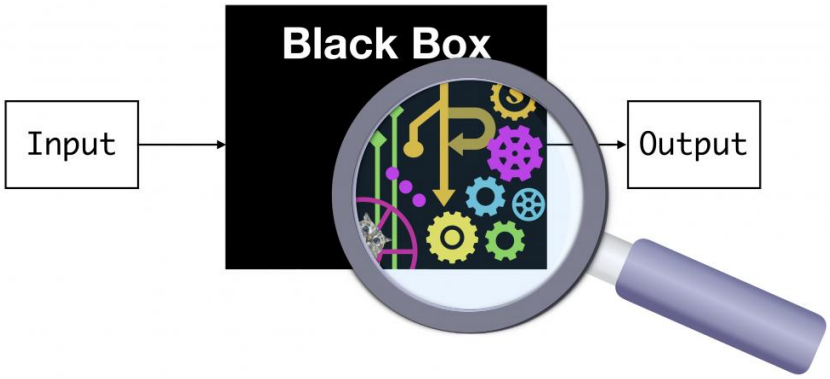
Neural Additive Experts: Context-Gated Experts for Controllable Model Additivity

Guangzhi Xiong, Sanchit Sinha, Aidong Zhang

Interpretability in Machine Learning (ML)

- Interpretability is the degree to which a human can understand the cause of a decision[1].
- When a model is fully interpretable, it will facilitate
  - Fairness: Ensuring that predictions are unbiased.
  - Privacy: Ensuring that sensitive information in the data is protected.
  - Robustness: Ensuring that small changes in the input don't lead to large changes in the prediction.
  - ……



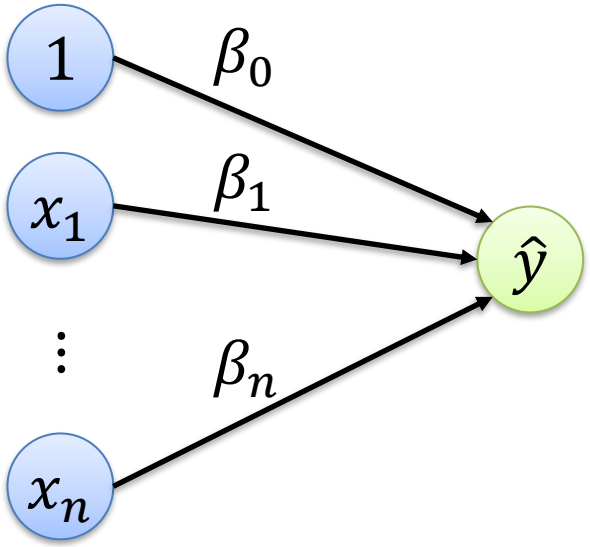
[1]Explanation and Justification in Machine Learning: A Survey. IJCAI Workshop. 2017.

Linear Model: A Fully Interpretable Model

- Given features $x_1, \ldots, x_n$, the prediction $\hat{y}$ can be written as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

where $\beta_1, \ldots, \beta_n$ are the learned feature weights and β_0 is the intercept.



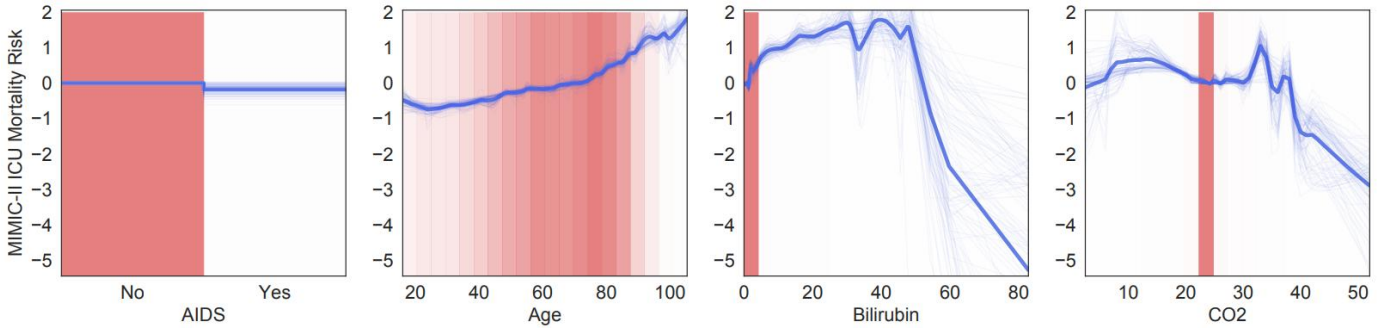
Generalized Additive Models (GAMs)

- In a GAM, the predicted output for a sample with n features $x_1, \cdots, x_n$ is modeled as an additive combination:

$$\hat{y} = \omega_0 + \sum_{i=1}^{n} f_i(x_i),$$

where each function f_i captures the influence of feature x_i and ω_0 is a learnable intercept.

- A key advantage of GAMs is their interpretability[1]:

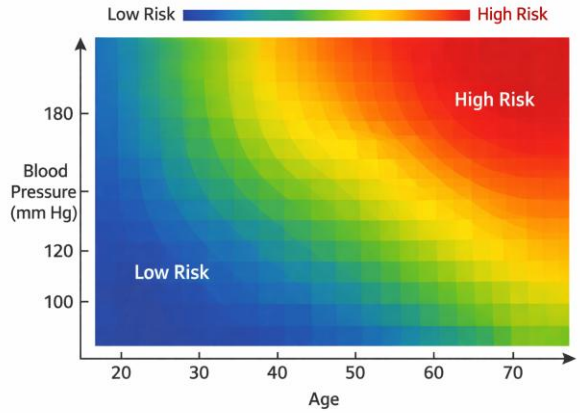


[1]Neural Additive Models: Interpretable Machine Learning with Neural Nets. NeurIPS. 2021.

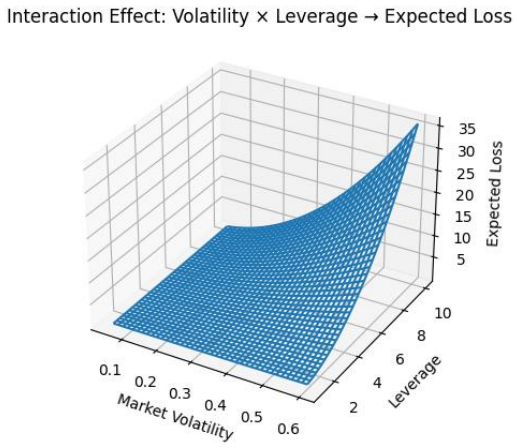
Core Limitation of GAMs

- Strict additive structure constrains the learning of feature interactions that are useful in decision making.

