

Regionally Additive Models: Explainable-by-design models minimizing feature interactions

Vasilis Gkolemis^{1,2} Anargiros Tzerefos¹ Theodore Dalamagas¹
Eirini Ntoutsis³ Christos Diou²

¹ATHENA Research and Innovation Center

²Harokopio University of Athens

³Universitat der Bundeswehr Munchen

September 2023, Turin, Italy

Generalized Additive Models (GAMs)

Wikipedia says:

In statistics, a generalized additive model (GAM) is a generalized linear model in which the response variable depends linearly on unknown smooth functions of some predictor variables.

Generalized Additive Models (GAMs)

Wikipedia says:

*In statistics, a generalized additive model (GAM) is a generalized linear model in which the **response** variable depends linearly on unknown smooth functions of some predictor variables.*

y

Generalized Additive Models (GAMs)

Wikipedia says:

*In statistics, a generalized additive model (GAM) is a generalized linear model in which the **response** variable depends **linearly** on unknown smooth functions of some predictor variables.*

$$y = \cdot + \dots + \cdot$$

Generalized Additive Models (GAMs)

Wikipedia says:

*In statistics, a generalized additive model (GAM) is a generalized linear model in which the **response** variable depends **linearly** on unknown **smooth functions of some predictor variables**.*

$$y = f_1(x_1) + \dots + f_D(x_D)$$

Introductory Example

Output/target variable:

- $y_{\text{bike-rentals}}$: the expected number of bike rentals per hour

Input/covariates:

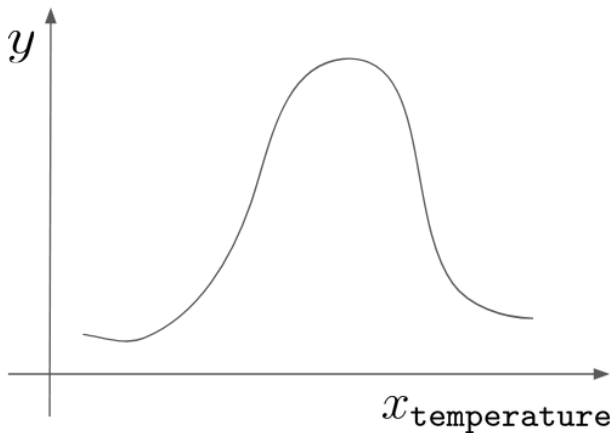
- $x_{\text{temperature}}$: temperature per hour
- x_{humidity} : humidity per hour
- $x_{\text{is_weekday}}$: if it is weekday or weekend

Let's fit a GAM:

$$y = f_1(x_{\text{temperature}}) + f_2(x_{\text{humidity}}) + f_3(x_{\text{is_weekday}})$$

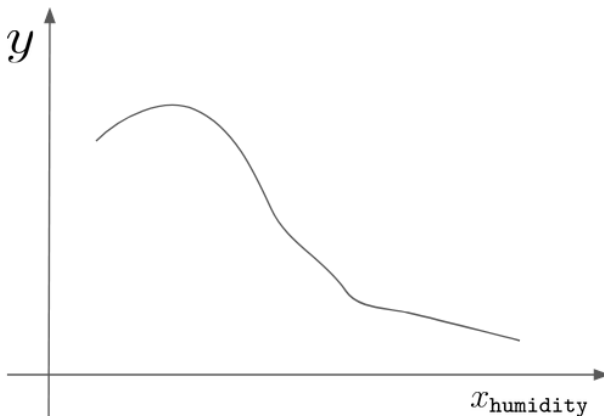
GAMs - Interpretability (1)

