Neural Basis Models for Interpretability

Filip Radenovic
Meta AI

Abhimanyu Dubey
Meta AI

Dhruv Mahajan
Meta AI

Abstract

Due to the widespread use of complex machine learning models in real-world applications, it is becoming critical to explain model predictions. However, these models are typically black-box deep neural networks, explained post-hoc via methods with known faithfulness limitations. Generalized Additive Models (GAMs) are an inherently interpretable class of models that address this limitation by learning a non-linear shape function for each feature separately, followed by a linear model on top. However, these models are typically difficult to train, require numerous parameters, and are difficult to scale. We propose an entirely new subfamily of GAMs that utilizes basis decomposition of shape functions. A small number of basis functions are shared among all features, and are learned jointly for a given task, thus making our model scale much better to large-scale data with high-dimensional features, especially when features are sparse. We propose an architecture denoted as the Neural Basis Model (NBM) which uses a single neural network to learn these bases. On a variety of tabular and image datasets, we demonstrate that for interpretable machine learning, NBMs are the state-of-the-art in accuracy, model size, and, throughput and can easily model all higher-order feature interactions. Source code is available at `github.com/facebookresearch/nbm-spam`.

1 Introduction

Real world machine learning models [14, 61] are mostly used as a *black-box*, *i.e.*, it is very difficult to analyze and understand why a specific prediction was made. In order to *explain* such black-box models, an instance-specific local interpretable model is often learned [38, 50]. However, these approaches tend to be unstable and unfaithful [1, 52], *i.e.*, they often misrepresent the model's behavior. On the other hand, a family of models known as generalized additive models (GAMs) [24] have been used for decades as an inherently interpretable alternative to black-box models.

GAMs learn a *shape function* for each feature independently, and outputs of such functions are added (with a bias term) to obtain the final model prediction. All models from this family share an important trait: the impact of any specific feature on the prediction does not rely on the other features, and can be understood by visualizing its corresponding shape function. Original GAMs [24] were fitted using splines, which have since been improved in explainable boosting machines (EBMs) [36] by fitting boosted decision trees, or very recently in neural additive models (NAMs) [2] by fitting deep neural networks (DNNs). A drawback for all the aforementioned approaches is that for each shape function, they require either millions of decision trees [36], or a DNN with tens of thousands of parameters [2], making them prohibitively expensive for learning datasets with a large number of features.

In this work, we propose a novel subfamily of GAMs, which, unlike previous approaches, learn to decompose each feature's shape function into a small set of basis functions *shared* across all features. The shape functions are fitted as the feature-specific linear combination of these shared bases, see Figure 1. At an abstract level, our approach is motivated by signal decomposition using traditional basis functions like the Fourier basis [9] or Legendre polynomials [45], where a weighted

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

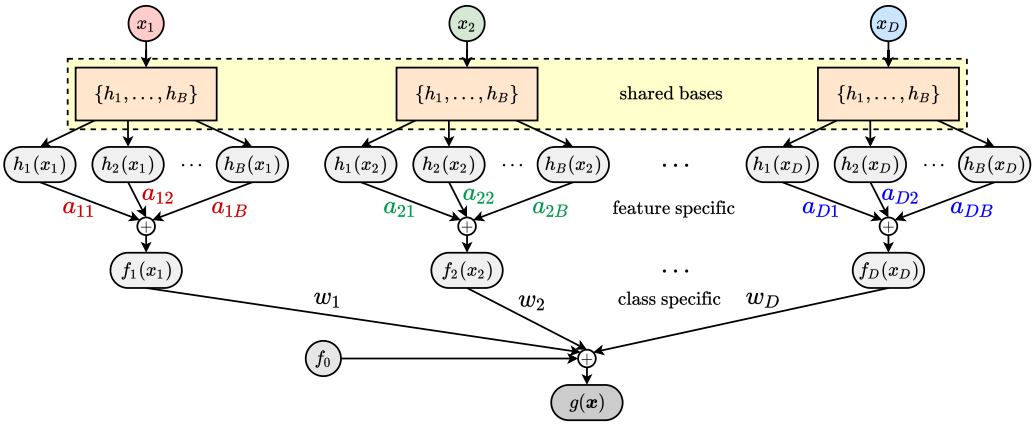

Figure 1: Neural Basis Model (NBM) architecture for a binary classification task.

combination of a few basis functions suffice to reconstruct complex signal shapes. However, in contrast to these approaches, our basis decomposition is not fixed *a priori*. In fact, it is learnt specifically for the prediction task. Consequently, we maintain the most important feature of GAMs, *i.e.*, their interpretability, as the contribution of single feature does not depend on the other features. At the same time, we gain scalability, as the number of basis functions needed in practice is much smaller than the number of input features. Moreover, we show that the usage of basis functions can increase computational efficiency by several orders of magnitude when the input features are sparse. Additionally, we propose an approach to learning the basis functions using a single DNN. We call this solution the Neural Basis Model (NBM). Using neural networks allows for even higher scalability, as training and inference are performed on GPUs or other specialized hardware, and can be easily implemented in any deep learning framework using standard, already developed, building blocks.

Our contributions are as follows: (i) We propose a novel subfamily of GAMs whose shape functions are composed of shared basis functions, and propose an approach to learn basis functions via DNNs, denoted as Neural Basis Model (NBM). This architecture is suitable for mini-batch gradient descent training on GPUs and easy to plug-in into any deep learning framework. (ii) We demonstrate that NBMs can be easily extended to incorporate pairwise functions, similar to GA^{2}Ms [37], by learning another set of bases to model the higher order interactions. This approach effectively only linearly increases the parameters, while other models such as EB^{2}Ms [36, 37] and NA^{2}Ms [2] suffer from quadratic growth of parameters, and often require heuristics and repeated training to select the most important interactions before learning [37]. (iii) Through extensive evaluation of regression, binary classification, and multi-class classification, with both tabular and computer vision datasets, we show that NBMs and NB^{2}Ms outperform state-of-the-art GAMs and GA^{2}Ms, while scaling significantly better, *i.e.*, fitting much fewer parameters and having higher throughput. For datasets with more than ten features, using NBMs result in around 5×–50× reduction in parameters over NAMs [2], and 4×–7× better throughput. (iv) We propose an efficient extension of NBMs to sparse datasets with more than a hundred thousand features, where other GAMs do not scale at all.

2 Related work

Shape functions in GAMs [24, 64] have many different representations in the literature, including: splines [24], trees or random forests [36], deep neural networks [2], neural oblivious decision trees [12]. Popular methods of fitting GAMs are: backfitting [24], gradient boosting [36], or mini-batch gradient descent [2, 12]. Our work falls under the GAM umbrella, however, it differs from these approaches by not learning shape functions independently, but rather, learning a set of shared basis functions that are used to compose each shape function. The bases themselves can be complex non-linear functions that operate on one feature at a time, *e.g.*, decision trees or splines, however, to keep the scope of the work concise and to ensure straightforward scalability, we represent them with deep neural networks and use mini-batch stochastic gradient descent to learn the bases. This makes our work most closely related to neural additive models (NAMs) [2], however, crucially, our method learns far fewer parameters by sharing bases compared to NAMs.

Methods have been proposed to model pairwise interactions [12, 37, 58], and they are commonly denoted as GA^{2}Ms. However, they usually require sophisticated feature selection using: heuristics [37], back-propagation [12], or, two-stage iterative training [58]. We note that one of the main goals of our work is to analyze the scalability of the newly proposed NBM architecture in extreme scenarios, hence, we do not apply feature selection algorithms. That being said, before-mentioned algorithms are complementary and can be applied to NB^{2}Ms as well.

Finally, GAMs are a popular choice in high-risk applications for healthcare [11, 56], finance [6], forensics [55], *etc.* Another line of work applicable to these domains are interpretable surrogate models such as LIME [50], SHAP [38], tree-based surrogates [5, 57], that give a *post-hoc* explanation of a high-complexity model. However, there are almost no theoretical guarantees that the simple surrogate model is highly representative of the more complex model [1]. This issue is completely resolved when using inherently transparent models such as NBMs. We hope that our approach sparks wider usage of GAMs in mission-critical applications with large-scale data, where the ability to interpret, understand, and correct the model is of utmost importance.

3 Method

3.1 Background

Generalized Additive Model (GAM) [24]. Given a D-dimensional interpretable input $\boldsymbol{x}=\{x_i\}_{i=1}^{D}$, $\boldsymbol{x} \in \mathbb{R}^D$, a target label y, a link function g (*e.g.*, logistic function), a univariate shape function f_i, corresponding to the input feature x_i, a bivariate shape function f_{ij}, corresponding to the feature interaction, and a bias term f_0, the prediction function in GAM and GA^{2}M is expressed as

$$\textbf{GAM} : g(\boldsymbol{x}) = f_0 + \sum_{i=1}^{D} f_i(x_i); \qquad \textbf{GA}^2\textbf{M} : g(\boldsymbol{x}) = f_0 + \sum_{i=1}^{D} f_i(x_i) + \sum_{i=1}^{D}\sum_{j>i}^{D} f_{ij}(x_i, x_j). \quad (1)$$

Interpreting GAMs is straightforward as the impact of a feature on the prediction does not rely on the other features, and can be understood by visualizing its corresponding shape function, *e.g.*, plotting x_i on the x-axis and $f_i(x_i)$ on the y-axis. A certain level of interpretability is sacrificed for accuracy when modeling interactions, as f_{ij} shape functions are harder to visualize. Shape function visualization through heatmaps [12, 37] is commonly used towards that purpose. Note that, the graphs visualizations of GAMs are an exact description of how GAMs compute a prediction.

3.2 Our model architecture

We observe that, typically, input features of high-dimensional data are correlated with each other. As a result, it should be possible to decompose each shape function f_i into a small set of basis functions shared among all the features. This is the core idea behind our approach.

Neural Basis Model (NBM). We propose to represent shape functions f_i as

$$f_i(x_i) = \sum_{k=1}^{B} h_k(x_i) a_{ik}; \quad (2)$$

where $\{h_1, h_2, ..., h_B\}$ represents a set of B shared basis functions that are independent of feature indices, and coefficients a_{ik} are the projection of each feature to the shared bases. We additionally propose to learn basis functions using a DNN, *i.e.*, a single *one*-input B-output multi-layer perceptron (MLP) for all $\{h_k; k = 1, \ldots, B\}$. The resulting architecture is shown in Figure 1.