Longitude & Latitude



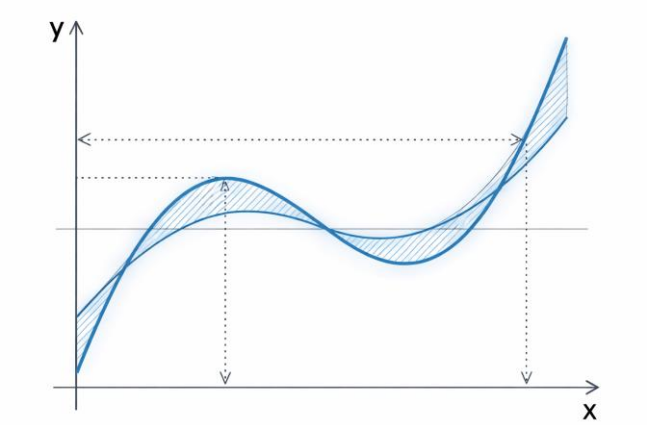
Age & Blood Pressure



Volatility & Leverage

The Interpretability-Accuracy Trade-off

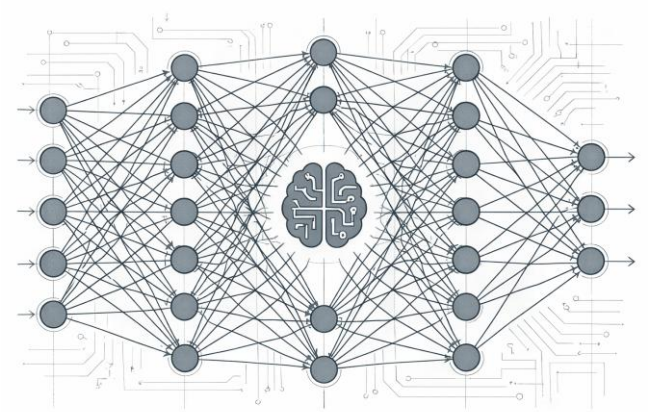
The Glass Box (GAMs)



Pros
- Clear feature attribution
- Transparent decision-making

Cons
- Fails on complex interactions
- Lower predictive performance

The Black Box (DNNs)



Pros
- High accuracy
- Captures intricate dependencies

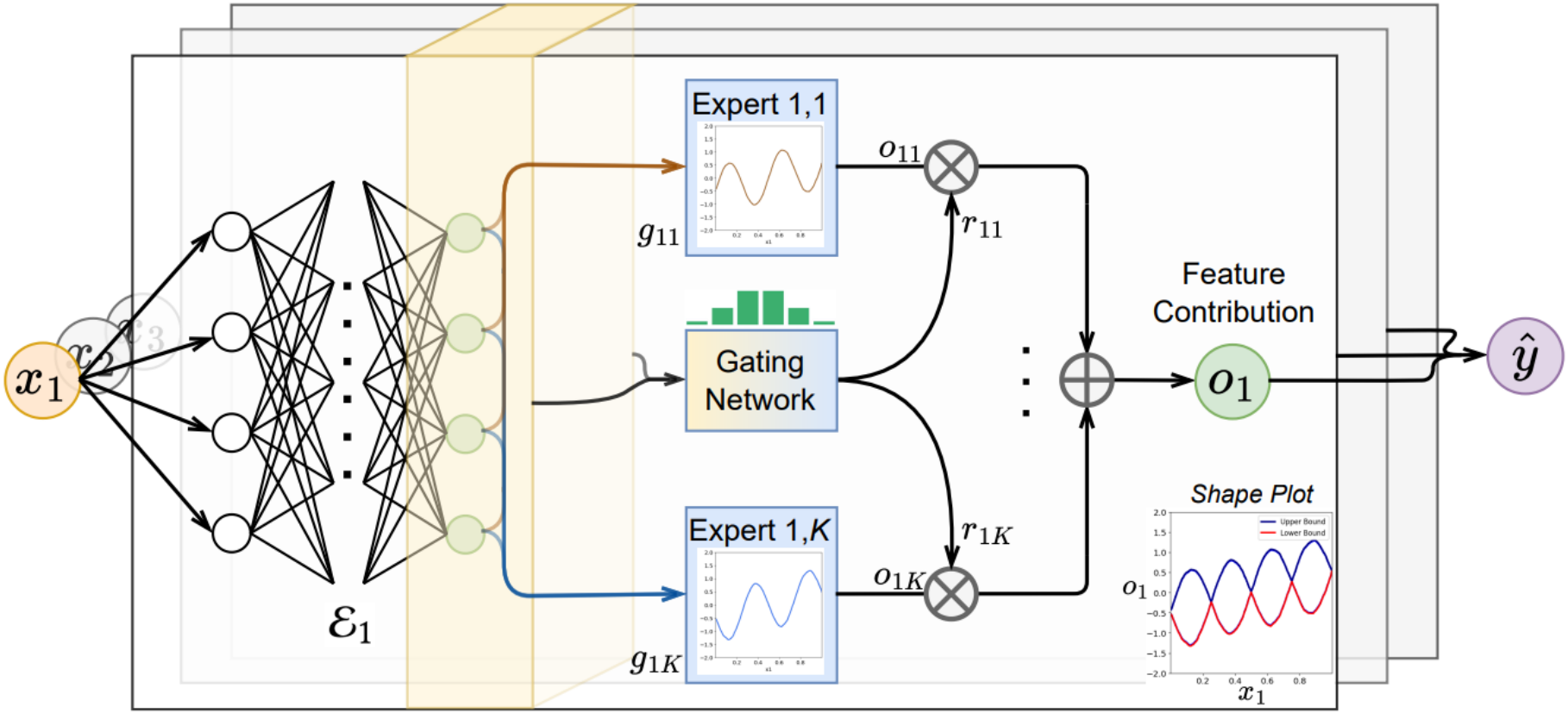
Cons
- Opaque decision process
- Trust issues in high-stakes fields

Motivation: Mitigating the Additive Constraint

- Wanted: Enabling feature interaction modeling in GAMs is necessary to improve their performance.
- Unwanted: However, explicitly adding interaction terms to GAMs (e.g., GA^{2}Ms) breaks the interpretability for single-feature attributions.

➢ A new model architecture is desired to enable interaction modeling while keeping the overall additive structure of GAMs for interpretability.

Neural Additive Experts



Training Objective

$$\text{minimize} \quad \frac{1}{N}\sum_{t=1}^{N}\mathcal{L}(y^t,\hat{y}^t)+\frac{\lambda}{nNK}\sum_{t=1}^{N}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(o_{ik}^t-\frac{1}{K}\sum_{l=1}^{K}o_{il}^t\right)^2$$

Theoretical Analysis of NAEs

- Theorem 1 (GAM containment). For any f ∈ GAM there exists K = 1 and parameters of an NAE such that $\hat{y}(x) = f(x)$ for all x.
- Theorem 2 (GA²M containment up to arbitrary precision). For any f ∈ GA²M and any $\epsilon > 0$ there exists K and an NAE such that $\sup_x |\hat{y}(x) - f(x)| < \epsilon$.