$$f_1(x_{\text{temperature}})$$



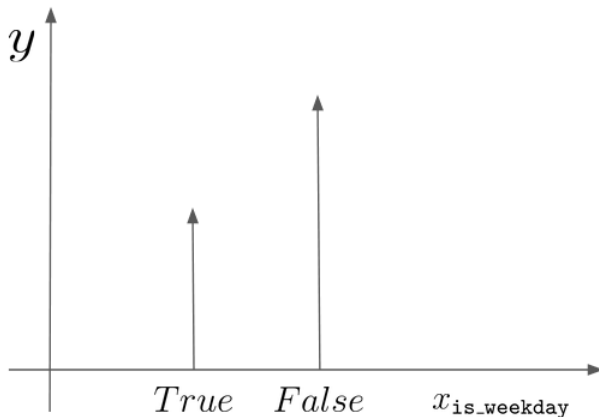
GAMs - Interpretability (2)

$$f(x_{\text{humidity}})$$



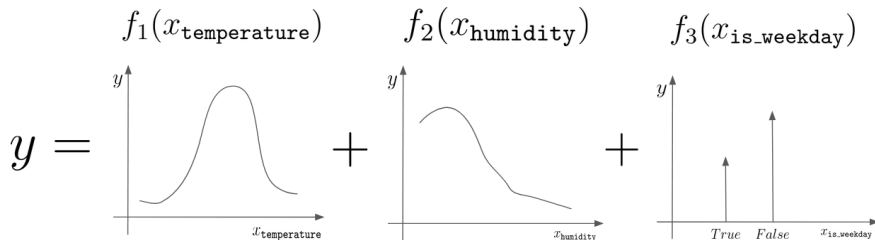
GAMs - Interpretability (3)

$$f(x_{\text{is_weekday}})$$



GAMs - Interpretability (4)

GAMs is explainable!



GAMs - Limitations/Extensions

Limitations:

Extensions:

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends

Extensions:

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends
- Cause: go to work vs go sightseeing

Extensions:

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends
- Cause: go to work vs go sightseeing
- Solution 1: Add pairwise term $f(x_{\text{temperature}}, x_{\text{is_weekday}})$

Extensions:

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends
- Cause: go to work vs go sightseeing
- Solution 1: Add pairwise term $f(x_{\text{temperature}}, x_{\text{is_weekday}})$
- Solution 2: Model two conditional terms
 - ▶ $f(x_{\text{temperature}} | \text{weekday})$
 - ▶ $f(x_{\text{temperature}} | \text{weekend})$

Extensions:

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends
- Cause: go to work vs go sightseeing
- Solution 1: Add pairwise term $f(x_{\text{temperature}}, x_{\text{is_weekday}})$
- Solution 2: Model two conditional terms
 - ▶ $f(x_{\text{temperature}} | \text{weekday})$
 - ▶ $f(x_{\text{temperature}} | \text{weekend})$

Extensions:

- Solution 1: $GA^2M = \text{GAM} + \text{pairwise interactions}$ (Yin Lou et. al)

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends
- Cause: go to work vs go sightseeing
- Solution 1: Add pairwise term $f(x_{\text{temperature}}, x_{\text{is_weekday}})$
- Solution 2: Model two conditional terms
 - ▶ $f(x_{\text{temperature}} | \text{weekday})$
 - ▶ $f(x_{\text{temperature}} | \text{weekend})$

Extensions:

- Solution 1: $GA^2M = \text{GAM} + \text{pairwise interactions}$ (Yin Lou et. al)
- Solution 2: $RAM = \text{GAM}$ at subregions

GAMs - Limitations/Extensions

Limitations:

- temperature has different effect on week-days vs weekends
- Cause: go to work vs go sightseeing
- Solution 1: Add pairwise term $f(x_{\text{temperature}}, x_{\text{is_weekday}})$ Explainable
- Solution 2: Model two conditional terms
 - ▶ $f(x_{\text{temperature}} | \text{weekday})$ Explainable
 - ▶ $f(x_{\text{temperature}} | \text{weekend})$ Explainable

Extensions:

- Solution 1: $GA^2M = \text{GAM} + \text{pairwise interactions}$ (Yin Lou et. al)
- Solution 2: $RAM = \text{GAM at subregions}$

$RA^{(2)}Ms$ go even beyond

GA^2Ms Limitations:

$RA^{(2)}Ms$ solve that:

$RA^{(2)}$ Ms go even beyond

GA^2 Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?