Multi-class / multi-task architecture. Let l correspond to the target class y_l in the multi-class setting. Similar to Equation 1, the prediction function g_l for class y_l in GAMs can be written as:

$$g_l(\boldsymbol{x}) = f_{0l} + \sum_{i=1}^{D} f_i(x_i) w_{il}, \quad (3)$$

where feature shape functions $f_i(x_i)$ are shared among the classes and are linearly combined using class specific weights w_{il}. Combining Equations 2 and 3, multi-class NBM can be represented as:

$$\textbf{Multi-class NBM} : g_l(\boldsymbol{x}) = f_{0l} + \sum_{i=1}^{D}\sum_{k=1}^{B} h_k(x_i) a_{ik} w_{il}. \quad (4)$$

Extension to NB^{2}M. Similar to NBM, we represent GA^{2}M shape functions f_{ij} in Equation 1 as:

$$f_{ij}(x_i, x_j) = \sum_{k=1}^{B} u_k(x_i, x_j) b_{ijk}; \tag{5}$$

where $\{u_1, u_2, ..., u_B\}$ represents a set of B shared bi-variate basis functions that are independent of feature indices and coefficients b_{ijk} are the projection of pair-wise features to the shared bases. We learn an additional *two*-input B-output MLP for all $\{u_k; k = 1, \ldots, B\}$ to learn the bases. Extension to multi-class setting can be done in the same way as for NBMs.

Sparse architecture. Typically, datasets with high-dimensional features are sparse in nature. For example, in the Newsgroups dataset [32], news articles are represented by *tf-idf* features, and, for a given instance, most of the features are absent due to the vocabulary being of the order of 100K words. Since NBM uses a single DNN to learn all the bases, we can simply append the single value representing the absent feature to the batch, to compute the corresponding basis function values. The subsequent linear projection to feature indices via a_{ik} is a computationally inexpensive operation.

In contrast, typical GAMs (*e.g.*, Neural Additive Model (NAM) [42]) need to pass the absent value through every shape function f_i which makes it compute-intensive as well as difficult to implement.

Training and regularization. We use mean squared error (MSE) for regression, and cross-entropy loss for classification. To avoid overfitting, we use the following regularization techniques: (i) L_2-normalization (weight decay) [31] of parameters; (ii) batch-norm [28] and dropout [54] on hidden layers of the basis functions network; (iii) an L_2-normalization penalty on the outputs f_i to incentivize fewer strong feature contributions, as done in [2]; (iv) basis dropout to randomly drop individual basis functions in order to decorrelate them. Similar techniques have been used for other GAMs [2, 12].

Selecting the number of bases. One can use the theory of Reproducing Hilbert Kernel Spaces (RKHS, [7]) to devise a heuristic for selecting the number of bases B. Specifically, we demonstrate that any NBM model lies on a subspace within the space spanned by a complete GAM if the GAM shape functions reside within a ball in an RKHS. Assuming a regularity property in the data distribution, one can then demonstrate that $B = \mathcal{O}(\log D)$ bases are sufficient to obtain competitive performance. We present this formally in Appendix Sec. A.5. This provides the alternate interpretation of NBM as learning a "principal components" decomposition in the L_2-space of functions, as we learn a set of (preferably orthogonal) basis functions to approximate the decision boundary.

3.3 Discussion

In this section, we contrast NBMs with closely related GAMs: Neural Additive Models (NAMs) [2].

Neural Additive Model (NAM) [2]. NAMs learn a linear combination of networks that each attend to a single input feature: each f_i in (1) is parametrized by a deep neural network, *i.e.*, a *one*-input *one*-output multi-layer perceptron (MLP). These MLPs are trained jointly using backpropagation and can learn arbitrarily complex shape functions.

Number of parameters. We compare number of weight parameters needed to learn NAM *vs.* NBM for the binary-classification task. See Appendix Section A.3 for multi-class analysis. Let us denote with M the number of parameters in MLP for each feature in NAM, and with N the number of parameters in MLP for bases in NBM. In most experiments the optimal NAM has 3 hidden layers with 64, 64 and 32 units ($M = 6401$), and, NBM has 3 hidden layers with 256, 128, 128 units ($N = 62820$) and $B = 100$ basis functions. Then the ratio of number of parameters in NAM *vs.* NBM is given by,

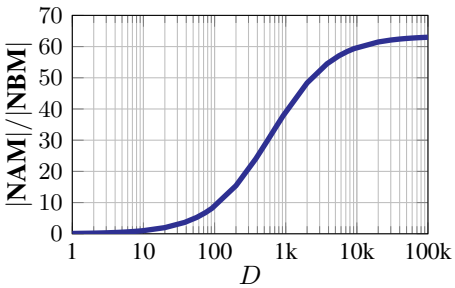

Figure 2: NAM *vs.* NBM: #parameters.

$$\frac{|\textbf{NAM}|}{|\textbf{NBM}|} = \frac{D \cdot M + D}{N + D \cdot B + D} = \frac{6402}{\frac{62820}{D} + 101}. \tag{6}$$

Figure 2 shows this ratio for different values of feature dimensionality D. For $D = 10$, NBMs and NAMs have roughly equal number of parameters. For most textual and vision datasets, feature dimensionality is significantly higher, thus leading to $10\times$–$50\times$ reduction in parameters. We also observe that as a result, for many datasets, specialized regularization discussed in the previous section gives incremental gains for NBMs, while they are very crucial for NAMs to give good performance. Additional analysis on number of parameters is given in Section 4.4 and Table 2.

Throughput. One of the main challenges in approach like NAMs is the low throughput rate, that is the number of data instances processed per second, which directly affects the training speed. Since NAMs use separate MLP per dimension, it is much more challenging to implement efficiently. NBMs on the other hand are much more efficient since feature specific linear layer on the top of bases is very fast. For example, for datasets with around hundred features, the original NAM implementation [2] is around $20\times$ slower in training compared to NBMs. We optimized the speed of NAMs using group convolutions (see Appendix Section A.3), but even this optimized version is around $5\times$ slower.

Stability. The interpretability of models and their explanations is tightly coupled to their stability. For example, post-hoc explanations of black-box models are known to be unstable, and produce different explanations for small input changes [53]. GAMs are exactly explained by visualizing shape functions that represent the model, however, a desirable property is to have similar shape functions when repeatedly training the model on the same data while varying random initialization. Because of the over-parameterization in NAMs, we observe that shape functions of different runs are often unstable (*i.e.* they often diverge), especially for feature regions where the data density is low. On the other hand, NBMs train a single set of shared bases for the entire dataset, which makes shape functions significantly more stable, see Section 4.6 and Figure 3.

NA^{2}M vs. NB^{2}M. NAMs can trivially be extended to NA^{2}Ms by learning additional *two*-input *one*-output MLPs for each pairwise feature interaction f_{ij}. Since the numbers of parameters grow quadratically, this setting further exaggerates the parameters and throughput discrepancies. In fact, as we show in Section 4.4, for high dimensional datasets the NA^{2}M approach does not scale at all.

4 Experiments

4.1 Datasets

Tabular datasets. We report performance on **CA Housing** [10, 46], **FICO** [22], **CoverType** [8, 16, 20], and **Newsgroups** [32, 43] tabular datasets. We perform one-hot encoding for categorical features, and min-max scaling of features to $[0, 1]$ range. Data is split to have $70/10/20$ ratio for training, validation, and, testing, respectively; except for Newsgroups where test split is fixed, so we only split the train part to $85/15$ ratio for train and validation.

We also report performance on **MIMIC-II** [41, 51], **Credit** [17, 19], **Click** [15], **Epsilon** [21], **Higgs** [3, 26], **Microsoft** [40, 49], **Yahoo** [60], and **Year** [66] tabular datasets. For these datasets, we follow [12, 47] to use the same training, validation, and, testing splits, and to perform the same feature normalization: target encoding for categorical features, quantile transformation with 2000 bins for all features to Gaussian distribution.

Additional details for all tabular datasets are given in Table 1 and Appendix Section A.1.

Image datasets (classification). We experiment with two bird classification datasets: **CUB** [18, 62] and **iNaturalist Birds** [27, 59]. CUB images are annotated with keypoint locations of 15 bird parts, and each location is associated with one or more part-attribute labels. iNaturalist Birds contains significantly more bird classes and more challenging scenes compared to CUB, but lacks keypoint annotations. We perform the concept-bottleneck-style [29] interpretable pre-processing: (i) Images are randomly cropped and resized to 448×448 size, and passed through a ResNet50 model [25] until the last pooling layer to extract 2048-D features on the 14×14 spatial grid. (ii) Part-attribute linear classifiers are trained on the extracted features using part locations and part-attribute labels from CUB training split. (iii) Max-pooling of part-attribute scores over spatial locations is performed to extract 278 interpretable features (*e.g.*, orange legs, striped wings, needle-shaped beak) for each image in both CUB and iNaturalist, using the part-attribute classifiers trained on CUB only. Splits are preset and results are reported on the validation split, which is a common practice in the computer vision community. Additional details are given in Table 1 and Appendix Section A.1.

Table 1: Datasets overview.

Dataset	Task	#Train	#Val	#Test	Sparse	#Feat	#Class
Tabular dataset							
CA Housing	Regression	14,447	2,065	4,128	No	8	–
FICO	Binary	7,321	1,046	2,092	No	39	2
CoverType	Multi-class	406,707	58,102	116,203	No	54	7
Newsgroups	Multi-class	9,899	1,415	7,532	99.9%	146,016	20
MIMIC-II	Binary	17,155	2,451	4,902	No	17	2
Credit	Binary	199,364	28,481	56,962	No	30	2
Click	Binary	800,000	100,000	100,000	No	11	2
Epsilon	Binary	320,000	80,000	100,000	No	2,000	2
Higgs	Binary	8,400,000	2,100,000	500,000	No	28	2
Microsoft	Regression	580,539	142,873	241,521	No	136	–
Yahoo	Regression	473,134	71,083	165,660	No	699	–
Year	Regression	370,972	92,743	51,630	No	90	–
Image dataset							
CUB	Classification	5,994	5,794	–	No	278	200
iNaturalist Birds	Classification	414,847	14,860	–	No	278	1,486
Common Objects	Detection	2,645,488	58,525	–	97.0%	2,618	115

Image dataset (object detection). For this task we use a proprietary object detection dataset, denoted as **Common Objects**, that contains 114 common objects, (plus a background class) with bounding box locations, 200 parts and 54 attributes. Each box for each image is pre-processed using compositions of parts and attributes to extract 2,618 interpretable features and 100k pairwise feature interactions. For the purpose of evaluating models in this paper, one data instance in this dataset is equivalent to one bounding box. As with previous computer vision datasets, results are reported on the validation split. Additional details are given in Table 1 and Appendix Section A.1.

4.2 Baselines

We implement the following baselines in PyTorch [48], and train using mini-batch gradient descent:

Linear. Linear / logistic regression are interpretable models that make a prediction decision based on the value of a linear combination of the features.

NAM [2]. We experiment with two proposed NAM architectures [2]: (i) MLPs containing 3 hidden layers with 64, 64 and 32 units and ReLU [23] activation, and (ii) single hidden layer MLPs with 1,024 ExU units and ReLU-1 activation. We reimplement the original NAM implementation [42] (details in Appendix Section A.3) achieving around $\times 2$–$\times 10$ speedup at training, depending on the dataset. Finally, we extend the implementation to NA^{2}Ms, as well. We released our NAM implementation together with the rest of our code at `github.com/facebookresearch/nbm-spam`.