➢ NAEs are more expressive than vanilla GAM/GA²M but remain feature-decomposed at prediction time.

Theoretical Analysis of NAEs

- A key feature of NAEs is to interpolate between strictly additive and highly flexible models. We quantify additivity with

$$\text{Additivity} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{Var}\big(\mathbb{E}(o_i|x_i)\big)+\delta}{\text{Var}(o_i)+\delta}$$

- Theorem 3 (Monotone additivity and additive limit). Let A(λ) be the additivity metric of an NAE trained with penalty λ ≥ 0. Then A(λ) is nondecreasing in λ. Moreover, $\lim_{\lambda\to\infty} A(\lambda) = 1$, and any limit point of minimizers as $\lambda \to \infty$ is a GAM.

Experiments on Simulated Data

- Key Research Questions
  - Can NAEs accurately recover underlying shape functions in both unimodal (additive) and multimodal (non-additive) scenarios?
  - How does the expert variation penalty parameter λ influence the trade-off between model flexibility and additivity?
- Simulation Setup
  - We generate two synthetic datasets, each with 10,000 samples:

Unimodal Distribution (Additive). With $x_1 \sim U(0,1)$ and $\sigma = 0.1$, the response is sampled as:

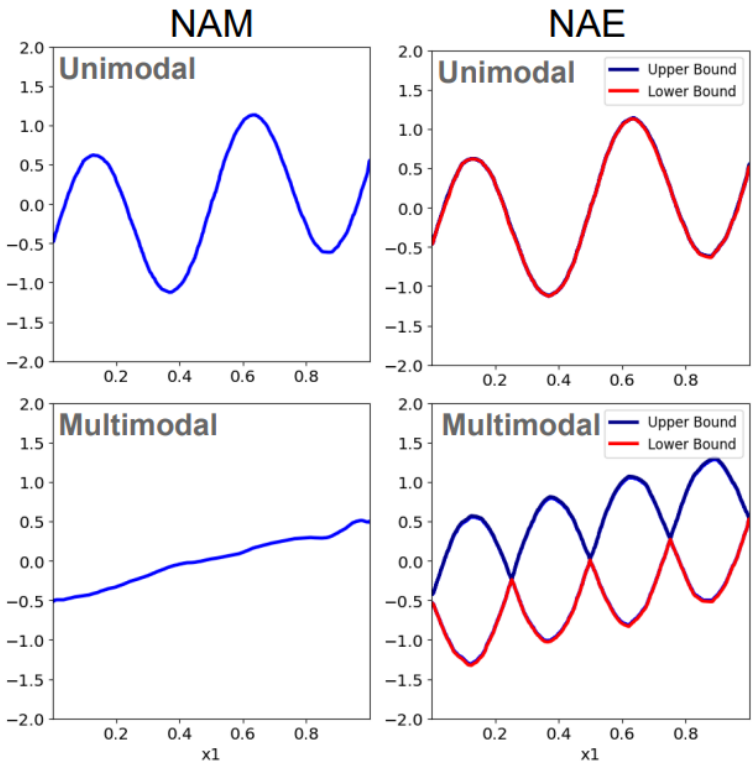
$$y \sim \mathcal{N}\left(x_1 - \frac{1}{2} + \sin(4\pi x_1), \sigma^2\right), \tag{16}$$

Multimodal Distribution (Non-Additive). With x_2 uniformly sampled from $\{-1, 1\}$, the response is generated following:

$$y \sim \begin{cases} \mathcal{N}\left(x_1 - \frac{1}{2} - \sin(4\pi x_1), \sigma^2\right), & \text{if } x_2 = -1, \\ \mathcal{N}\left(x_1 - \frac{1}{2} + \sin(4\pi x_1), \sigma^2\right), & \text{if } x_2 = 1, \end{cases} \tag{17}$$

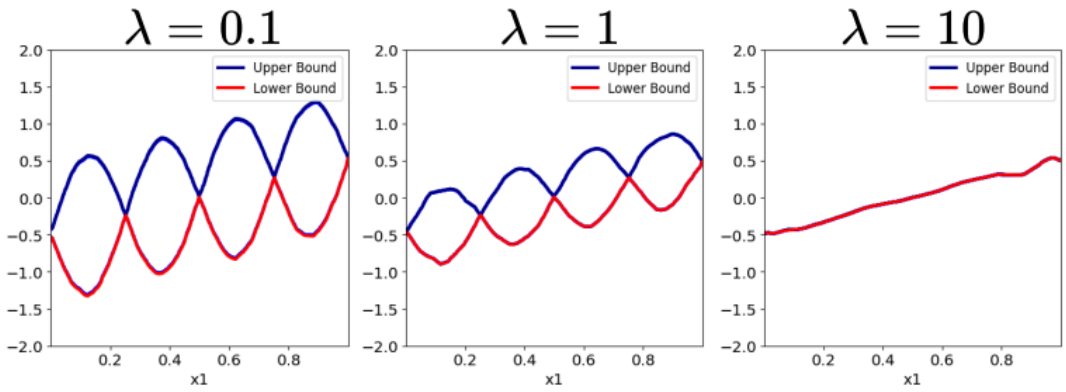
Experiments on Simulated Data

- Can NAEs accurately recover underlying shape functions?
  - In the unimodal setting, both NAM and NAE accurately recover the underlying pattern.
  - In the multimodal setting, NAM fails to capture the interaction, while NAE recovers the multimodal structure.



