to get more reasonable parameter functions for each continuous neuron, opening up opportunities for interpretation and a better understanding of our approach. We check the performance for different early stopping strategies using 5-fold cross-validation but the results are similar with no clear winner as shown in Tables 5 and 6 (“Appendix”).

We would also like to highlight the computational efficiency of various methods used, which is important for data analysis. FFLM proved to be the most time-efficient, with each iteration taking a mere 20 s, while CAM method, although slightly more time-consuming at 1.5 min per simulation, still maintained reasonable computational efficiency. The run times of the deep learning approaches vary significantly, contingent upon the architectural intricacies and other hyperparameters. We endeavored to maintain a consistent experimental environment, keeping the deep learning models as similar as possible. The standard neural network (NN) model required 45 s per simulation, whereas FNN needed 3 min, primarily due to the additional computational overhead incurred during the computation of functional principal component analysis (FPCA), which is integral to the FNN methodology. For our approaches, FFDNN needed 45 s for each iteration and the FFDBNN required 40 s per iteration. FFBNN runs slightly faster than FFDNN because the computational complexity is decreased with the use of basis expansion. Note that these experiments were conducted using R Software on a DELL XPS laptop equipped with an Intel Core i7 processor with 8 GB of memory.

3.2 Real data

We analyze two different real data sets. The first example deals with the relation between electricity demand and temperature in Adelaide, Australia (Magnano et al. 2008), while the second examines how temperature for a day affects the bike rental patterns for that day (Fanaee-T and Gama 2013). In both examples, the main motivation is to leverage weather patterns (which are readily available forecasts) to understand how temperature affects electricity demand and bike rentals.

In the Adelaide data (available in fds package in R), learning the variability of electricity demand in the city over time is of significance, as it can help to improve electricity storage, leading to cost reduction and sustainability. Several factors contribute to electricity demand but temperature can be considered as one of the most important variable related to it. Electricity demand has different dynamics on weekends compared to weekdays. We restrict our study to Saturday. Our focus is on how Saturday’s electricity demand relates to the temperature. Therefore, we model the daily temperature to the daily electricity demand from 7/6/1997 to 3/31/2007 (508 days). Both temperature and electricity demand (Megawatts) are measured half-hourly, i.e. the functional predictor and the functional response are dense functional data measured at 48 time points in a day. Figure 5 shows the half-hourly temperature and electricity demand for a sample of 50 days.
### Table 3
Comparison of the RMSE on the test set for our method with early stopping and regularization under different setting between the functional predictor and the functional response

| Mapping        | FFDNN | FFDNN_R | FFBNN | FFBNN_R |
|----------------|-------|---------|-------|---------|
| Linear         | 1.009 | 1.010   | 1.003 | 1.004   |
| CAM            | 1.225 | 1.328   | 1.213 | 1.335   |
| Single index   | 1.019 | 1.043   | 1.047 | 1.116   |
| Multiple index | 1.422 | 1.549   | 1.552 | 1.631   |
| Quadratic      | 1.025 | 1.026   | 1.054 | 1.065   |
| Complex quadratic | 2.293 | 2.133   | 2.337 | 2.783   |

### Fig. 5
Sample curves of temperatures in degrees (left panel) and the electricity demand (right panel) recorded half-hourly on Saturdays.

The bike data (available in the UCI data repository) consists of bike rental information for Washington, D.C., collected from the Capital Bike Share system. Bicycle rentals have become popular in recent years and are an attractive option to owning one. The goal is to learn the relation between bike demand and temperature so that the company avoids deploying too few or too many bikes than needed.

We are interested in casual rentals per hour for a day, which are rentals to cyclists without membership. The bike rental dynamics vary on weekends compared to weekdays. We consider only Saturday rentals similar to the study in Kim et al. (2018). Hence, we model for the Saturday casual rentals using temperature information. We have the hourly rental and temperature information for the period of 01/01/2011 to 12/31/2012 (105 Saturdays). Since the measurements are hourly, we have 24 time points per day. Figure 6 shows the counts of casual bike rentals and hourly temperature for 50 samples.

For the Adelaide data, we divide it into 408 training (80 validation) and 100 testing samples, and for the bike data, we have 85 training (16 validation) and 20 testing samples. Also, the bike data has a small amount of missing values, which requires smoothing the temperature curves using FPCA. We can see from Table 4 that the true nature between the functional predictor and the functional response is non-linear and our methods perform the best when comparing the RMSE with other methods.

### 4 Conclusions
In this paper, we presented a novel approach for complex functional regression with a functional response. This approach was a natural extension of a neural network to functional data. We capture the temporal dependency and complicated relationships between the functional predictors and the functional response through the use of continuous hidden layers with continuous neurons. FFDNN and FFBNN don’t assume any particular structure that limits the kind of non-linear patterns learned, especially, interactions within
time points for the predictors. We also explored regularization techniques to get smoother results while maintaining the performance.