MLP. Multi-layer perceptron (MLP) is a non-interpretable black-box model capturing high-order interaction between the features: $g(\boldsymbol{x})=f(x_1, x_2, \ldots, x_D)$. For most datasets, MLP sets the upper bound on the performance, and gives an idea of the trade-off between accuracy and interpretability. We experiment with the following architectures: (i) 5 hidden layers with 128, 128, 64, 64, 64 units; (ii) 3 hidden layers with 1024, 512, 512 units; (iii) 2 hidden layers with 2048, 1024 units. We have observed that increasing the depth by adding more layers for any of the three architectures has no additional accuracy gain. For all datasets, we report the best performing architecture across the three. For the following baselines, we use the available implementations:

EBM [36, 37]. Explainable Boosting Machines (EBMs) are another state-of-the-art GAM which use gradient boosting of millions of shallow trees to learn a shape function for each input feature. These models support automatic pairwise interactions through their EB^{2}M implementation, but only for regression and binary classification tasks. We use the `interpretml` library [44].

XGBoost [13]. EXtreme Gradient Boosted trees (XGBoost) are another non-interpretable black-box model that learn high-order feature interactions. We use the `xgboost` library [13].

Table 2: Number of parameters (#par.) and throughput as inputs per second ($\boldsymbol{x}$/sec). Here NAM and NA^{2}M refers to our optimized implementation; †refers to sparse NBM optimization.

Model	CA Housing		FICO		CoverType		Newsgroups		iNat. Birds	
	#par.	$\boldsymbol{x}$/sec	#par.	$\boldsymbol{x}$/sec	#par.	$\boldsymbol{x}$/sec	#par.	$\boldsymbol{x}$/sec	#par.	$\boldsymbol{x}$/sec
NAM	54K	0.5M	262K	123K	363K	80K	984M	23	2.3M	15K
NBM	65K	3.4M$\times$6.8	68K	821K$\times$6.7	70K	530K$\times$6.6	18M	†9K$\times$391	0.5M	74K$\times$4.9
NA^{2}M	243K	119K	5.3M	6K	10M	3K	–	–	320M	99
NB^{2}M	161K	641K$\times$5.4	0.3M	30K$\times$5.0	0.5M	15K$\times$5.0	–	–	66M	374$\times$3.8

4.3 Implementation details

NBM. We use the following architecture for NBMs and NB^{2}Ms: MLP containing 3 hidden layers with 256, 128, and 128 units, ReLU [23], $B = 100$ basis outputs for NBMs and $B = 200$ for NB^{2}Ms. Source code is available at `github.com/facebookresearch/nbm-spam`.

Training details. Linear, NAM, NBM, and MLP models are trained using the Adam with decoupled weight decay (AdamW) optimizer [35], on 8$\times$V100 GPU machines with 32 GB memory, and a batch size of at most 1024 per GPU (divided by 2 every time a batch cannot fit in the memory). We train for 1,000, 500, 100, or, 50 epochs, depending on the size and feature dimensionality of the dataset. The learning rate is decayed with cosine annealing [34] from the starting value until zero. For NBMs on all datasets, we tune the starting learning rate in the continuous interval $[1\text{e}{-5}, 1.0)$, weight decay in the interval $[1\text{e}{-10}, 1.0)$, output penalty coefficient in the interval $[1\text{e}{-7}, 100)$, dropout and basis dropout coefficients in the discrete set $\{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. We find optimal hyper-parameters using validation set and random search. Similar hyper-parameter search is performed for Linear, NAM, and MLP baselines. Finally, for EBMs and XGBoost, CPU machines are used, with hyper-parameter search using the guidelines in the original works. See Appendix Section A.2 for more training and hyper-parameter search details.

Evaluation details. Performance metrics: (i) for regression we report mean-squared error (MSE) or root MSE (RMSE); (ii) for binary classification we report area under the ROC curve (AUROC) or error rate (Error); (iii) for multi-class classification in tabular and image domains we report accuracy@1 (acc@1); and, (iv) for object detection in image domain we report mean average precision (mAP) averaged over IoU thresholds from 0.5 to 0.9 as per MS-COCO [33] definitions. We report average performance and standard deviation over 10 runs with different random seeds.

4.4 Comparison with baselines

In this section we compare NBM, as well as the pairwise interactions NB^{2}M counterpart, with popular and widely used GAMs and non-interpretable black-box models.

First, we perform extensive comparison on the number of parameters and throughput against the most similar architecture, *i.e.* Neural Additive Model (NAM) [2]. Both NAMs and NBMs utilize deep networks and are able to run on GPUs, which helps to scale the models on large datasets. Results on representative datasets are presented in Table 2. The throughput is measured as the number of input instances that we can process per second ($\boldsymbol{x}$ / sec) on one 32 GB V100 GPU, in inference mode. For each model, we take the largest batch size (up to 8,192) that fits on the GPU and calculate the average time over 100 runs to process that batch. With that, we calculate the number of input instances processed per second. Throughput can vary depending on the implementation, library used, *etc.* Hence, in order to be as fair as possible, we compare both models with our own optimized implementation in the same deep learning library, *i.e.*, PyTorch [48]. The only dataset where NAM narrowly beats NBM in the number of parameters is CA Housing, which has the smallest number of input features $D = 8$. On other datasets, NBM has around 5$\times$–50$\times$ fewer parameters than NAM, while having 4$\times$–7$\times$ smaller runtime. Finally, only the sparse version of NBM model can efficiently run on Newsgroups and Common Objects (with pairwise interactions) datasets, where standard NAMs (and other GAMs) have throughput too low for any practical application. Note that these comparisons are with our optimized implementation of NAM.

Table 3: Performance comparison with baselines. ↓: lower is better; ↑: higher is better.

Model	CA Housing	FICO	CoverType	News.	CUB	iNat. Birds	Common Objects
	RMSE ↓	AUROC ↑	acc@1 ↑	acc@1 ↑	acc@1 ↑	acc@1 ↑	mAP ↑
Linear	0.7354 ±0.0004	0.7909 ±0.0001	0.7254 ±0.0000	0.8238 ±0.0005	0.7451 ±0.0003	0.3932 ±0.0004	0.1917 ±0.0001
EBM	0.5586 ±0.0002	0.7985 ±0.0001	0.7392 ±0.0004	—	—	—	—
NAM	0.5721 ±0.0054	0.7993 ±0.0004	0.7359 ±0.0003	—	0.7632 ±0.0010	0.4194 ±0.0007	0.2056 ±0.0005
NBM	0.5638 ±0.0013	0.8048 ±0.0005	0.7369 ±0.0002	0.8446 ±0.0012	0.7683 ±0.0007	0.4227 ±0.0007	0.2168 ±0.0006
EB^{2}M	0.4919 ±0.0004	0.7998 ±0.0005	—	—	—	—	—
NA^{2}M	0.4921 ±0.0078	0.7992 ±0.0003	0.8872 ±0.0006	—	0.7713 ±0.0011	0.4591 ±0.0010	—
NB^{2}M	0.4779 ±0.0020	0.8029 ±0.0003	0.8908 ±0.0008	—	0.7770 ±0.0006	0.4684 ±0.0009	0.2378 ±0.0006
XGBoost	0.4428 ±0.0006	0.7925 ±0.0008	0.8860 ±0.0003	0.7677 ±0.0009	0.7186 ±0.0008	—	—
MLP	0.5014 ±0.0061	0.7936 ±0.0013	0.9694 ±0.0002	0.8494 ±0.0021	0.7684 ±0.0007	0.4584 ±0.0008	0.2376 ±0.0007

Next, we compare performance with GAM baselines, and non-interpretable black-box models in Table 3. We notice that NBMs achieve the best performance among all GAMs, even outperforming full complexity MLPs on several datasets. The difference in performance is more pronounced on larger datasets, such as iNaturalist and Common Objects, where NBMs generalize better. EBMs have a big downside as they do not support pairwise interactions on multi-class tasks, and do not scale at all to very large datasets. Finally, on datasets with large number of features, *i.e.* Newsgroups and Common Objects (with pairwise interactions), NBMs are the only GAMs that scale, as we did not manage to complete a successful training with NAMs or EBMs.

We finally observe that pairwise interactions are enough to match or beat black-box models on 6 out of 7 datasets. The only exception is CoverType, where MLP has 0.9694 acc@1 *vs.* NB^{2}M with 0.8908. However, we scale NBMs further and train NB^{3}M (*i.e.*, including triplet interactions) that gets 0.9634 acc@1. Even though this is an interesting result that helps us understand the data behavior, we argue that triplet feature interactions are very hard to visualize and hence *not* interpretable.

4.5 Comparison with state of the art

We finally compare NBM, as well as the pairwise interactions NB^{2}M counterpart, with neural-based state-of-the-art GAMs. Namely, we compare with Neural Additive Model (NAM) [2] and Neural Oblivious Decision Trees GAM (NODE-GAM) [12], and their pairwise interactions versions NA^{2}M and NODE-GA^{2}M. Results are presented in Table 4.

We compute NAM and NBM results using our codebase, while NODE-GAM results are reproduced from [12]. For a fair comparison on these datasets, we use training, validation, and, testing splits from NODE-GAM [12], and perform the same feature normalization, see Section 4.1 for details. Without any additional model selection (architecture tuning, stochastic weight averaging, *etc.*) NBMs outperform NODE-GAMs on 7 out of 8 presented datasets (albeit marginally), and outperform NAM on all datasets. Note that, the idea of NBM is perpendicular to that of NODE-GAM and it is possible that bigger gains can be achieved by combining them, especially for Epsilon and Yahoo datasets, where NB^{2}Ms do not scale due to a high dimensionality and non-sparse nature of features.

Table 4: Performance comparison with state-of-the-art GAMs. NODE-GAM results are reproduced from [12]. ↓: lower is better; ↑: higher is better. State-of-the-art performance shown in **bold**.

Model	MIMIC-II	Credit	Click	Epsilon	Higgs	Microsoft	Yahoo	Year
	AUROC ↑	AUROC ↑	Error ↓	Error ↓	Error ↓	MSE ↓	MSE ↓	MSE ↓
NAM	0.8539 ±0.0004	0.9766 ±0.0027	0.3317 ±0.0005	0.1079 ±0.0002	0.2972 ±0.0001	0.5824 ±0.0002	0.6093 ±0.0003	85.25 ±0.01
NODE GAM	0.8320 ±0.0110	0.9810 ±0.0110	0.3342 ±0.0001	0.1040 ±0.0003	0.2970 ±0.0001	0.5821 ±0.0004	0.6101 ±0.0006	**85.09** ±0.01
NBM	**0.8549** ±0.0004	**0.9829** ±0.0014	**0.3312** ±0.0002	**0.1038** ±0.0002	**0.2969** ±0.0001	**0.5817** ±0.0001	**0.6084** ±0.0001	85.10 ±0.01
NA^{2}M	0.8639 ±0.0011	0.9824 ±0.0032	0.3290 ±0.0005	—	0.2555 ±0.0003	0.5622 ±0.0003	—	79.80 ±0.05
NODE GA^{2}M	0.8460 ±0.0110	**0.9860** ±0.0100	0.3307 ±0.0001	0.1050 ±0.0002	0.2566 ±0.0003	**0.5618** ±0.0003	**0.5807** ±0.0004	79.57 ±0.12
NB^{2}M	**0.8690** ±0.0010	0.9856 ±0.0017	**0.3286** ±0.0002	—	**0.2545** ±0.0002	**0.5618** ±0.0002	—	**79.01** ±0.03

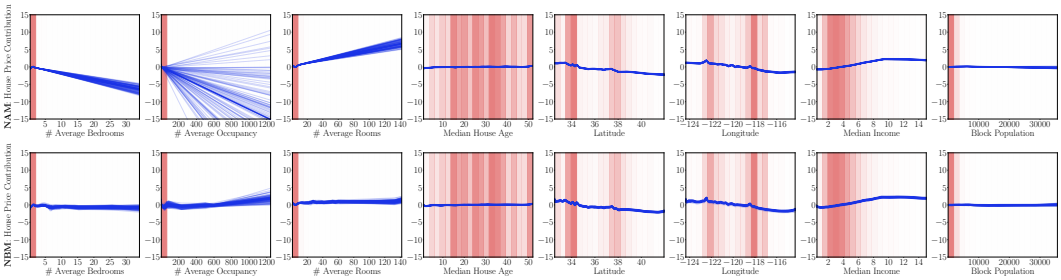

Figure 3: NAM (upper row) and NBM (bottom row) shape functions f_i for the CA Housing dataset.