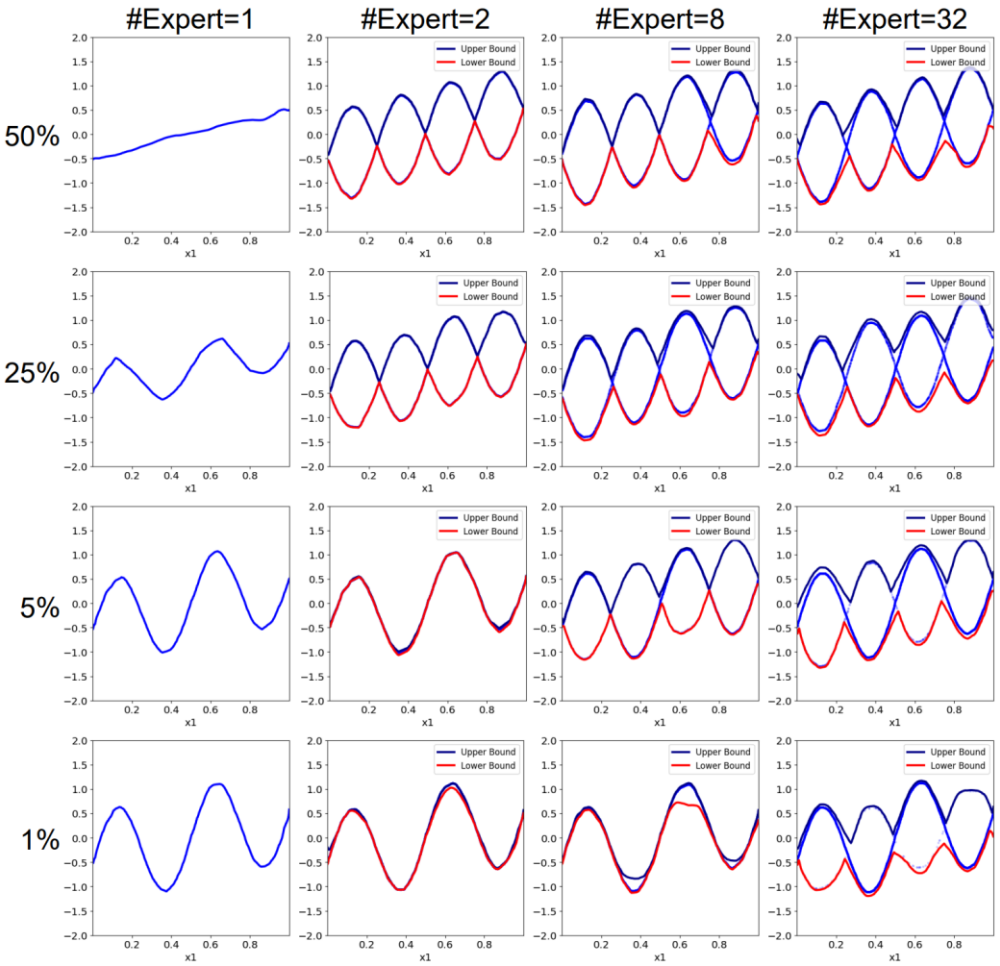
Experiments on Simulated Data

- Effect of Variation Penalty
  - We train NAEs on the multimodal dataset with $\lambda = 0.1, 1, 10$ and visualize the learned shape functions for x_1
  - As λ increases, the bounds converge, enforcing additivity.



Experiments on Simulated Data

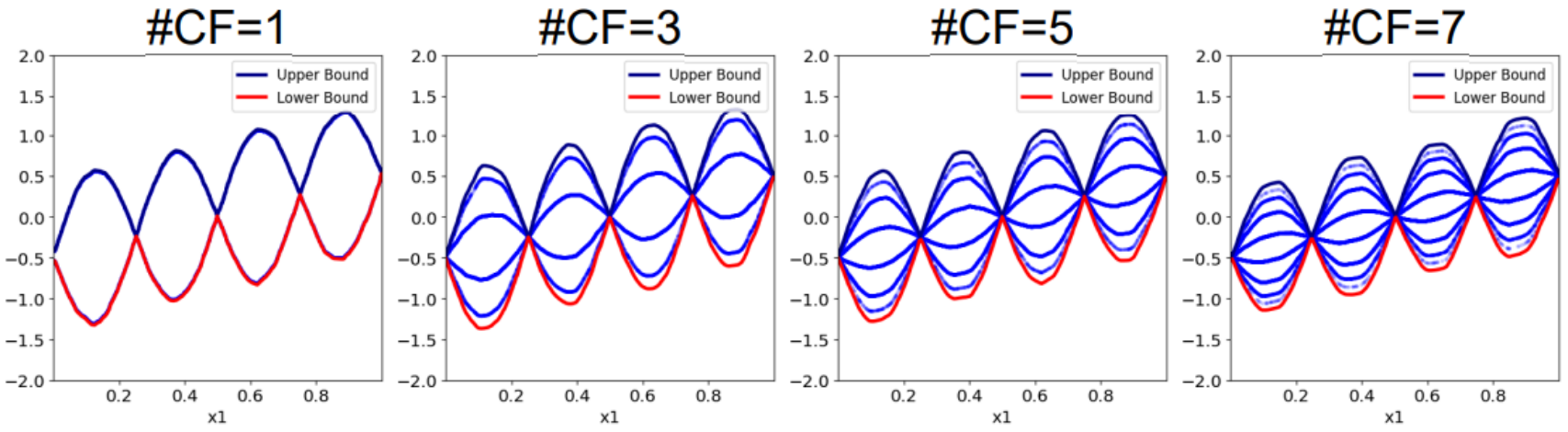
- Impact of Feature Sparsity
  - To assess robustness to imbalanced feature distributions, we simulate datasets where the proportion of instances with $x_2 = 1$ is set to 50%, 25%, 5%, and 1%



Experiments on Simulated Data

- Impact of Distribution Modality
  - We further test NAEs on data with increasing numbers of modes by introducing multiple categorical features:

$$y = \varepsilon + x_1 - \frac{1}{2} + \frac{1}{CF} \sum_{i=2}^{CF+1} x_i \sin(4\pi x_1),$$



Experiments on Real-world Data

- Datasets
  - Classification: MIMIC-II, MIMIC-III, Income, Credit
  - Regression: Housing, Year

Dataset	Task	#Instances	#Features	#Categorical Features	#Classes
Housing	regression	20,640	8	0	–
MIMIC-II	classification	24,508	17	7	2
MIMIC-III	classification	27,348	57	0	2
Income	classification	32,561	14	8	2
Credit	classification	284,807	30	0	2
Year	regression	515,345	90	0	–

- Baselines
  - GAM: Linear, Spline, NAM, NBM, EBM, NODE-GAM
  - GA^{2}M: NA^{2}M, NB^{2}M, EB^{2}M, NODE-GA^{2}M
  - Black box: MLP, NODE, XGBoost