FFDNN and FFBNN are working not only on par with the best existing methods under simple cases but are outperforming in the case of non-linear settings. These approaches are appealing in practice as it provides a way for researchers to systematically conduct deep learning tasks when they have functional predictors and functional response. The main difference between FFDNN and FFBNN is that FFDNN is more straightforward to apply on dense and nicely gridded data with less tuning, whereas, FFBNN is more flexible and allows for greater parsimony with very dense grids as it only learns the basis weights. A deeper theoretical investigation of these approaches is needed and is an avenue for future work, to develop the statistical theory to give these methods a sound backing. Also, the performance of these approaches under a sparse setting can be of interest but requires nontrivial extensions.

Deep learning methods have been helping to make discoveries in multiple fields. In this paper, we merged the idea of neural network and functional data. There are multiple other methods in machine learning and deep learning that can be tailored for functional data like Clustering, Autoencoders, convolutional neural network (CNN), recurrent neural network (RNN), and many more. The applicability of such models is of huge significance as they can be used in many areas including Time Series, Natural Language Processing, and Computer Vision. Also, there is a lot of ongoing research for interpreting these black-box methods in deep learning. A major hurdle is defining interpretability for the problem in hand. While the roughness penalty is a step in the right direction, a deep dive into our approach for better understanding and interpretation can lead to more meaningful insights.

Acknowledgements This research was supported in part by NSF SES-1853209.

Author Contributions ARR performed the computations. MR supervised the findings of this work. All authors discussed the results and Aniruddha wrote the manuscript with support from MR.
Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Appendix A: Cross-validation (CV) performance

In Tables 5 and 6, we compare the results of different early stopping strategies using 5-fold cross-validation (CV) without a validation set against the results from early stopping with validation set. The strategies we use are, taking the mean, median, maximum and minimum of the early stopping iteration values between the 5-folds and also the average of the functional parameter values across the 5-folds denoted by Wavg. As seen in the tables below, the results are inconsistent and there is no clear winner. Also, interestingly the difference in performance is more for FFBNN compared to FFDNN. For now, the mean and median strategy looks to be the most promising among the other options. A more holistic investigation is required to find the best option. One reason might be the black-box nature of our approach, as depending on the architecture of the network, step size, and initialization of the functions, different folds are learning at a different rate and are stopping early or much later. This could be causing under-fitting in the overall model.

| Table 5 | Comparison of the RMSE for FFDNN method with multiple early stopping strategies using CV under different setting between the functional predictor and the functional response |
|---------|-------------------------------------------------|
| Mapping | FFDNN Mean | FFDNN Med | FFDNN max | FFDNN min | FFDNN Wavg | FFDNN |
| Linear  | 3.054      | 3.054     | 1.949     | 3.054     | 1.034     | 1.009  |
| CAM     | 1.083      | 1.099     | 1.089     | 1.087     | 1.114     | 1.225  |
| Single index | 1.026  | 1.027     | 1.043     | 1.019     | 1.168     | 1.028  |
| Multiple index | 1.550   | 1.563     | 1.573     | 1.692     | 2.164     | 1.422  |
| Quadratic | 1.094   | 1.175     | 1.093     | 1.202     | 1.420     | 1.025  |
| Complex quadratic | 3.037  | 3.030     | 3.017     | 2.293     | 3.010     | 2.316  |

| Table 6 | Comparison of the RMSE for FFBNN method with multiple early stopping strategies using CV under different setting between the functional predictor and the functional response |
|---------|-------------------------------------------------|
| Mapping | FFBNN Mean | FFBNN Med | FFBNN Max | FFBNN Min | FFBNN Wavg | FFBNN |
| Linear  | 1.324      | 1.352     | 1.197     | 1.208     | 1.625     | 1.003  |
| CAM     | 1.060      | 1.060     | 1.059     | 1.064     | 1.356     | 1.213  |
| Single index | 1.378  | 1.378     | 1.378     | 1.378     | 1.381     | 1.047  |
| Multiple index | 2.069   | 2.085     | 2.043     | 2.426     | 2.467     | 1.552  |
| Quadratic | 1.717   | 1.717     | 1.717     | 1.633     | 1.608     | 1.052  |
| Complex quadratic | 3.045  | 3.002     | 2.034     | 2.962     | 2.958     | 2.633  |
References

Cai, T.T., Yuan, M.: Minimax and adaptive prediction for functional linear regression. J. Am. Stat. Assoc. 107(499), 1201–1216 (2012)

Chen, D., Hall, P., Müller, H.: Single and multiple index functional regression models with nonparametric link. Ann. Stat. 39, 11 (2012)

Conan-Guez, B., Rossi, F.: Multi-layer perceptrons for functional data analysis: a projection based approach. In: ICANN (2002)

Eilers, P.H., Li, B., Marx, B.D.: Multivariate calibration with single-index signal regression. Chemom. Intell. Lab. Syst. 96(2), 196–202 (2009)

Fan, Y., James, G.M., Radchenko, P.: Functional additive regression. Ann. Stat. 43(5), 2296–2325 (2015)

Fanane-T., H., Gama, J.: Event labeling combining ensemble detectors and background knowledge. Prog. Artif. Intell. 2, 1–15 (2013)