4.6 Stability and interpretability

The interpretability of GAMs comes from the fact that the learned shape functions can be easily visualized. In the same manner as the other GAM approaches, each feature's importance in an NBM can be represented by a unique shape function that *exactly* describes how the NBM computes a prediction. We demonstrate this on the CA Housing dataset because it has a small number of input features ($D = 8$) and the full model can conveniently be visualized in a single row, see Figure 3.

Towards this purpose, we train an ensemble of 100 models by running different random seeds with optimal hyper-parameters, in order to analyze when the models learn the same shape and when they diverge. Following Agarwal et al. [2], we set the average score for each shape function to be zero by subtracting the respective mean feature score. Next, we plot each shape function as $f_i(x_i)$ *vs.* x_i for each model in the ensemble using a semi-transparent line, and an average ensemble shape function using a thick line. Finally, the x-axis is divided by bars depicting the normalized data density, *i.e.*, darker areas contain more data points. Figure 3 depicts an ensemble of NAMs [2] (upper row) and NBMs (bottom row). Although the shape functions are correlated for most features, we observe that NBMs diverge much less in the cases where there are only few data points (light / white areas in the graphs). This is due to the fact that the basis network in NBM is trained jointly with only linear composing weights being trained for each feature separately. In contrast, each feature in NAM has its own network, which becomes unstable and diverges in cases of few training data points.

Finally, to quantify stability, we compute the standard deviation for each visualized shape function over 100 models, and report the mean standard deviation over all features. For NAMs, this mean standard deviation is 0.9921, while for NBMs it is 0.1987.

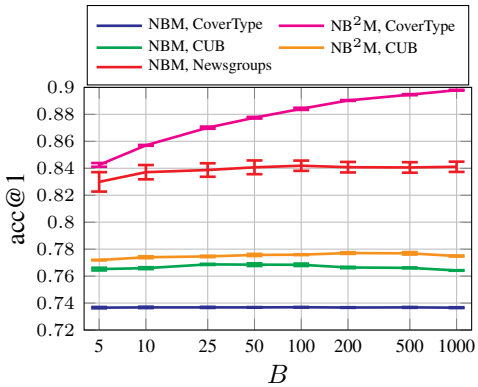

Figure 4: Ablation on the number of bases B.

Table 5: Ablation on the number of subnets S.

Model	S	CoverType		
		#param	$\boldsymbol{x}$/sec	acc@1
NAM	1	363K	80K	0.7359
NBM	1	70K	530K	0.7369
NAM	5	1.8M	16K	0.7417
NBM	5	350K	88K	0.7435
NB^{2}M	1	463K	15K	0.8908

4.7 Ablation study

Number of basis functions. We evaluate the robustness of NBMs *w.r.t.* the choice of the number of basis functions B. Results are presented in Figure 4 for Newsgroups, CoverType, and, CUB. We observe that NBMs are not overly sensitive to the choice of B. Rather than tuning this hyperparameter, we recommend setting $B = 100$ for NBMs and $B = 200$ for NB^{2}Ms as it performs well across a large variety of datasets we experimented with. Although, *e.g.*, for NB^{2}Ms on CoverType, using a larger number of basis functions leads to some performance gains, it comes with a throughput trade-off, as computing the linear combination of bases starts becoming the bottleneck. Thus, we suggest using the recommended values as a great trade-off between accuracy and throughput.

Multiple subnets. Agarwal et al. [2] propose to extend NAMs to multi-task / multi-class setups by associating each feature with multiple subnets. This method is directly applicable on NBMs as well, by using S different networks to learn S sets of basis functions. We compare few options on the multi-class CoverType predictions with the number of subnets S=1 and S=5 in Table 5. Both NAMs and NBMs have a similarly small relative accuracy improvement when using 5 subnets over the 1 subnet, however that comes at a 5$\times$ increase in the number of parameters and 5$\times$ lower throughput. With the accuracy-complexity trade-off in mind, we did not see much benefit of using $S = 5$ on our datasets, so we keep $S = 1$ across all other experiments. Interestingly, NB^{2}M with S=1 has a comparable throughput to NAM with S=5, while achieving an impressive accuracy gain, see Table 5.

5 Conclusion and future work

This work describes novel Neural Basis Models (NBMs), which belong to a new subfamily of Generalized Additive Models (GAMs), and utilize deep learning techniques to scale to large datasets and high-dimensional features. Our approach addresses several scalability and performance issues associated with GAMs, while still preserving their interpretability compared to black-box deep neural networks (DNNs). We show that our NBMs and NB^{2}Ms achieve state-of-the-art performance on a large variety of datasets and tasks, while being much smaller and faster than other neural-based GAMs. As a result, they can be used as a drop-in replacements for the black-box DNNs.

We do recognize that our approach, though highly scalable, has limitations *w.r.t.* number of features. Beyond 10,000 dense (or 1 million sparse) features, we would need to apply some form of feature selection [12, 37, 58] to scale further. Scalability issue is even more pronounced when modeling pairwise interactions in NB^{2}M. However, NBMs can still handle an order of magnitude more than what can be handled by NAM or other GAM approaches that do not perform feature selection.

There are many future directions for this line of work. First, for the computer vision domain, we assume an intermediate, interpretable concept layer [29] on which NBMs and in general GAMs can be applied. It would be interesting to explore visual interpretability by either directly going to the pixel space with NBMs or learning visual features that can do recognition with lower order interactions in NBM framework (for example, NB^{2}Ms). Finally, the idea of using shared basis is generic. Although, we used neural networks to learn these basis, we can enhance the model interpretability further for higher order interactions, by using more interpretable learning functions such as polynomials.

References

[1] A. Adadi and M. Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 2018.

[2] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton. Neural additive models: Interpretable machine learning with neural nets. In *NeurIPS*, 2021.

[3] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 2014.

[4] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 2002.

[5] O. Bastani, C. Kim, and H. Bastani. Interpretability via model extraction. In *FAT/ML*, 2017.

[6] D. Berg. Bankruptcy prediction by generalized additive models. *ASMBI*, 2007.

[7] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[8] J. A. Blackard and D. J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 1999.

[9] R. N. Bracewell. *The Fourier transform and its applications*. McGraw Hill, 1986.

[10] CA Housing. California Housing Dataset. https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing, 1997.

[11] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *SIGKDD*, 2015.

[12] C.-H. Chang, R. Caruana, and A. Goldenberg. NODE-GAM: Neural Generalized Additive Model for Interpretable Deep Learning. In *ICLR*, 2022.

[13] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *SIGKDD*, 2016.

[14] G. Chowdhary. Natural language processing. *Fundamentals of Artificial Intelligence*, 2020.

[15] Click. KDD Cup 2012, Track 2. https://www.kaggle.com/c/kddcup2012-track2, 2012.

[16] CoverType. Forest CoverType Dataset. https://archive.ics.uci.edu/ml/datasets/Covertype, 1999.

[17] Credit. Credit Card Fraud Detection. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud, 2015.

[18] CUB. Caltech-UCSD Birds-200-2011 Dataset. https://www.vision.caltech.edu/datasets/cub_200_2011, 2011.

[19] A. Dal Pozzolo. Adaptive machine learning for credit card fraud detection. *PhD Thesis, Universite libre de Bruxelles*, 2015.

[20] D. Dua and C. Graff. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml, 2019.

[21] Epsilon. Large Scale Learning Challenge. https://www.k4all.org/project/large-scale-learning-challenge, 2008.

[22] FICO HELOC. FICO Explainable Machine Learning Challenge. https://community.fico.com/s/explainable-machine-learning-challenge, 2018.

[23] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *AISTATS*, 2011.

[24] T. J. Hastie and R. J. Tibshirani. Generalized additive models. *Statistical Science*, 1986.

[25] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[26] Higgs. HIGGS Dataset. https://archive.ics.uci.edu/ml/datasets/HIGGS, 2014.

[27] iNaturalist. iNaturalist 2021 Competition. https://github.com/visipedia/inat_comp/tree/master/2021, 2021.

[28] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[29] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. Concept bottleneck models. In *ICML*, 2020.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[31] A. Krogh and J. Hertz. A simple weight decay can improve generalization. In *NeurIPS*, 1991.

[32] K. Lang. Newsweeder: Learning to filter netnews. In *ICML*, 1995.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[34] I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2016.

[35] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[36] Y. Lou, R. Caruana, and J. Gehrke. Intelligible models for classification and regression. In *SIGKDD*, 2012.

[37] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate intelligible models with pairwise interactions. In *SIGKDD*, 2013.

[38] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

[39] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society*, 1909.

[40] Microsoft. Microsoft Learning to Rank Dataset. https://www.microsoft.com/en-us/research/project/mslr, 2010.

[41] MIMIC-II. Multiparameter Intelligent Monitoring in Intensive Care II. https://archive.physionet.org/mimic2, 2011.

[42] NAMs. Neural Additive Models. https://github.com/google-research/google-research/tree/master/neural_additive_models, 2021.

[43] Newsgroups. The 20 Newsgroups Dataset. http://qwone.com/~jason/20Newsgroups, 1995.

[44] H. Nori, S. Jenkins, P. Koch, and R. Caruana. InterpretML: A unified framework for machine learning interpretability. In *arXiv:1909.09223*, 2019.

[45] F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark. *NIST handbook of mathematical functions*. Cambridge University Press, 2010.

[46] R. K. Pace and R. Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 1997.

[47] S. Popov, S. Morozov, and A. Babenko. Neural oblivious decision ensembles for deep learning on tabular data. In *ICLR*, 2020.

[48] PyTorch. PyTorch: From research to production. https://pytorch.org, 2021.

[49] T. Qin and T.-Y. Liu. Introducing LETOR 4.0 datasets. In *arXiv:1306.2597*, 2013.

