Massachusetts Institute of Technology

GRAND-SLAMIN’ Interpretable Additive Modeling with Structural Constraints

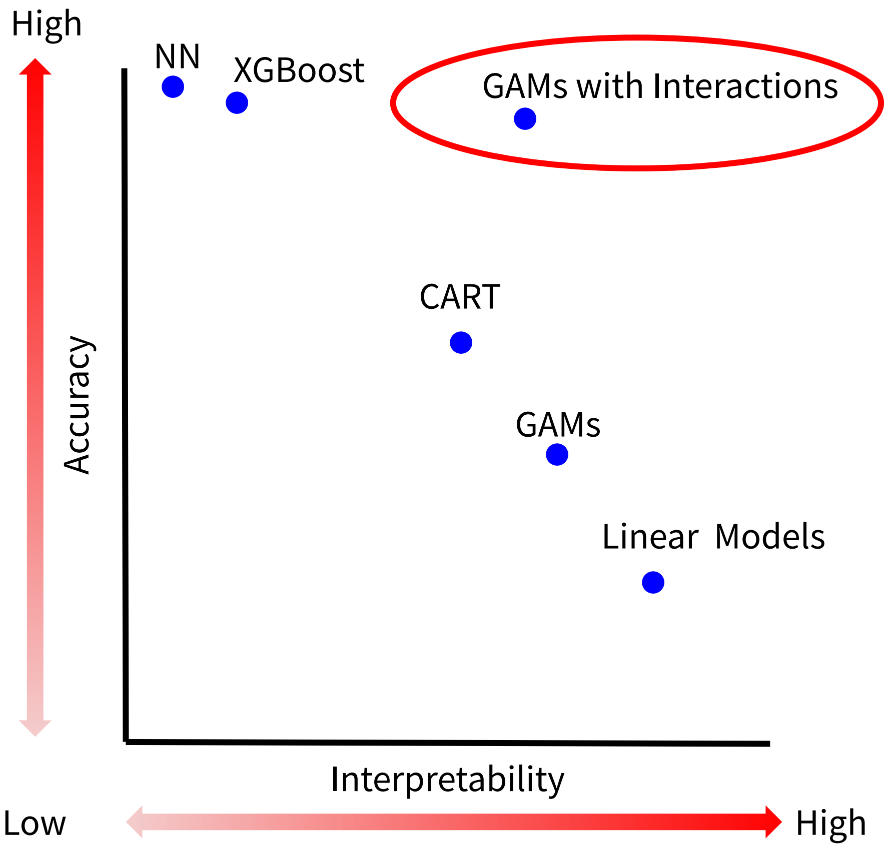
Shibal Ibrahim* Gabriel I. Afriat Kayhan Behdin Rahul Mazumder

Generalized Additive Models (GAMs) with Interactions

- GAMs with Interactions [Hastie (1987)] consider a model of the form:

$$g(\mathbb{E}[y]) = \sum_{j \in [p]} f_j(x_j) + \sum_{(j,k) \in \mathcal{I}} f_{j,k}(x_{j,k})$$

- Class of flexible models
  - Highly Interpretable
  - Provide good performance comparable to black-box methods



Challenges for GAMs with Interactions

- GAMs with Interactions consider a model of the form:

$$g(\mathbb{E}[y]) = \sum_{j\in[p]} f_j(x_j) + \sum_{(j,k)\in\mathcal{I}} f_{j,k}(x_{j,k})$$

- Challenges:
    - Learning all pairwise interaction effects (order ~ p^2) computationally challenging.
    - End-to-end Component selection (few components $\{f_j\}$ and $\{f_{j,k}\}$ to be nonzero) is a hard combinatorial optimization problem.
        - Benefit: Component selection aids interpretability.
    - Structural constraints on the interaction effects, e.g., hierarchy, make optimization more complex
        - Benefit: Improve interpretability, practical sparsity and reduce variance.