Ferraty, F.: Recent advances in functional data analysis and related topics (2011)

Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistics, Springer, Berlin (2006)

Ferraty, F., Goia, A., Salinelli, E., Vieu, P.: Functional projection pursuit regression. TEST 22, 293–320 (2013)

Horváth, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer, New York (2012)

James, G., Silverman, B.: Functional adaptive model estimation. J. Am. Stat. Assoc. 100, 565–576 (2004)

Jiang, C.-R., Wang, J.-L.: Functional single index models for longitudinal data. Ann. Stat. 39(1), 362–388 (2011)

Kadri, H., Duffos, E., Preux, P., Canu, S., Davy, M.: Nonlinear functional regression: a functional RKHS approach. In: Teh, Y.W., Titterington, M. (eds.) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, vol. 9, pp. 374–380. Chia Laguna Resort, Sardinia (2010)

Kang, H.B., Reimherr, M., Shriver, M., Claes, P.: Manifold data analysis with applications to high-frequency 3D imaging (2017)

Kim, J.S., Staicu, A.-M., Maitty, A., Carroll, R.J., Ruppert, D.: Additive function-on-function regression. J. Comput. Graph. Stat. 27(1), 234–244 (2018)

Kokoszka, P., Reimherr, M.: Introduction to Functional Data Analysis. CRC, New York (2018)

Lian, H.: Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. Can. J. Stat. 35(4), 597–606 (2007)

Ma, H., Zhu, Z.: Continuously dynamic additive models for functional data. J. Multivar. Anal. 150, 1–13 (2016)

Magnano, L., Boland, J.W., Hyndman, R.J.: Generation of synthetic sequences of half-hourly temperature. Environmetrics 19(8), 818–835 (2008)

McLean, M.W., Hooker, G., Staicu, A., Scheipl, F., Ruppert, D.: Functional generalized additive models. J. Comput. Graph. Stat. 23(1), 249–269 (2014)

Morris, J.: Functional regression. Annu. Rev. Stat. Appl. 2, 321–359 (2014)

Muller, H., Yao, F.: Functional additive models. J. Am. Stat. Assoc. 103, 1534–1544 (2008)

Muller, H., Wu, Y., Yao, F.: Continuously additive models for nonlinear functional regression. Biometrika 100(3), 607–622 (2013)

Olver, P.J.: Introduction to the calculus of variations (2019)

Preda, C.: Regression models for functional data by reproducing kernel Hilbert spaces methods. J. Stat. Plan. Inference 137(3), 829–840 (2007). https://doi.org/10.1016/j.jspi.2006.06.011. (Special Issue on Nonparametric Statistics and Related Topics: In honor of M.L. Puri)

Ramsay, J., Silverman, B.: Functional Data Analysis. Springer, Berlin (1997)

Rao, A.R., Reimherr, M.: Non-linear functional modeling using neural networks (2021)

Reimherr, M., Sriperumbudur, B., Taoufik, B.: Optimal prediction for additive function-on-function regression (2017)

Reiss, P., Goldsmith, J., Shang, H.L., Ogden, R.: Methods for scalar-on-function regression. Int. Stat. Rev. 85, 228–249 (2016)

Rossi, F., Conan-Guez, B.: Theoretical properties of projection based multi-layer perceptrons with functional inputs. Neural Process. Lett. 23, 55–70 (2006)

Rossi, F., Conan-Guez, B., Fleuret, F.: Functional data analysis with multi layer perceptrons. In: Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN’02 (cat. no.02ch37290), vol. 3, pp. 2843–2848 (2002)

Scheipl, F., Staicu, A.-M., Greven, S.: Functional additive mixed models. J. Comput. Graph. Stat. 24(2), 477–501 (2015)

Stoker, T.M.: Consistent estimation of scaled coefficients. Econometrica 54(6), 1461–1481 (1986)

Sun, X., Du, P., Wang, X., Ma, P.: Optimal penalized function-on-function regression under a reproducing kernel Hilbert space framework. J. Am. Stat. Assoc. 113(524), 1601–1611 (2018)

Sun, Y., Wang, Q.: Function-on-function quadratic regression models. Comput. Stat. Data Anal. 142, 106814 (2020)

Wang, X., Ruppert, D.: Optimal prediction in an additive functional model. Stat. Sin. 2, 567–589 (2013)

Wang, Q., Zheng, S., Farahat, A., Serita, S., Gupta, C.: Remaining useful life estimation using functional data analysis (2019a)

Wang, Q., Zheng, S., Farahat, A., Serita, S., Saeki, T., Gupta, C.: Multi-layer perceptron for sparse functional data. In: 2019 International joint conference on neural networks (IJCNN), pp. 1–10 (2019b)

Wang, Q., Wang, H., Gupta, C., Rao, A.R., Khorasgani, H.: A non-linear function-on-function model for regression with time series data (2020)

Yao, F., Muller, H.: Functional quadratic regression. Biometrika 97(1), 49–64 (2010)

Zhu, H., Yao, F., Zhang, H.: Structured functional additive regression in reproducing kernel Hilbert spaces. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 76, 581–603 (2013)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.