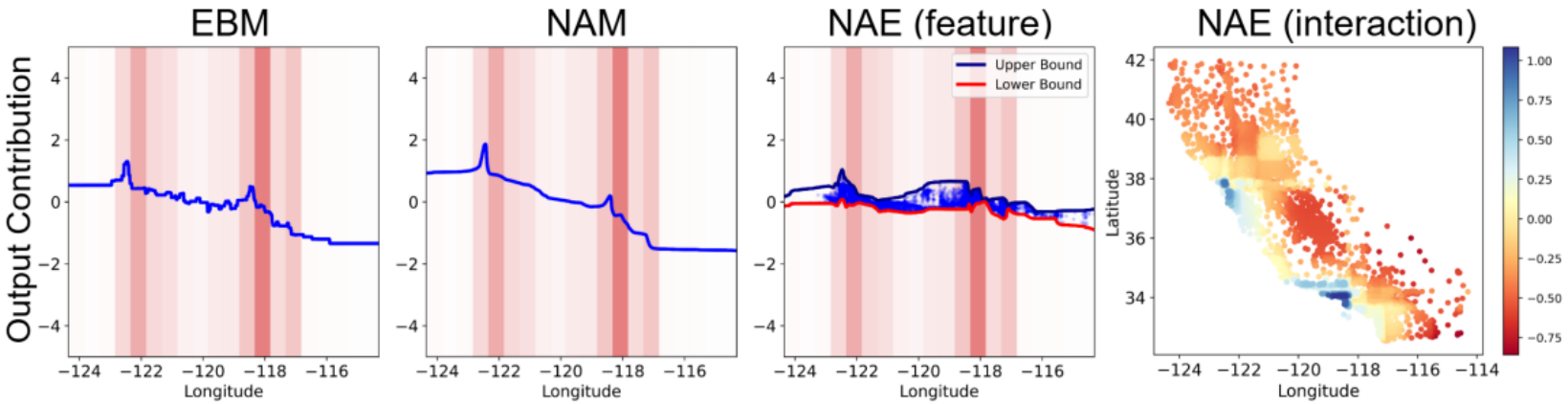
Experiments on Real-world Data

- NAEs outperforms all additive models, matching or exceeding the performance of more complex models while maintaining feature-level interpretability.

Model	Interpret-ability	Complex.	Housing	MIMIC-II	MIMIC-III	Income	Credit	Year
			RMSE ↓	AUC ↑	AUC ↑	AUC ↑	AUC ↑	MSE ↓
Black-box Models								
MLP	No	N/A	0.501 ±0.006	0.835 ±0.014	0.815 ±0.009	0.914 ±0.003	**0.981 ±0.007**	78.48 ±0.56
NODE	No	N/A	0.523 ±0.000	0.843 ±0.011	**0.828 ±0.007**	0.919 ±0.003	0.981 ±0.009	**76.21 ±0.12**
XGBoost	No	N/A	**0.443 ±0.000**	**0.844 ±0.012**	0.819 ±0.004	**0.928 ±0.003**	0.978 ±0.009	78.53 ±0.09
Interaction-explaining Models								
EB^{2}M	Yes	$O(n^2)$	0.492 ±0.000	**0.848 ±0.012**	0.821 ±0.004	**0.928 ±0.003**	0.982 ±0.006	83.16 ±0.01
NA^{2}M	Yes	$O(n^2)$	0.492 ±0.008	0.843 ±0.012	**0.825 ±0.006**	0.912 ±0.003	0.985 ±0.007	79.80 ±0.05
NB^{2}M	Yes	$O(n^2)$	0.478±0.002	**0.848 ±0.012**	0.819 ±0.010	0.917 ±0.003	0.978 ±0.007	79.01 ±0.03
NODE-GA^{2}M	Yes	$O(n^2)$	**0.476 ±0.007**	0.846 ±0.011	0.822 ±0.007	0.923 ±0.003	**0.986 ±0.010**	**79.57 ±0.12**
Feature-explaining Models								
Linear	Yes	$O(n)$	0.735 ±0.000	0.796 ±0.012	0.772 ±0.009	0.900 ±0.002	0.976 ±0.012	88.89 ±0.40
Spline	Yes	$O(n)$	0.568 ±0.000	0.825 ±0.011	0.812 ±0.004	0.918 ±0.003	0.982 ±0.011	85.96 ±0.07
EBM	Yes	$O(n)$	0.559 ±0.000	0.835 ±0.011	0.809 ±0.004	**0.927 ±0.003**	0.974 ±0.009	85.81 ±0.11
NAM	Yes	$O(n)$	0.572 ±0.005	0.834 ±0.013	0.813 ±0.003	0.910 ±0.003	0.977 ±0.015	85.25 ±0.01
NBM	Yes	$O(n)$	0.564 ±0.001	0.833 ±0.013	0.806 ±0.003	0.918 ±0.003	0.981 ±0.007	85.10 ±0.01
NODE-GAM	Yes	$O(n)$	0.558 ±0.003	0.832 ±0.011	0.814 ±0.005	**0.927 ±0.003**	0.981 ±0.011	85.09 ±0.01
NAE	Yes	$O(n)$	**0.451 ±0.002**	**0.847 ±0.014**	**0.825 ±0.006**	**0.927 ±0.003**	**0.982 ±0.009**	**78.66 ±0.21**

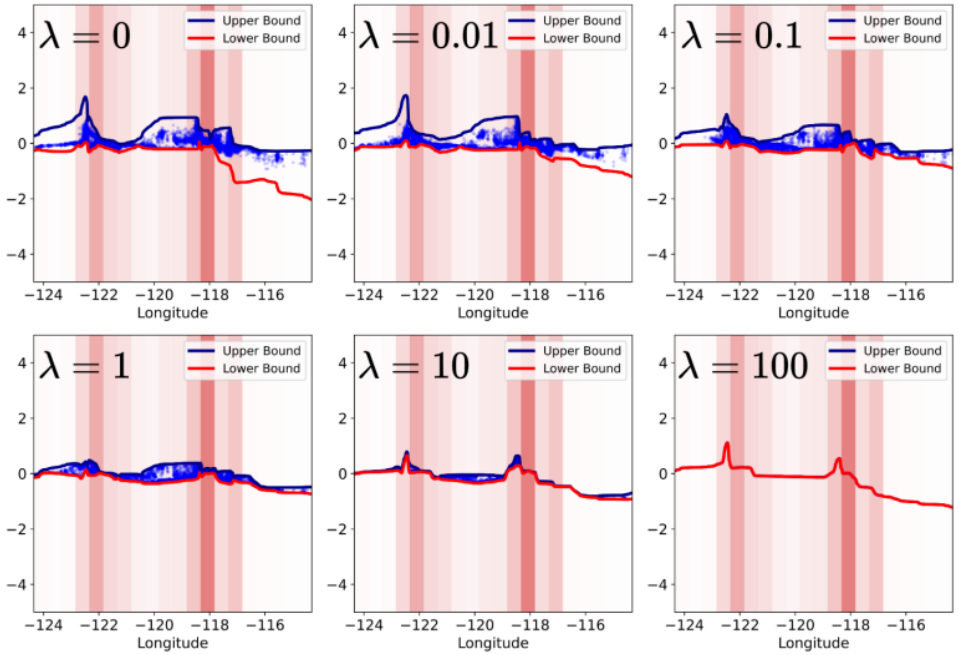
Experiments on Real-world Data

- NAE has feature-level interpretability
  - It captures similar geographic trends as in NAM/EBM, where key longitudes (e.g., -122.5 for San Francisco) show higher property values.
  - NAE further reveals a wide range of possible outcomes between -120 and -119, with most data points near the lower bound.
    - *coastal regions positively influence house prices, while other areas in the same longitude range may have a negative effect*



Experiments on Real-world Data

- NAE can control the trade-off between accuracy and interpretability
  - As λ increases, the bounds on feature effects tighten, enforcing additivity but potentially reducing flexibility.