Key literature on GAMs with Interactions

Existing methods have the following limitations:

- Not flexible (require customized algorithms to adapt)
  - EBM [Lou et al. (2013), Nori et al. (2019)] , ELAAN [Ibrahim et al. (2021)]
- Do not support component selection in an end-to-end fashion
  - GAMI-Net [Yang et al. (2020)], SIAN [Enouen et al. (2022)]
- Do not support structural constraints
  - EBM [Lou et al. (2013), Nori et al. (2019)], NODE-GAM [Chang et al. (2022)]
- Slow when fitting interactions
  - EBM [Lou et al. (2013), Nori et al. (2019)], SIAN [Enouen et al. (2022)], GAMI-Net [Yang et al. (2020)]

Proposal

1. **GRAND-SLAMIN**: a general optimization framework, which
    a. Works in an **end-to-end** fashion for any differentiable loss function.
    b. Supports **component selection** i.e., selects a sparse subset of main and interaction effects.
    c. Supports **structural constraints** e.g., weak hierarchy and strong hierarchy.
    d. Has **statistical guarantees** — we provide novel non-asymptotic error bounds.
    e. Is GPU-compatible **sparse back-propagation** for efficient training.

Goal

$$f = \sum_{j \in [p]} f_j(x_j) + \sum_{(j,k) \in \mathcal{I}} f_{j,k}(x_{j,k})$$

Component Selection:

$$\mathbf{nnz}(f_j, f_{j,k}) \leq K$$

Structural Constraints:

Weak Hierarchy: $f_{j,k} \neq 0 \implies f_j \neq 0 \ \text{ OR } \ f_k \neq 0$

Strong Hierarchy: $f_{j,k} \neq 0 \implies f_j \neq 0 \ \text{AND} \ f_k \neq 0$

Optimization Formulation

$$\min_{\{f_j\},\,\{f_{j,k}\},\{z_j\}\in\{0,1\}^p,\{z_{j,k}\}\in\{0,1\}^{|\mathcal{I}|}} \hat{\mathbb{E}}[l(y,f)] + \lambda\left(\sum_{j\in[p]} z_j + \alpha \sum_{(j,k)\in\mathcal{I}} z_{j,k}\right)$$

$$f = \sum_{j\in[p]} f_j(x_j)z_j + \sum_{(j,k)\in\mathcal{I}} f_{j,k}(x_{j,k})q(z_j, z_k, z_{j,k})$$

No structural constraint: $q(z_j, z_j, z_{j,k}) = z_{j,k}$

Weak Hierarchy: $q(z_j, z_j, z_{j,k}) = (z_j + z_k - z_j z_k)z_{j,k}$

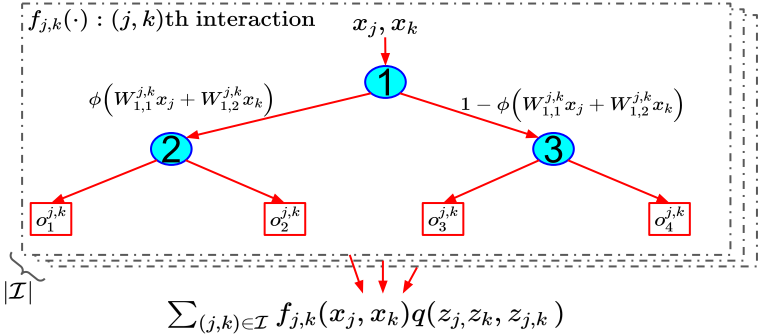
Strong Hierarchy: $q(z_j, z_j, z_{j,k}) = z_j z_k z_{j,k}$

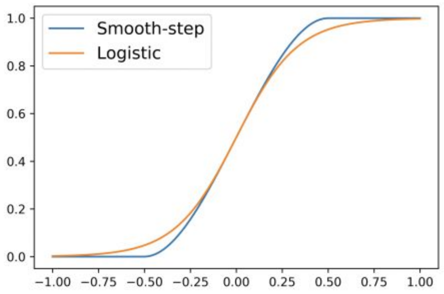
Smooth Reformulation

$$f = \sum_{j \in [p]} f_j(x_j) z_j + \sum_{(j,k) \in \mathcal{I}} f_{j,k}(x_{j,k}) q(z_j, z_k, z_{j,k})$$

Parameterize as follows:

- Components, i.e., f_j and $f_{j,k}$ with Soft trees [Jordan and Jacob (1993)]
- Smooth binary variables, i.e., z_j, z_k and $z_{j,k}$ with Smooth-Step function [Hazimeh et al. (2020)]





Allows optimization with first-order methods e.g., SGD!