[50] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *SIGKDD*, 2016.

[51] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. *Critical care medicine*, 2011.

[52] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *AIES*, 2020.

[53] D. Slack, A. Hilgard, S. Singh, and H. Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. In *NeurIPS*, 2021.

[54] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[55] A. M. Tarone and D. R. Foran. Generalized additive models and Lucilia sericata growth: assessing confidence intervals and error rates in forensic entomology. *JFS*, 2008.

[56] B. Thelen, N. H. French, B. W. Koziol, M. Billmire, R. C. Owen, J. Johnson, M. Ginsberg, T. Loboda, and S. Wu. Modeling acute respiratory illness during the 2007 San Diego wildland fires using a coupled emissions-transport system and generalized additive modeling. *Environmental Health*, 2013.

[57] J. J. Thiagarajan, B. Kailkhura, P. Sattigeri, and K. N. Ramamurthy. TreeView: Peeking into deep neural networks via feature-space partitioning. In *NeurIPS*, 2016.

[58] M. Tsang, H. Liu, S. Purushotham, P. Murali, and Y. Liu. Neural interaction transparency (NIT): Disentangling learned interactions for improved interpretability. In *NeurIPS*, 2018.

[59] G. Van Horn, E. Cole, S. Beery, K. Wilber, S. Belongie, and O. Mac Aodha. Benchmarking representation learning for natural world image collections. In *CVPR*, 2021.

[60] Verizon Media Webscope. Yahoo! Learning to Rank Challenge. https://webscope.sandbox.yahoo.com/catalog.php?datatype=c, 2010.

[61] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis. Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018.

[62] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. *California Institute of Technology*, 2011.

[63] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

[64] S. N. Wood. *Generalized additive models: an introduction with R*. CRC, 2006.

[65] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[66] Year. Year Prediction MSD Dataset. https://archive.ics.uci.edu/ml/datasets/yearpredictionmsd, 2011.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sec. A.5.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Sec. A.5.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] No new assets except code, see 3(a).
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Dataset descriptions

CA Housing [10, 46]. The target to **regress** is the median house value for California districts (1990 U.S. census), expressed in hundreds of thousands of dollars, from 8 household-related variables.

FICO [22]. Anonymized dataset of line of credit applications made by real homeowners. The customers in this dataset have requested a credit line in the range of \$5,000 – \$150,000. The task is to make a **binary prediction** whether applicants will repay their account within 2 years.

CoverType [8, 16, 20]. The samples in this dataset correspond to 30×30m patches of forest in the U.S., collected for the task of predicting each patch's cover type, *i.e.*, the dominant species of tree. There are seven covertypes, making this a **multi-class classification** problem.

Newsgroups [32, 43]. This dataset is a collection of news documents, partitioned across 20 different newsgroups, making this a **multi-class classification** problem. Each article is represented by a *tf-idf* term for each word in the training split word vocabulary. Newsgroups dataset has sparsity 99.9%, *i.e.*, only around 150 words from a vocabulary of 150k words appears per a given article, on average.

MIMIC-II [41, 51]. The task is to make a **binary prediction** on mortality of Intensive Care Unit (ICU) patients. Contains physiologic signals and vital signs time series captured from patient monitors for tens of thousands of ICU patients.

Credit [17, 19]. This dataset contains anonymized credit card transactions labeled as fraudulent or genuine, and the task is to make a **binary prediction** between them.

Click [15]. Subset of data from the KDD Cup 2012. Namely, 500,000 objects of a positive class and 500,000 objects of a negative class were randomly sampled to create a **binary prediction** task. We would like to gratefully acknowledge the organizers of KDD Cup 2012 as well as Tencent Inc. for making the dataset available.

Epsilon [21]. Dataset for **binary prediction** from the PASCAL Large Scale Learning Challenge.

Higgs [3, 26]. The problem is to **binary predict** whether the given event produces Higgs bosons.

Microsoft [40, 49]. Ranking dataset, where features are extracted from query-url pairs. Each pair has relevance judgment labels to **regress**, which take values from 0 (irrelevant) to 4 (very relevant).

Yahoo [60]. Another ranking dataset with query-url pairs that have labels to **regress** from 0 to 4.

Year [66]. Subset of Million Song Dataset. The task is to **regress** the release year of the song by using the audio features. It contains tracks from 1922 to 2011.

CUB [18, 62]. This **image classification** dataset consists of images of 200 bird classes. All images are annotated with keypoint locations of 15 bird parts (*e.g.*, beak, wing, crown) and each location is associated with one or more part-attribute labels (*e.g.*, orange leg, striped wing). Some of the keypoint annotations distinguish between the left-right instances of parts, *e.g.*, 'left wing' / 'right wing', 'left eye' / 'right eye'. We treat these as the same part, *i.e.*, 'left wing' and 'right wing' as 'wing'.

iNaturalist Birds [27, 59]. Another **image classification** dataset that contains 1,486 bird classes. The full iNaturalist 2021 dataset consists of various super-categories (*e.g.*, plants, insects, birds), covering 10K species in total. The Birds super-category contains 1,486 bird classes and more challenging scenes compared to CUB. Therefore, the iNaturalist dataset is a challenging testbed for any image classification method. However, note that this dataset lacks keypoint annotations.

Common Objects. Proprietary **object detection** dataset created by collecting public images from Instagram[1]. Dataset contains 114 common household objects, (*e.g.*, stove, bed, table), plus a background class, with bounding box locations, 200 parts and 54 attributes. Each bounding box for each image is pre-processed using compositions of parts (*e.g.*, leg, handle, top) and attributes (*e.g.*, colors, textures, shapes) to extract 2,618 interpretable features and 100k pairwise feature interactions. Common Objects dataset has sparsity 97%, *i.e.*, only around 76 part-attribute compositions from a vocabulary of 2618 compositions are active for a given object, on average.

[1] www.instagram.com

Table A.1: Optimal hyper-parameters for NBMs and NB^{2}Ms on all datasets.

NBM: $[256, 256, 128]$ hidden units, 100 basis functions

Dataset	Number of epochs	Batch size	Learning rate	Weight decay	Dropout	Basis dropout	Output penalty
CA Housing	1,000	1,024	0.00197	1.568e-5	0.0	0.05	1.439e-4
FICO	1,000	1,024	0.02176	1.684e-5	0.3	0.7	2.462e-4
CoverType	500	1,024	0.01990	5.931e-7	0.0	0.0	0.05533
Newsgroups	500	512	3.133e-4	1.593e-8	0.1	0.3	4.578
MIMIC-II	1,000	1,024	0.01460	3.177e-6	0.5	0.1	2.318
Credit	500	1,024	0.00391	1.574e-6	0.0	0.9	0.03737
Click	500	1,024	2.745e-4	7.21e-10	0.0	0.5	20.085
Epsilon	500	1,024	3.776e-5	1.507e-7	0.3	0.4	0.00273
Higgs	50	1,024	1.792e-4	1.087e-9	0.0	0.0	3.906e-5
Microsoft	500	1,024	1.677e-4	1.969e-7	0.1	0.3	1.986e-4
Yahoo	500	1,024	0.00446	1.399e-8	0.1	0.3	0.01688
Year	500	1,024	8.780e-5	1.580e-7	0.1	0.1	2.592e-5
CUB	500	128	0.01173	0.12910	0.7	0.3	4.739
iNaturalist Birds	100	1,024	0.00140	3.548e-5	0.0	0.2	1.423e-5
Common Objects	100	1,024	0.12480	1.001e-5	0.1	0.0	0.0

NB^{2}M: $[256, 256, 128]$ hidden units, 200 basis functions

Dataset	Number of epochs	Batch size	Learning rate	Weight decay	Dropout	Basis dropout	Output penalty
CA Housing	1,000	1,024	0.00190	7.483e-9	0.0	0.05	1.778e-6
FICO	1,000	1,024	2.287e-4	3.546e-7	0.1	0.7	0.19330
CoverType	500	512	0.00268	1.660e-7	0.0	0.0	0.00155
MIMIC-II	1,000	1,024	1.796e-4	3.494e-4	0.1	0.5	0.05964
Credit	500	1,024	3.745e-4	4.610e-5	0.5	0.0	0.25280
Click	500	1,024	9.614e-4	0.00159	0.0	0.5	0.05773
Higgs	50	1,024	0.00201	2.202e-4	0.0	0.1	1.969e-7
Microsoft	100	128	1.640e-4	1.552e-8	0.0	0.9	2.928e-6
Year	100	256	3.180e-4	1.696e-8	0.0	0.9	4.454e-4
CUB	500	32	2.629e-4	0.03209	0.0	0.0	96.894
iNaturalist Birds	100	32	6.735e-5	9.870e-5	0.05	0.0	3.785
Common Objects	100	64	0.03127	1.013e-4	0.0	0.2	8.126

A.2 Hyper-parameters

Linear, MLP, NAM, and NBM are trained using the Adam with decoupled weight decay (AdamW) optimizer [35], on 8×V100 GPU machines with 32 GB memory, and a batch size of at most 1024 per GPU. We train for 1,000, 500, 100, or, 50 epochs, depending on the size and feature dimensionality of the dataset. The learning rate is decayed with cosine annealing [34] from the starting value until zero. We find optimal hyper-parameters for all models using validation set and random search, following the detailed guidelines.

Linear. We tune the starting learning rate in the continuous interval $[1\text{e}{-5}, 100)$, weight decay in the interval $[1\text{e}{-10}, 1.0)$.

MLP. We tune the starting learning rate in the continuous interval $[1\text{e}{-5}, 1.0)$, weight decay in the interval $[1\text{e}{-10}, 1.0)$, dropout coefficients in the discrete set $\{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

NAM. We tune the starting learning rate in the continuous interval $[1\text{e}{-5}, 1.0)$, weight decay in the interval $[1\text{e}{-10}, 1.0)$, output penalty coefficient in the interval $[1\text{e}{-7}, 100)$, dropout and feature dropout coefficients in the discrete set $\{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

NBM. We tune the starting learning rate in the continuous interval $[1\text{e}{-5}, 1.0)$, weight decay in the interval $[1\text{e}{-10}, 1.0)$, output penalty coefficient in the interval $[1\text{e}{-7}, 100)$, dropout and basis dropout coefficients in the discrete set $\{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. Optimal hyper-parameters for NBMs and NB^{2}Ms on all datasets are given in Table A.1.