Experiments on Real-world Data

- NAE can control the trade-off between accuracy and interpretability
  - Quantitatively, we measure the tightness of bounds with

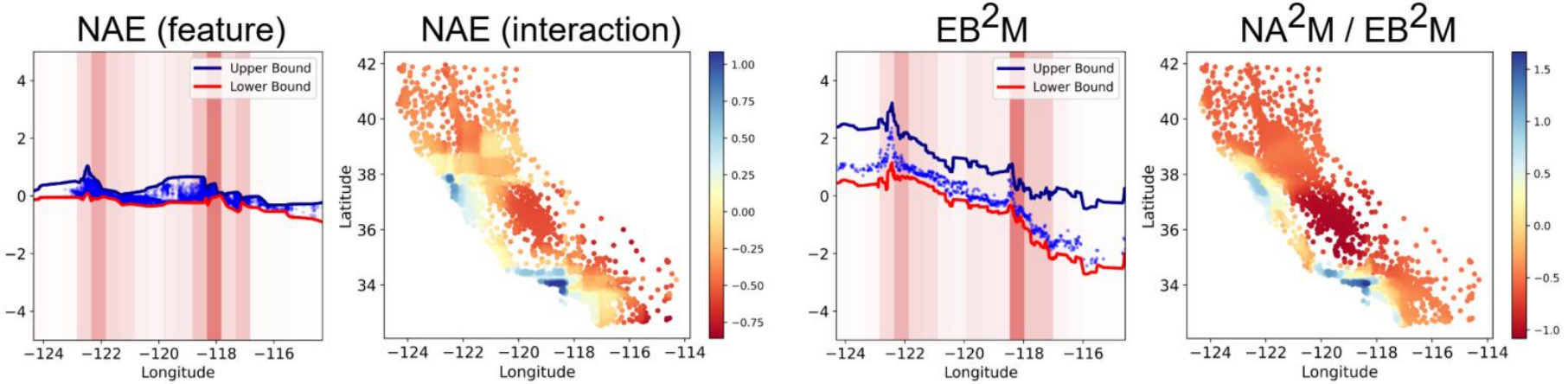
$$\text{Tightness} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\max(o_i|x_i) - \min(o_i|x_i) + \delta}{\text{upper}(o_i|x_i) - \text{lower}(o_i|x_i) + \delta}\right]$$

  - As λ increases, additivity and tightness improve, but predictive performance (RMSE) may degrade.

$\lambda = 0$	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 10$	$\lambda = 100$
Additivity					
0.522 ±0.000	0.537 ±0.000	0.562 ±0.012	0.639 ±0.000	0.897 ±0.000	1.000 ±0.000
Tightness					
0.860 ±0.000	0.895 ±0.018	0.953 ±0.011	0.988 ±0.000	0.999 ±0.000	1.000 ±0.000
RMSE					
0.451 ±0.003	0.450 ±0.002	0.451 ±0.003	0.458 ±0.003	0.515 ±0.008	0.582 ±0.008

Discussion: Comparing NAE and GA²M

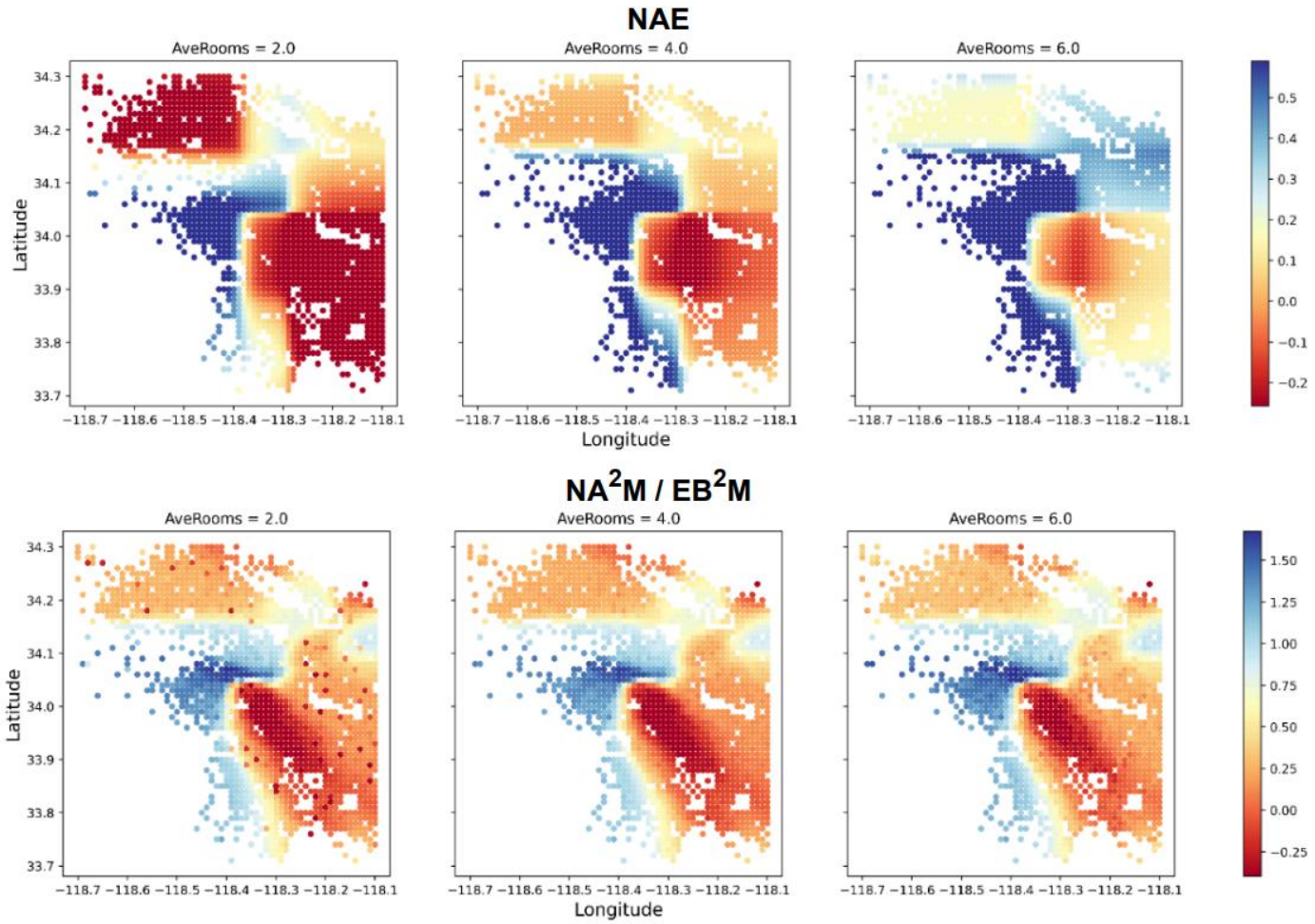
- A key advantage of NAE is its ability to control model additivity, which is not available in GA²M.
- Compared to GA²M, NAE provides more interpretable single-feature attributions, where the bounds tightly fit the actual predictions.