Statistical Theory Takeaways:

- First to discuss statistical properties of GAMs with interactions with **tree-shape functions**

- Under a well-specified model, non-asymptotic prediction error **rates of $n^{-2/3}$** and **$n^{-1/(2+a)} \approx n^{-0.42}$** are achievable for main effects and interaction models, respectively.
  - Prediction error (resulting from the noise in observations) converges to zero as we increase the total number of samples, n.

- Asymptotically, when n → ∞ and other parameters in the problem stay constant, an **error rate of $n^{-0.5}$** is achievable for the interactions model

Results

Comparison with Sparse GAMs with interactions

- Competitive with EB^2M and $NODE\text{-}GA^2M$
- Our key advantages:
  - Hierarchical interactions (not supported by $NODE\text{-}GA^2M$ and EB^2M).
  - faster training times
  - Improved variable selection.

Dataset	**EB^2M**	**$NODE\text{-}GA^2M$**	**GRAND-SLAMIN (ours)**
Magic	93.12 ± 0.001	**94.27** ± 0.13	93.86 ± 0.3
Adult	91.41 ± 0.0004	**91.75** ± 0.14	91.54 ± 0.14
Churn	91.97 ± 0.005	89.62 ± 5.61	**92.40** ± 0.41 (SH)
Satimage	97.65 ± 0.0007	98.7 ± 0.07	**98.81** ± 0.04
Texture	99.81 ± 0.0004	**100.0** ± 0.0	**100.0** ± 0.0
MiniBooNE	97.86 ± 0.0001	**98.44** ± 0.02	97.77 ± 0.05 (WH)
Covertype	90.08 ± 0.0003	95.39 ± 0.12	**98.11** ± 0.08
Spambase	**98.84** ± 0.01	98.78 ± 0.06	98.55 ± 0.07 (SH)
News	73.03 ± 0.002	**73.53** ± 0.06	73.24 ± 0.04 (SH)
Optdigits	99.79 ± 0.0003	99.93 ± 0.02	**99.98** ± 0.0
Bankruptcy	**93.85** ± 0.01	92.02 ± 1.03	92.51 ± 0.54 (WH)
Madelon	88.04 ± 0.02	60.07 ± 0.82	**89.25** ± 1.03 (WH)
Activity	74.96 ± 8.77	99.86 ± 0.04	**99.24** ± 1.45
Multiple	**99.96** ± 0.0002	99.94 ± 0.02	99.95 ± 0.02

Comparison with Sparse Hierarchical interactions

- Our models outperform GAMI-Net and SIAN in many datasets.
- Our key advantages:
    - Our Hierarchical interactions is end-to-end.
    - Faster training times
    - Improved variable selection.

	Weak Hierarchy		Strong Hierarchy	
Dataset	**GAMI-Net**	**GRAND-SLAMIN**	**SIAN**	**GRAND-SLAMIN**
Magic	91.72 ± 0.05	**93.16** ± 0.55	93.02 ± 0.06	**93.37** ± 0.16
Adult	91.01 ± 0.04	**91.34** ± 0.32	90.67 ± 0.05	**91.46** ± 0.15
Churn	90.05 ± 0.77	**92.28** ± 0.75	**92.98** ± 0.20	92.40 ± 0.41
Spambase	**98.67** ± 0.04	98.45 ± 0.15	98.28 ± 0.04	**98.55** ± 0.07
MiniBooNE	96.11 ± 0.41	**97.77** ± 0.05	95.9	**97.62** ± 0.30
News	72.54 ± 0.05	**73.15** ± 0.08	72.28	**73.24** ± 0.04
Bankruptcy	92.46 ± 0.12	**92.51** ± 0.54	**90.71**	90.45 ± 1.87
Madelon	88.14 ± 0.94	**89.25** ± 1.03	83.18	**86.23** ± 1.89