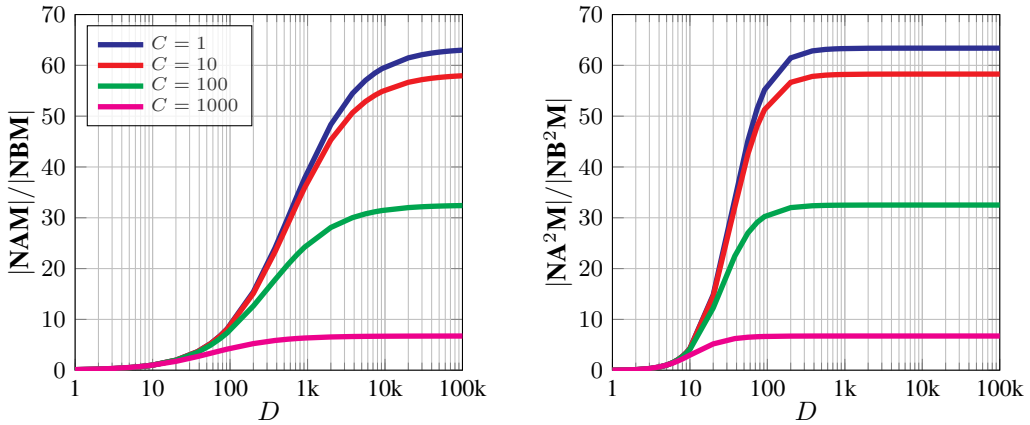

Figure A.1: NAM *vs.* NBM: #parameters (left); NA^{2}M *vs.* NB^{2}M: #parameters (right).

Finally, for EBMs and XGBoost, CPU machines are used, with hyper-parameter search as follows.

EBM. We tune the maximum bins from the set $\{8, 16, 32, 64, 128, 256, 512\}$, number of interactions from $\{0, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$ (they are set to 0 for EBMs and ≥ 0 for EB^{2}Ms), learning rate in the continuous range from $[1\text{e}{-6}, 100)$, the maximum rounds from the set $\{1000, 2000, 4000, 8000, 16000\}$, the minimum samples in a leaf node from the set $\{1, 2, 4, 8, 10, 15, 20, 25, 50\}$, and the same range is used for the maximum leaves parameter. For binning, we search within the set {"quantile", "uniform", "quantile_humanized"}. The inner bags and outer bags are selected from the range $\{1, 2, 4, 8, 16, 32, 64, 128\}$.

XGBoost. We tune the number of estimators from $\{1, 2, 4, 8, 10, 20, 50, 100, 200, 250, 500, 1000\}$, the max-depth from the set $\{\infty, 2, 5, 10, 20, 25, 50, 100, 2000\}$, η over a continuous range $[0.0, 1.0)$, and use the same for the `subsample` and `colsample_bytree` parameters.

A.3 Additional discussion *w.r.t.* NAM

Number of parameters for multi-class and pairwise feature interactions. We compare number of weight parameters needed to learn NAM *vs.* NBM for the multi-class task. This discussion is an extension of the discussion in Section 3.3. Let us denote with M the number of parameters in MLP for each feature in NAM, and with N the number of parameters in MLP for bases in NBM. In most experiments the optimal NAM has 3 hidden layers with 64, 64 and 32 units ($M = 6401$), and, NBM has 3 hidden layers with 256, 128, 128 units ($N = 62820$) and $B = 100$ basis functions. Finally, let us denote with D the input feature dimensionality, and with C the number of classes in the multi-class task. Then the ratio of number of parameters in NAM *vs.* NBM is given by,

$$\frac{|\textbf{NAM}|}{|\textbf{NBM}|} = \frac{D \cdot M + D \cdot C}{N + D \cdot B + D \cdot C} = \frac{6401 + C}{\frac{62820}{D} + 100 + C}. \tag{A.1}$$

Similarly, in the case of pairwise feature interactions in NA^{2}M *vs.* NB^{2}M, this ratio is given by,

$$\frac{|\textbf{NA}^2\textbf{M}|}{|\textbf{NB}^2\textbf{M}|} = \frac{\frac{D(D-1)}{2} \cdot M + \frac{D(D-1)}{2} \cdot C}{N + \frac{D(D-1)}{2} \cdot B + \frac{D(D-1)}{2} \cdot C} = \frac{6401 + C}{\frac{125640}{D(D-1)} + 100 + C}. \tag{A.2}$$

Figure A.1 shows both ratios for different values of feature dimensionality D and number of classes C.

For unary features (NAM *vs.* NBM, Figure A.1 left), the conclusion is the same as in the case of binary classification, *i.e.*, for $D = 10$, NBMs and NAMs have roughly equal number of parameters, for any given value of C. For higher number of classes, NBMs still provide significant gain over NAMs, however, that gain starts decreasing, due to the fact that the final linear classifier starts becoming the most memory hungry part of the model. Nevertheless, even with $C = 1{,}486$ and $D = 278$ in iNaturalist Birds, which has the most classes from the datasets we used, NBMs have around $5\times$ less parameters than NAMs, see Table 2.

For pairwise feature interactions (NA^{2}M *vs.* NB^{2}M, Figure A.1 right), the ratio is much more pronounced, *i.e.*, already at $D = 5$, NB^{2}Ms and NA^{2}Ms have equal number of parameters, and the growth of the ratio *w.r.t.* D is much more significant. Already after few hundred feature dimensions the ratio is at peak.

Throughput optimization of NAMs. Neural Additive Models (NAMs) [2] learn an MLP network for each input feature, followed by a linear combination to make a prediction. Official implementation [42] runs a `for loop` over all networks, which results in a poor GPU utilization. More precisely, this implementation requires extremely large batches (>>1,024) per GPU to make the training efficient, which is impractical. We do recognize that efficiency was not of highest priority to the authors [2], but in our case we are scaling GAMs to multi-class datasets with order of million data points. Thus, to facilitate a fair comparison against our NBMs, we reimplement NAMs using grouped convolutions [30, 65], which are readily available in standard deep learning libraries. Namely, we stack corresponding hidden layers of all MLPs (*e.g.*, first hidden layer of all MLPs) into a grouped 1-D convolution, where number of groups equals the number of features. The computation performed is identical to original NAMs, *i.e.* there is no feature interaction, while achieving around $\times 2$–$\times 10$ speedup, depending on the dataset. We perform the same implementation trick to NA^{2}Ms, as well.

A.4 Additional visualization

The interpretability of GAMs comes from the fact that the learned shape functions can be easily visualized. In the same manner as the other GAM approaches, each feature's importance in an NBM can be represented by a unique shape function that *exactly* describes how the NBM computes a prediction. For an example, see Figure 3 visualization on the CA Housing dataset. We additionally demonstrate this on the CUB image classification dataset that consists of 200 bird classes, where each image is represented by interpretable features, *e.g.*, a "bird" image can be represented with "striped wings", "needle-shaped beak", "long legs", *etc.*, that are predicted from the image using a convolutional-neural-network (CNN) model. We present visualizations of shape functions with highest positive or negative contribution to 6 randomly selected bird classes, see Figures A.2 and A.3.

Towards this purpose, we train an ensemble of 20 models by running different random seeds with optimal hyper-parameters, in order to analyze when the models learn the same shape and when they diverge. Following Agarwal et al. [2], we set the average score for each shape function to be zero by subtracting the respective mean feature score. Next, we plot each shape function as $f_i(x_i)$ *vs.* x_i for each model in the ensemble using a semi-transparent line, and an average ensemble shape function using a thick line. Finally, the x-axis is divided by bars depicting the normalized data density, *i.e.*, darker areas contain more data points. Figures A.2 and A.3 depict few image examples for the respective class (upper row), and an ensemble of NBMs (bottom row). We visually observe that NBMs provide a strong interpretable overview of the respective bird class, and that the shape functions do not diverge significantly even in the cases where there are only few data points (light / white areas in the graphs).

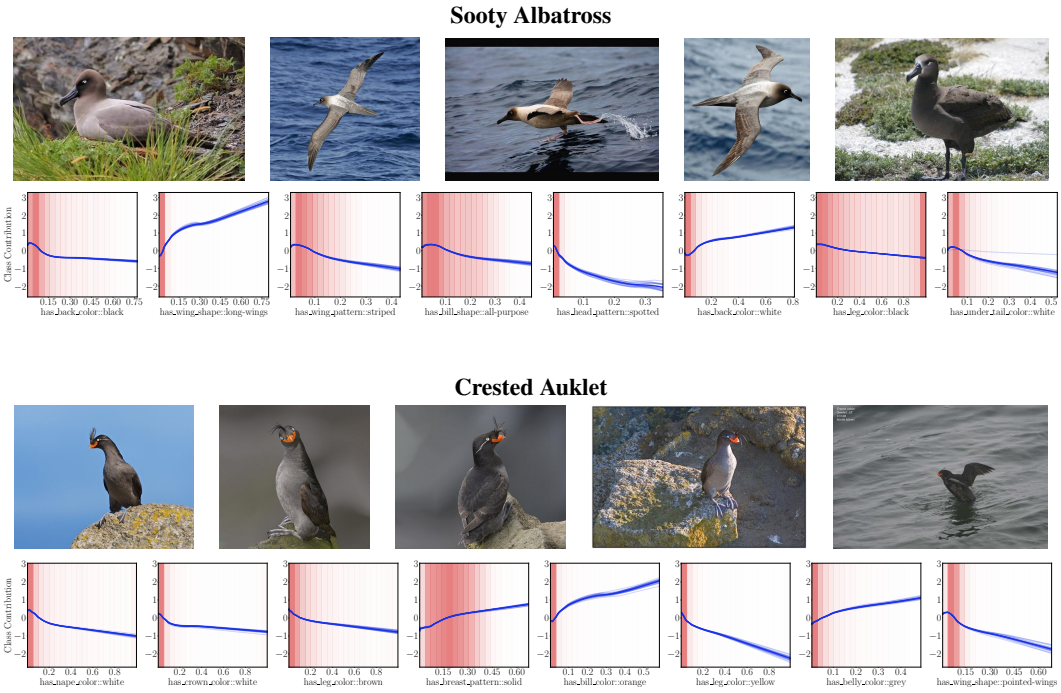

Figure A.2: CUB bird class image examples (upper row) and NBM shape functions f_i (bottom row) with highest positive or negative contribution to the respective bird class prediction.