Discussion: Comparing NAE and GA²M

- Additionally, NAE is not limited to capturing pairwise interactions.

"AveRooms-Latitude-Longitude" interaction



Discussion: NAE with Diagonal Routing Matrix

- We explored a special version of NAE, termed NAE-D, where the scoring matrices compose a block diagonal matrix. In NAE-D, the matrix A_{ij} becomes 0 if $i \neq j$:

$$\varphi_j = \mu_j + \sum_{i=1}^{n} \mathcal{A}_{ij}^{\top} \mathcal{E}_i(x_i)$$

- To encourage multi-expert learning, we introduced randomness into with Gumbel-softmax:

$$\hat{r}_{ik} = \exp\left(\frac{\log(r_{ik}) + g_i}{\tau}\right) \Big/ \sum_{l=1}^{K} \exp\left(\frac{\log(r_{il}) + g_l}{\tau}\right) \qquad \text{where} \qquad g_i, g_l \sim Gumbel(0,1)$$

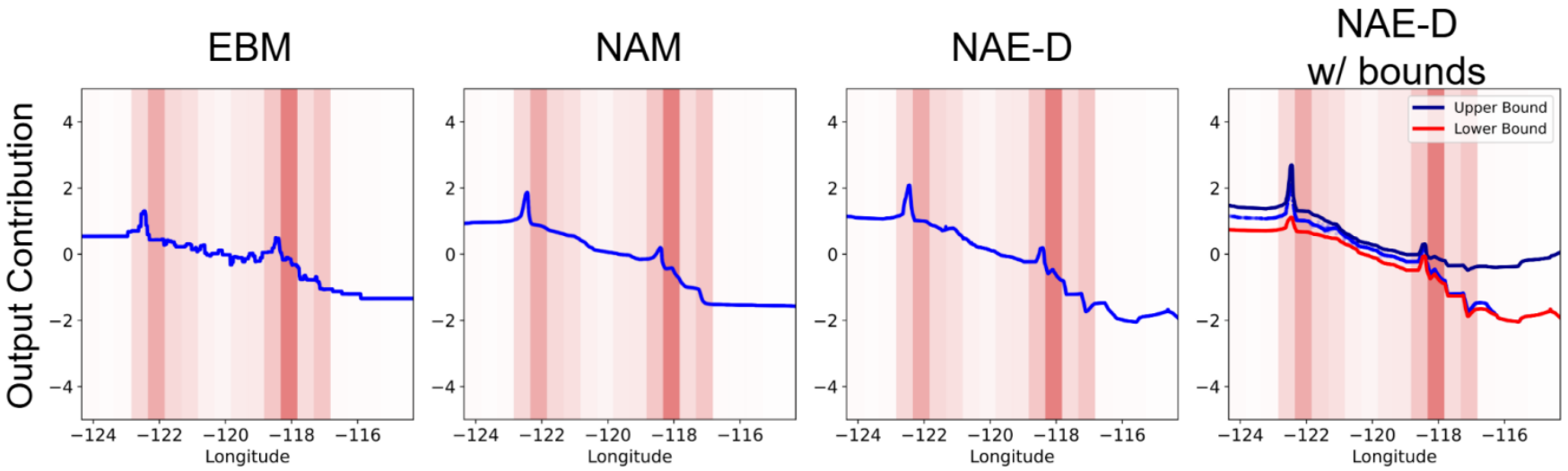
Discussion: NAE with Diagonal Routing Matrix

- NAE-D shows a performance close to traditional additive models, which have worse accuracies than complex models that capture feature interactions.

Model	Interpret-ability	Complex.	Housing	MIMIC-II	MIMIC-III	Income	Credit	Year
			RMSE ↓	AUC ↑	AUC ↑	AUC ↑	AUC ↑	MSE ↓
Black-box Models								
MLP	No	N/A	0.501 ± 0.006	0.835 ± 0.014	0.815 ± 0.009	0.914 ± 0.003	0.981 ± 0.007	78.48 ± 0.56
XGBoost	No	N/A	0.443 ± 0.000	0.844 ± 0.012	0.819 ± 0.004	0.928 ± 0.003	0.978 ± 0.009	78.53 ± 0.09
Interaction-explaining Models								
EB^2M	Yes	$O(n^2)$	0.492 ± 0.000	0.848 ± 0.012	0.821 ± 0.004	0.928 ± 0.003	0.982 ± 0.006	83.16 ± 0.01
NA^2M	Yes	$O(n^2)$	0.492 ± 0.008	0.843 ± 0.012	0.825 ± 0.006	0.912 ± 0.003	0.985 ± 0.007	79.80 ± 0.05
Feature-explaining Models								
EBM	Yes	$O(n)$	0.559 ± 0.000	0.835 ± 0.011	0.809 ± 0.004	0.927 ± 0.003	0.974 ± 0.009	85.81 ± 0.11
NAM	Yes	$O(n)$	0.572 ± 0.005	0.834 ± 0.013	0.813 ± 0.003	0.910 ± 0.003	0.977 ± 0.015	85.25 ± 0.01
NAE-D	Yes	$O(n)$	0.553 ± 0.001	0.830 ± 0.012	0.805 ± 0.006	0.927 ± 0.003	0.977 ± 0.010	85.60 ± 0.04

Discussion: NAE with Diagonal Routing Matrix

- The patterns captured by NAE-D are similar to the ones by EBM and NAM.
- NAE-D offers additional insights with its prediction bounds.