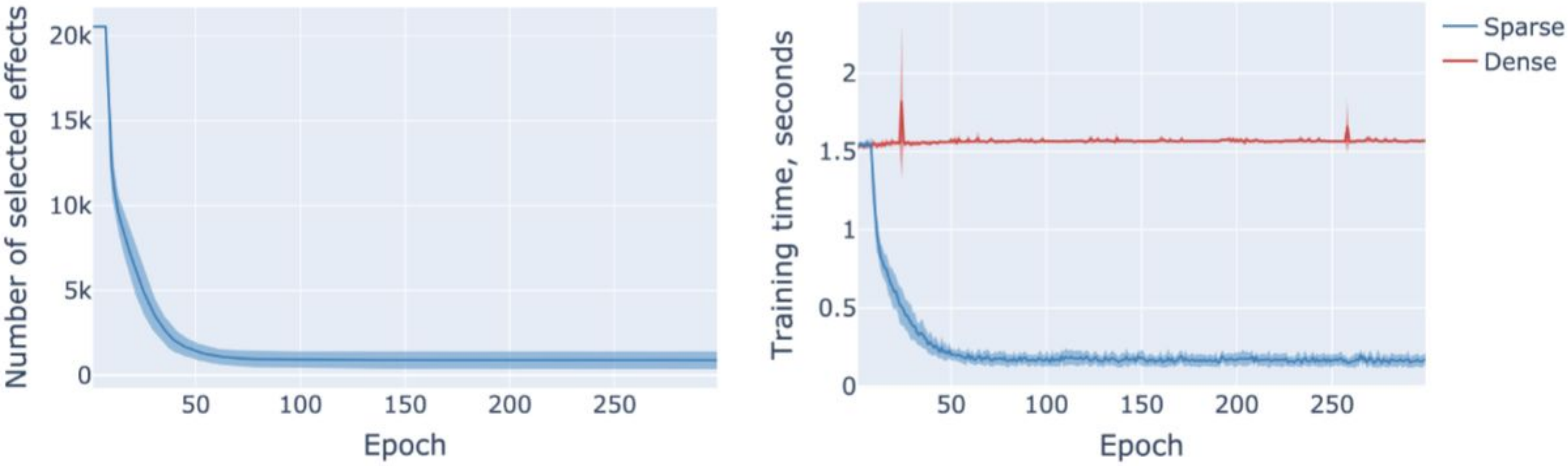
WH = Weak Hierarchy, SH=Strong Hierarchy

Variable Selection

GRAND-SLAMIN with structural constraints, in particular SH, can reduce the number of features selected.

	EB²M	NODE-GA²M	GAMI-Net	SIAN	GRAND-SLAMIN (ours)		
Dataset	**None**	**None**	**WH**	**SH**	**None**	**WH**	**SH**
Magic	10 ± 0	10 ± 0	10 ± 0	10 ± 0	10 ± 0	**9** ± 1	**7** ± 0
Adult	14 ± 0	14 ± 0	14 ± 1	14 ± 0	**13** ± 1	**11** ± 1	**11** ± 1
Churn	19 ± 0	19 ± 0	18 ± 2	19 ± 0	19 ± 0	**11** ± 1	**12** ± 2
Satimage	36 ± 0	36 ± 0	–	–	36 ± 0	36 ± 0	**22** ± 2
Texture	40 ± 0	40 ± 0	–	–	40 ± 0	**37** ± 2	**17** ± 2
MiniBooNE	50 ± 0	50 ± 0	**16** ± 12	34	50 ± 0	50 ± 0	28 ± 3
Covertype	54 ± 0	54 ± 0	–	–	**34** ± 1	54 ± 1	54 ± 0
Spambase	57 ± 0	57 ± 0	**52** ± 2	55 ± 1	57 ± 0	56 ± 3	54 ± 2
Bankruptcy	95 ± 0	95 ± 0	60 ± 15	69	95 ± 0	60 ± 26	**7** ± 16
Madelon	500 ± 0	500 ± 0	61 ± 56	490	**26** ± 19	**19** ± 15	**24** ± 9
Activity	533 ± 0	346 ± 6	–	–	**182** ± 15	440 ± 22	**159** ± 21
Multiple	649 ± 0	649 ± 0	–	–	648 ± 1	**629** ± 9	649 ± 0

Efficient Training with Sparse Backpropagation



(a) Number of selected effects at each epoch.

(b) Training time (seconds) for each epoch.