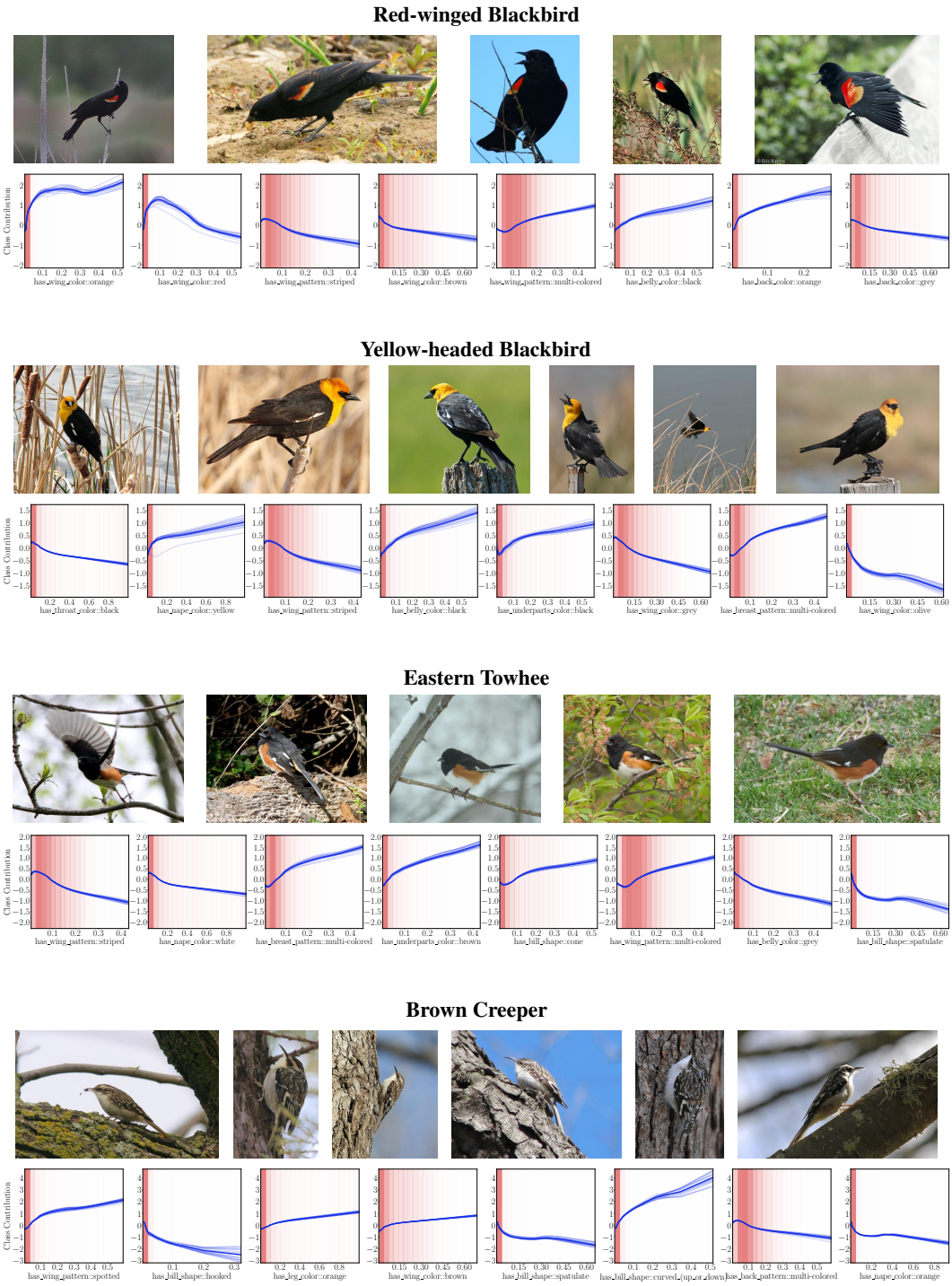

Figure A.3: CUB bird class image examples (upper row) and NBM shape functions f_i (bottom row) with highest positive or negative contribution to the respective bird class prediction.

A.5 Learning-Theoretic Guarantees for Basis Models in a RKHS

As discussed briefly in the main paper, it is possible to develop a more rigorous argument for the use of a small set of basis functions instead of a complete generalized additive model. To elucidate we first require establishing some notation: We represent matrices by uppercase boldface, *e.g.*, $\mathbf{X}$ and vectors by lowercase boldface, *i.e.*, $\boldsymbol{x}$. We assume that the covariates lie within the set $\mathcal{X} \subseteq \mathbb{R}^D$, and the labels lie within the finite set $\mathcal{Y}$. Data $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ are drawn following some unknown (but fixed) distribution $\mathfrak{P}$. We assume we are provided with n i.i.d. samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ as the train set.

Consider a generalized additive model (GAM) $g : \mathcal{X} \to \mathcal{Y}$:

$$g(\boldsymbol{x}) = \sum_{i=1}^{D} w_i \cdot f_i(x_i).$$

Assume that the shape functions $f_1, \ldots, f_D; f_i : \mathbb{R} \to \mathcal{Y}$ have a maximum norm $B_{\mathcal{H}} > 0$ in some Reproducing Kernel Hilbert Space (RKHS, [7]) $\mathcal{H}$ endowed with a PSD kernel $k(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and feature $\boldsymbol{\phi} : \mathbb{R} \to \mathbb{R}^{d_{\mathcal{H}}}$, *i.e.*, $\|f_i\|_{\mathcal{H}} \leq B_{\mathcal{H}}$, and $\boldsymbol{w} = \{w_i\}_{i=1}^D \in \mathbb{R}^D$, $k(x, y) = \boldsymbol{\phi}(x)^\top \boldsymbol{\phi}(y)$ such that $\|\boldsymbol{w}\|_2 \leq B_{\boldsymbol{w}}$ for $B_{\boldsymbol{w}} > 0$. This characterization corresponds to a family of functions $\mathcal{H}_A$, *i.e.*,

$$\mathcal{H}_A = \{g \mid g(\boldsymbol{x}) = \sum_{i=1}^{D} w_i f_i(x_i), \|f_i\|_{\mathcal{H}} \leq B_{\mathcal{H}}, \|\boldsymbol{w}\|_2 \leq B_{\boldsymbol{w}}\} \tag{A.3}$$

The idea behind the basis decomposition approach highlighted in this paper is to only use a fixed number of bases, B, to model each f_i. Observe that one can obtain rigorous guarantees for f_i that lie within an RKHS using Mercer's Theorem [39]. We have that if the kernel k associated with the RKHS $\mathcal{H}$ is continuous, positive-definite and symmetric, there exist a set of eigenvalues $\{\lambda_i\}_{i=1}^\infty$ and eigenfunctions (basis functions) $\{\boldsymbol{\omega}_i\}_{i=1}^\infty$ that form an orthonormal basis for k, *i.e.*, for any $x, y \in \mathbb{R}$,

$$k(x, y) = \sum_{i=1}^{\infty} \lambda_i \boldsymbol{\omega}_i(x) \boldsymbol{\omega}_i(y). \tag{A.4}$$

Where the bases are orthonormal, *i.e.*, $\int_{x \in \mathbb{R}} \boldsymbol{\omega}_i(x) \boldsymbol{\omega}_j(x) dx = 0$ for $i \neq j$ and 1 otherwise. This representation naturally gives a form for $\boldsymbol{\phi}(\cdot) = [\sqrt{\lambda_i} \boldsymbol{\omega}_i(\cdot)]_{i=1}^\infty$. Furthermore, we have that for each $f \in \mathcal{H}$ there exists $\boldsymbol{f} \in L^2$ such that $f(x) = \langle \boldsymbol{f}, \boldsymbol{\phi}(x) \rangle_{\mathcal{H}} \forall x \in \mathbb{R}$. Note once again that the reproducing kernel Hilbert space $\mathcal{H}$ corresponds to the feature-wise functions f, whereas the space $\mathcal{H}_A$ corresponds to the overall function g. Now, we can define, for the family $\mathcal{H}_A$ a **Generalized Basis Model** of order B (denoted as $\mathcal{H}_B$) as the following.

Definition 1. A Generalized Basis Model of order B for any function class $\mathcal{H}_A$ that satisfies the characterization in Equation A.3 for some $(\mathcal{H}, B_{\mathcal{H}}, B_{\boldsymbol{w}})$ is given by the family $\mathcal{H}_B$:

$$\mathcal{H}_B = \left\{ g \;\middle|\; \begin{array}{c} g(\boldsymbol{x}) = \sum_{i=1}^{D} w_i f_i(x_i), \\ f_i(\cdot) = \sum_{j=1}^{B} \beta_{ij} h_j(\cdot), \|f_i\|_{\mathcal{H}} \leq B_{\mathcal{H}}, \|\boldsymbol{w}\|_2 \leq B_{\boldsymbol{w}}, \\ h_i \in \mathcal{H}, h_i \perp h_j \forall\, i \neq j. \end{array} \right\}$$

Where orthogonality ($\perp$) is defined as $h_i \perp h_j \implies \int_{x \in \mathbb{R}} h_i(x) \cdot h_j(x) dx = 0$.

Next, note that by Mercer's Theorem, for each function $f \in \mathcal{H}$, there exists $\boldsymbol{f} = \{f_i\}_{i=1}^\infty, \boldsymbol{f} \in L^2$ such that $f(x) = \langle \boldsymbol{f}, \boldsymbol{\phi}(x) \rangle_{\mathcal{H}}$. Combining this statement with the basis representation for $\boldsymbol{\phi}$ gives us an alternate representation of any $f \in \mathcal{H}$, as

$$f(\cdot) = \sum_{i=1}^{\infty} f_i \sqrt{\lambda_i} \boldsymbol{\omega}_i(\cdot).$$

Under this representation, we can relate the two spaces $\mathcal{H}_A$ and $\mathcal{H}_B$ as follows.

Proposition 1. *For any $\mathcal{H}$, dimensionality D, and number of basis functions $B > 0$, $\mathcal{H}_B \subseteq \mathcal{H}_A$.*

Proof. Follows from Mercer's Theorem [39]. Any $g \in \mathcal{H}_B$ can be written as a linear combination of functions in $\mathcal{H}$ (and consequently $\mathcal{H}_A$), each of which admit a basis representation via Mercer's Theorem, where all but B components have coefficient 0. In the limit $B \to \infty$, $\mathcal{H}_B = \mathcal{H}_A$. □

Since the basis functions in $\mathcal{H}_B$ lie on a finite-dimensional subspace within $\mathcal{H}$ spanned by B basis vectors, we can without loss of generality, assume that these B basis vectors correspond to $\{\boldsymbol{\omega}_i\}_{i=1}^B$ obtained from Equation A.4. Now, to prove generalization bounds on the best function learnable in $\mathcal{H}_B$ and contrast that with $\mathcal{H}_A$, we require a "smoothing" assumption in $\{\boldsymbol{\omega}_i\}_{i=1}^\infty$ (and correspondingly on $\mathcal{H}$). The essence of this assumption is to ensure that the kernel $\mathcal{H}$ can be spanned without introducing much error by only with a few basis components, and is similar to smoothing kernel assumptions made in other areas as well, *e.g.*, in reinforcement learning.

Assumption 1 (γ-Exponential Spectral Decay of $\mathcal{H}$)**.** *For the decomposition of $\mathcal{H}$ as outlined in Equation A.4, we assume that there exist absolute constants $C_1 < 1$ and $C_2 = \mathcal{O}(1)$ and parameter γ such that $\lambda_i \leq C_1 \exp(-C_2 \cdot i^\gamma)$ for each $i \geq 1$.*

At a high level, our approach is to bound the *test error* of the *empirical risk minimizer* in $\mathcal{H}_A$, with the optimal risk minimizer in $\mathcal{H}_B$ to demonstrate that learning a generalized basis model does not incur significantly larger error compared to learning the full model. We first make these terms precise. Recall that the empirical risk for any function g is given by $\widehat{\mathcal{L}}_n(g) = \frac{1}{n}\sum_{i=1}^n \ell(g(\boldsymbol{x}_i), y_i)$. We denote $\hat{g}$ as the empirical risk minimizer within $\mathcal{H}_B$, *i.e.*,

$$\widehat{g} = \arg\min_{g \in \mathcal{H}_B} \widehat{\mathcal{L}}_n(g). \tag{A.5}$$

Similarly, the *expected risk* can be given, for any function g as $\mathcal{L} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{P}}\left[\ell(g(\boldsymbol{x}), y)\right]$. Then we can define the *optimal expected risk minimizer* $g_\star$ in $\mathcal{H}_A$ as,

$$g_\star = \arg\min_{g \in \mathcal{H}_A} \mathcal{L}(g). \tag{A.6}$$

We are now equipped to discuss our generalization bound.