Discussion: NAE with Evenly Distributed Expert Activation

- While the default implementation of NAE uses continuous expert relevance estimation:

$$r_{jk} = \frac{\exp\left(\varphi_j[k] + M_j[k]\right)}{\sum_{l=1}^{K} \exp\left(\varphi_j[l] + M_j[l]\right)}, \quad k = 1, \cdots, K,$$

  we also explored an adaptation that shifts this estimation to a discrete framework:

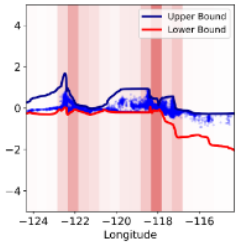
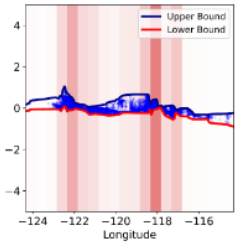
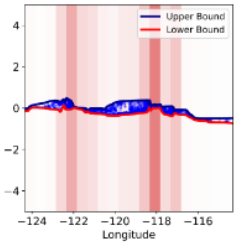
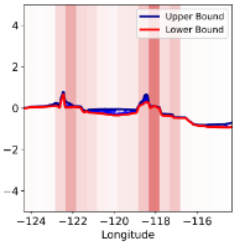
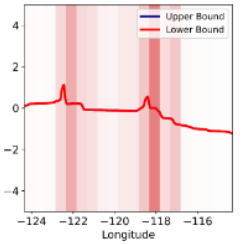
$$r_{ik} = \frac{\exp(\varphi_i[k] - \varphi_i^*[k] + M_i[k])}{\sum_{l=1}^{K} \exp(\varphi_i[l] - \varphi_i^*[l] + M_i[l])}$$

  where φ_i^* mirrors the values of φ_i but does not require gradients.

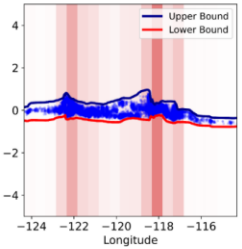
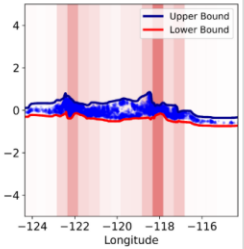
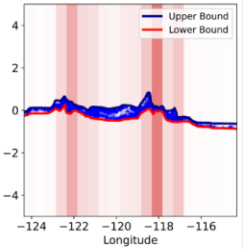
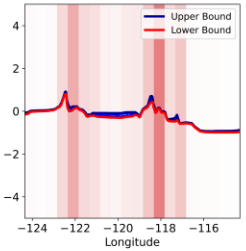
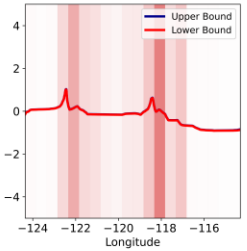
Discussion: NAE with Evenly Distributed Expert Activation

- Compared with NAE, the performance of NAE-E is slightly worse due to the limitations imposed by its discrete range.

NAE

$\lambda = 0$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 10$	$\lambda = 100$
RMSE (↓)				
0.451 ± 0.003	0.451 ± 0.003	0.458 ± 0.003	0.515 ± 0.008	0.582 ± 0.002
Shape Plot				
				

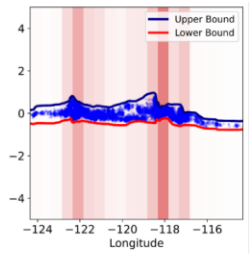
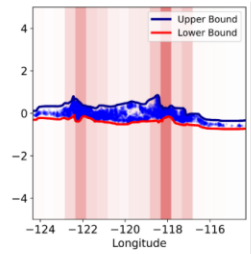
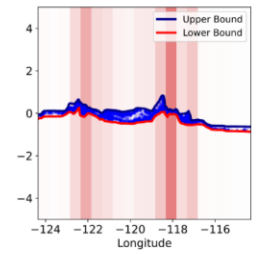
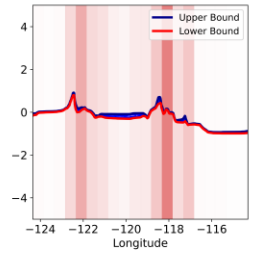
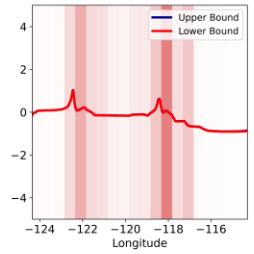
NAE-E

$\lambda = 0$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 10$	$\lambda = 100$
RMSE (↓)				
0.462 ± 0.002	0.460 ± 0.002	0.483 ± 0.001	0.551 ± 0.002	0.585 ± 0.003
Shape Plot				
				

Discussion: NAE with Evenly Distributed Expert Activation

- Compared with NAE, the performance of NAE-E is slightly worse due to the limitations imposed by its discrete range.

- However, even without any variation penalty, the estimated bounds by NAE-E still fit the actual score distribution tightly.

$\lambda = \mathbf{0}$	$\lambda = \mathbf{0.1}$	$\lambda = \mathbf{1}$	$\lambda = \mathbf{10}$	$\lambda = \mathbf{100}$
RMSE (↓)				
0.462 ± 0.002	0.460 ± 0.002	0.483 ± 0.001	0.551 ± 0.002	0.585 ± 0.003
Shape Plot				
				

Discussion: Scaling of Experts

- We explored how the number of total experts (K) and the number of activated experts (C) affect the overall model performance.

$C \setminus K$	NAE			
	2	**4**	**8**	**16**
2	0.479 ± 0.003	0.459 ± 0.003	0.455 ± 0.004	0.460 ± 0.004
4	–	$\underline{0.451} \pm 0.002$	$\underline{0.447} \pm 0.004$	0.455 ± 0.004
8	–	–	$\underline{0.449} \pm 0.004$	0.456 ± 0.003
16	–	–	–	0.456 ± 0.005
$C \setminus K$	NAE-E			
	16	**32**	**64**	**128**
2	0.486 ± 0.007	0.487 ± 0.005	0.486 ± 0.006	0.489 ± 0.005
4	0.469 ± 0.007	0.466 ± 0.005	0.468 ± 0.004	0.470 ± 0.004
8	0.470 ± 0.003	0.459 ± 0.004	$\underline{0.458} \pm 0.002$	0.458 ± 0.004
16	0.587 ± 0.003	0.462 ± 0.002	$\underline{0.451} \pm 0.002$	$\underline{0.453} \pm 0.002$

Discussion: Model Complexity

- Theoretical and empirical additional costs brought by NAE compared to NAM. n denotes the number of input features.
  - NAE is implemented with d = 128, K = 4.
  - NAE-D is implemented with d = 128, K = 64.

	n	NAE		NAE-D	
		Parameter Count	**Memory Usage**	**Parameter Count**	**Memory Usage**
Housing	8	37k	144k	132k	516k
MIMIC-II	17	157k	613k	281k	1.1M
MIMIC-III	57	1.7M	6.5M	941k	3.59M
Income	14	108k	420k	231k	903k
Credit	30	476k	1.8M	495k	1.89M
Year	90	4.2M	16.0M	1.5M	5.67M
Theoretical	–	$nK[(n+1)d+2]$		$nK(2d+2)$	