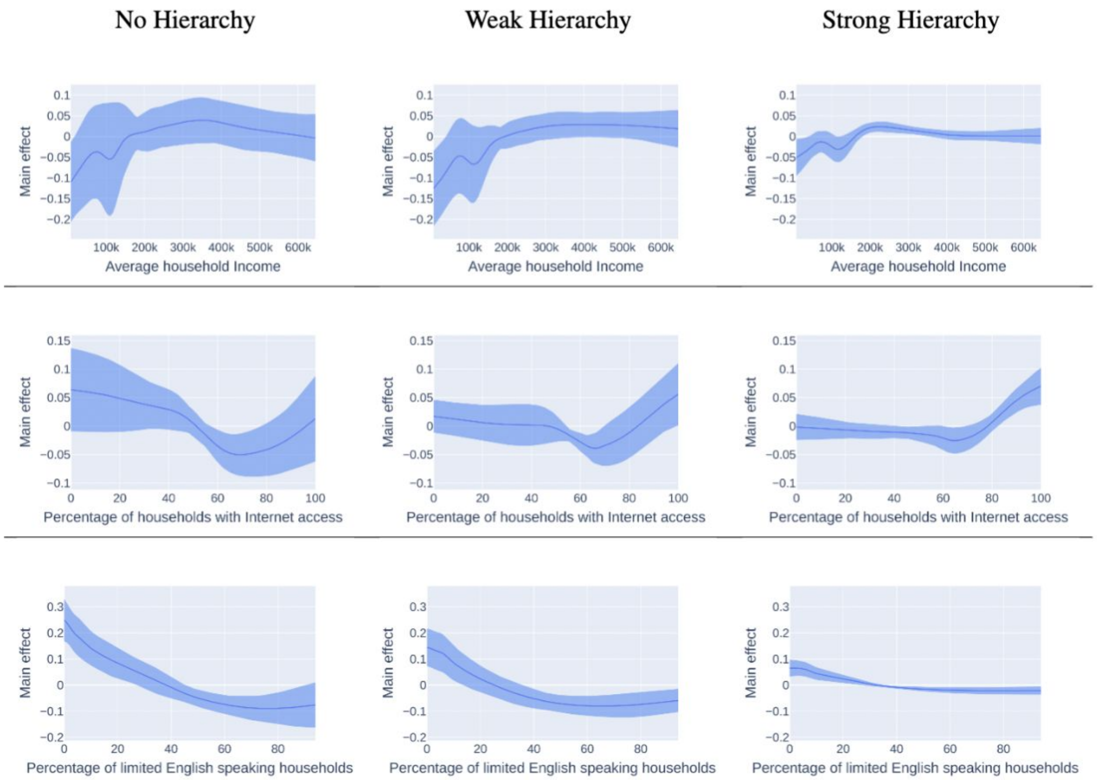
Sparse backpropagation up to 10× faster than standard backpropagation - no loss in accuracy

- Components with zero z’s are removed from computational graph during training

Variance reduction with structural constraints

Estimation of main effects (in the presence of interaction effects) is more stable with structural constraints

- Smaller error bars across seeds/runs!



Check out our paper!

Paper: https://openreview.net/pdf?id=F5DYsAc7Rt
GRAND-SLAMIN Code: https://github.com/mazumder-lab/grandslamin
Email: shibal@mit.edu

References

- Chun-Hao Chang, Rich Caruana, and Anna Goldenberg. NODE-GAM: Neural generalized additive model for interpretable deep learning. In ICLR 2022.
- James Enouen and Yan Liu. Sparse interaction additive networks via feature interaction detection and sparse selection. In NeurIPS, 2022.
- Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. Journal of the American Statistical Association, 82(398):371–386, 1987.
- Hussein Hazimeh, Natalia Ponomareva, Petros Mol, et al. The tree ensemble layer: Differentiability meets conditional computation. In ICML 2020.
- Shibal Ibrahim, Wenyu Chen, Hussein Hazimeh, Natalia Ponomareva, Zhe Zhao, and Rahul Mazumder. Comet: Learning cardinality constrained mixture of experts with trees and local search. In KDD '23.
- Shibal Ibrahim, Hussein Hazimeh, and Rahul Mazumder. Flexible modeling and multitask learning using differentiable tree ensembles. In KDD ’22.
- Shibal Ibrahim, Rahul Mazumder, Peter Radchenko, and Emanuel Ben-David. Predicting Census Survey Response Rates With Parsimonious Additive Models and Structured Interactions, arXiv, abs/2108.11328, 2021.
- Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the em algorithm. Neural Comput., 6(2):181–214, mar 1994.
- Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In KDD ’12.
- Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In KDD ’13.
- Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. ArXiv, abs/1909.09223, 2019.
- Zebin Yang, Aijun Zhang, and A. Sudjianto. Gami-net: An explainable neural network based on generalized additive models with structured interactions. Pattern Recognit., 120:108192, 2021.