Theorem 1. *Let ℓ be a 1-Lipschitz loss, $\delta \in (0, 1]$ and Assumption 1 hold with constants C_1, C_2, γ. Then we have that with probability at least $1 - \delta$ there exist absolute constants C_1, C_2 such that,*

$$\mathcal{L}(\hat{g}) - \mathcal{L}(g_\star) \leq 2B_{\boldsymbol{w}}\sqrt{\frac{B}{n}} + \frac{DC_2}{C_1}\exp(-B^\gamma) + 5\sqrt{\frac{\log(4/\delta)}{n}}.$$

Proof. We will denote the weights and singular values for $f_\star$ as $\boldsymbol{w}^\star$ and $\lambda_{ij}^\star$, *i.e.*, $g_\star(\boldsymbol{x}) = \sum_{i=1}^D w_i^\star h_i^\star(x_i)$ where $h_i^\star(x) = \sum_{j=1}^\infty \lambda_{ij}^\star \boldsymbol{\omega}_j(x)$. Note that this represnetation exists for some $\lambda_{ij}^\star$ by Mercer's Theorem, as discussed earlier. For any $\mathcal{H}_B, \mathcal{H}_A$, consider the function $\tilde{g} \in \mathcal{H}_B$ that is a truncated version of $g_\star$ up to b bases, *i.e.*, $\tilde{g}(\boldsymbol{x}) = \sum_{i=1}^D w_i^\star \widetilde{h}_i(x_i)$ where $\widetilde{h}_i(x) = \sum_{j=1}^b \lambda_{ij}^\star \boldsymbol{\omega}_j(x)$. Clearly, $\tilde{g} \in \mathcal{H}_B$. We can then rewrite the L.H.S. in the Theorem as,

$$\mathcal{L}(\hat{g}) - \mathcal{L}(g_\star) = \underbrace{\mathcal{L}(\hat{g}) - \widehat{\mathcal{L}}_n(\hat{g})}_{\text{Ⓐ}} + \underbrace{\widehat{\mathcal{L}}_n(\hat{g}) - \widehat{\mathcal{L}}_n(\tilde{g})}_{\leq 0} + \underbrace{\widehat{\mathcal{L}}_n(\tilde{g}) - \mathcal{L}(g)}_{\text{Ⓑ}}.$$

Note that the middle term $\widehat{\mathcal{L}}_n(\hat{g}) - \widehat{\mathcal{L}}_n(\tilde{g}) \leq 0$ since $\hat{g}$ is the empirical risk minimizer in $\mathcal{H}_B$. Hence, by bounding terms Ⓐ and Ⓑ, the proof will be complete. We can bound Ⓑ via Lemma 1. We have that with probability at least $1 - \delta/2$ for any $\delta \in (0, 1]$,

$$\left|\widehat{\mathcal{L}}_n(\tilde{g}) - \mathcal{L}(g_\star)\right| \leq \frac{L \cdot C_1}{C_2}\exp(-B^\gamma) + 2\sqrt{\frac{\log(2/\delta)}{n}}.$$

We bound Ⓐ via bounding the Rademacher complexity [63]. Since the loss function is Lipschitz and bounded, with probability at least $1 - \delta/2, \delta \in (0, 1]$, we have that by Theorem 12 and Theorem 8 of Bartlett and Mendelson [4],

$$\mathcal{L}(\hat{g}) - \widehat{\mathcal{L}}_n(\hat{g}) \leq \mathfrak{R}_n(\ell \odot \mathcal{H}_B) + \sqrt{\frac{8\log(4/\delta)}{n}}. \tag{A.7}$$

Where $\mathfrak{R}_n$ denotes the empirical Rademacher complexity at n samples [4]. Observe that each element of $\mathcal{H}_B$ is a linear combination of d elements that are represented by b basis vectors in $\mathcal{H}$. Hence, there exist weights $\{\{\alpha_{ij}\}_{i=1}^D\}_{j=1}^B$ such that any $f \in \mathcal{H}_B$ can be written as $\sum_{i,j}\alpha_{ij}\boldsymbol{\omega}_j(x_i)$, $\|\boldsymbol{\alpha}\|_2 \leq B_{\mathcal{H}}B_{\boldsymbol{w}}$ where $\boldsymbol{\alpha} = \{\{\alpha_{ij}\}_{i=1}^D\}_{j=1}^B$. Furthermore, we have that for any x ,$\boldsymbol{\phi}(x)^\top\boldsymbol{\phi}(x) = \sum_{j=1}^B \boldsymbol{\omega}_j(x_i)^2 \leq B$. We therefore have, by Theorem 12 of [4] that with probability at least $1 - \delta/2, \delta \in (0, 1]$,

$$\begin{aligned}
\mathcal{L}(\hat{g}) - \widehat{\mathcal{L}}_n(\hat{g}) &\leq \mathfrak{R}_n(\ell \odot \mathcal{H}_B) + \sqrt{\frac{8\log(4/\delta)}{n}} \\
&\leq 2L\mathfrak{R}_n(\mathcal{H}_B) + \sqrt{\frac{8\log(4/\delta)}{n}} \\
&\leq 2LB_{\boldsymbol{w}}\mathfrak{R}_n(\mathcal{H}) + \sqrt{\frac{8\log(4/\delta)}{n}} \\
&\leq 2LB_{\boldsymbol{w}}\sqrt{\frac{B}{n}} + \sqrt{\frac{8\log(4/\delta)}{n}}
\end{aligned}$$

The last inequality follows from Lemma 22 of [4]. Replacing the above result for k, we have that with probability at least $1 - \delta/2$, Using the bound for term Ⓑ and applying a union bound provides us the final result.

Lemma 1. *The following holds with probability at least* $1-\delta, \delta \in (0,1]$*, for some absolute constant* $C \ll 1$*,*

$$\left|\widehat{\mathcal{L}}_n(\tilde{g}) - \mathcal{L}(g_\star)\right| \leq \frac{LD \cdot C_1}{C_2} \exp(-B^\gamma) + 2\sqrt{\frac{\log(2/\delta)}{n}}.$$

Proof.

$$\begin{aligned}\widehat{\mathcal{L}}_n(\tilde{g}) - \mathcal{L}(g_\star) &= \widehat{\mathcal{L}}_n(\tilde{g}) - \mathcal{L}(\tilde{g}) + \mathcal{L}(\tilde{g}) - \mathcal{L}(g_\star) \\ &\leq \underbrace{\left|\widehat{\mathcal{L}}_n(\tilde{g}) - \mathcal{L}(\tilde{g})\right|}_{①} + \underbrace{\left|\mathcal{L}(\tilde{g}) - \mathcal{L}(g_\star)\right|}_{②}.\end{aligned}$$

To bound ①, we have that for any $(\boldsymbol{x}, y)$ within the training set, $\mathbb{E}[\ell(\tilde{g}(\boldsymbol{x}), y)] = \mathcal{L}(\tilde{g})$ and $0 \leq \ell(\cdot,\cdot) \leq 1$. By Azuma-Hoeffding, we obtain with probability at least $1-\delta, \delta \in (0,1]$,

$$\left|\widehat{\mathcal{L}}_n(\tilde{g}) - \mathcal{L}(\tilde{g})\right| \leq 2\sqrt{\frac{\log(2/\delta)}{n}}.$$

For ②, since ℓ is $L-$Lipschitz, we have for some $x_1, x_2, y \in \mathcal{Y}$,

$$\begin{aligned}|\ell(x_1, y) - \ell(x_2, y)| &\leq |L \cdot |x_1 - y| - L \cdot |x_2 - y|| \\ &\leq L \cdot |x_1 - x_2|.\end{aligned}$$

Therefore:

$$\begin{aligned}|\mathcal{L}(\tilde{g}) - \mathcal{L}(g_\star)| &\leq \left|\mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{P}}\left[\ell(\tilde{g}(\boldsymbol{x}), y) - \ell(g_\star(\boldsymbol{x}), y)\right]\right| \\ &\leq \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{P}}\left[|\ell(\tilde{g}(\boldsymbol{x}), y) - \ell(g_\star(\boldsymbol{x}), y)|\right] \\ &\leq L \cdot \mathbb{E}_{(\boldsymbol{x},y)\sim\mathfrak{P}}\left[|\tilde{g}(\boldsymbol{x}) - g_\star(\boldsymbol{x})|\right] \\ &\leq L \cdot \sup_{\boldsymbol{x}\in\mathcal{X}} |\tilde{g}(\boldsymbol{x}) - g_\star(\boldsymbol{x})|.\end{aligned}$$

Observe now that for any $\boldsymbol{x} \in \mathcal{X}$,

$$|\tilde{g}(\boldsymbol{x}) - g_\star(\boldsymbol{x})| = \left|\sum_{i=1}^{D}\sum_{j=1}^{\infty}(\lambda_{ij}^\star - \widetilde{\lambda}_{ij})\boldsymbol{\omega}_j(x_i)\right| \leq \sum_{i=1}^{D}\sum_{j=1}^{B}\left|(\lambda_{ij}^\star - \widetilde{\lambda}_{ij})\right| \leq \sum_{i=1}^{D}\sum_{j=B+1}^{\infty}|\lambda_{ij}^\star|.$$

Invoking Assumption 1, we have that

$$\sum_{i=1}^{D}\sum_{j=B+1}^{\infty}|\lambda_{ij}^\star| \leq D\sum_{j=B+1}^{\infty} C_1 \exp(-C_2 j^\gamma) \leq D\int_{j=B}^{\infty} C_1 \exp(-C_2 j^\gamma).$$

Since $\gamma \geq 1$, we have,

$$\int_{j=r_i}^{\infty} C_1 \exp(-C_2 j^\gamma) \leq \frac{C_1}{C_2}\exp\left(-B^\gamma\right).$$

A union bound for both parts finishes the proof. □

Discussion. The result holds when the target function class is a member of a Reproducing Kernel Hilbert Space (RKHS). While RKHSes include a variety of expressive machine learning function classes, *e.g.*, radial basis functions, polynomials, linear classifiers, it is not known whether arbitrarily initialized neural networks have a small norm in any RKHS with desirable properties. Most notably, however, it was shown recently that certain neural networks can be represented via the Neural Tangent Kernel (NTK), an example of where the theory can be applied as-is. More generally, however, this result demonstrates for arbitrary infinite-dimensional RKHS, we have an exponential dependence on the number of basis B required in the approximation error (second term). Observe that if we set $B = \mathcal{O}(\log D)$, the second term is $o(1)$ and goes to 0 as $n \to \infty$, which suggests that in practice, we only require a number of bases, B that grows logarithmically with the dimensionality D.