$RA^{(2)}$ Ms solve that:

$RA^{(2)}$ Ms go even beyond

GA^2 Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!

$RA^{(2)}$ Ms solve that:

$RA^{(2)}$ Ms go even beyond

GA^2 Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!
- But if it workday? and bike is the only transport?

$RA^{(2)}$ Ms solve that:

$RA^{(2)}$ Ms go even beyond

GA^2 Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!
- But if it workday? and bike is the only transport?
- model $f(x_{\text{temperature}}, x_{\text{humidity}}, x_{\text{is_weekday}})$?

$RA^{(2)}$ Ms solve that:

$RA^{(2)}$ Ms go even beyond

GA^2 Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!
- But if it workday? and bike is the only transport?
- model $f(x_{\text{temperature}}, x_{\text{humidity}}, x_{\text{is_weekday}})$? **Not explainable**

$RA^{(2)}$ Ms solve that:

$RA^{(2)}$ Ms go even beyond

GA^2 Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!
- But if it workday? and bike is the only transport?
- model $f(x_{\text{temperature}}, x_{\text{humidity}}, x_{\text{is_weekday}})$? **Not explainable**

$RA^{(2)}$ Ms solve that:

- $f(x_{\text{temperature}}, x_{\text{humidity}} | x_{\text{is_weekday}}) \rightarrow RA^2M$

$RA^{(2)}$ Ms go even beyond

GA^2 Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!
- But if it workday? and bike is the only transport?
- model $f(x_{\text{temperature}}, x_{\text{humidity}}, x_{\text{is_weekday}})$? **Not explainable**

$RA^{(2)}$ Ms solve that:

- $f(x_{\text{temperature}}, x_{\text{humidity}} | x_{\text{is_weekday}}) \rightarrow RA^2M$
- $f(x_{\text{temperature}} | x_{\text{humidity}} = \{high, low\}, x_{\text{is_weekday}}) \rightarrow$ RAM with two conditions

$RA^{(2)}$ Ms go even beyond

GA^2 Ms Limitations:

- Have you ever ridden a bike in a cold day with humidity?
- If it is weekend, let's see a movie instead!
- But if it workday? and bike is the only transport?
- model $f(x_{\text{temperature}}, x_{\text{humidity}}, x_{\text{is_weekday}})$? **Not explainable**

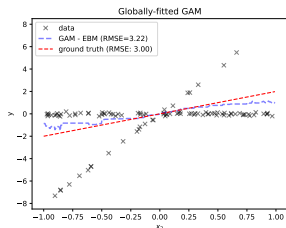
$RA^{(2)}$ Ms solve that:

- $f(x_{\text{temperature}}, x_{\text{humidity}} | x_{\text{is_weekday}}) \rightarrow RA^2M$ **Explainable**
- $f(x_{\text{temperature}} | x_{\text{humidity}} = \{high, low\}, x_{\text{is_weekday}}) \rightarrow$ RAM with two conditions **Explainable**

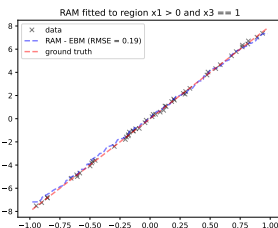
RAM on toy example

$$f(\mathbf{x}) = 8x_2 \mathbb{1}_{x_1 > 0} \mathbb{1}_{x_3 = 0}$$

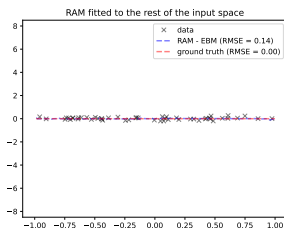
$$x_1, x_2 \sim \mathcal{U}(-1, 1), x_3 \sim \text{Bernoulli}(0, 1)$$



(a) $f_2(x_2)$



(b) $f_2(x_2) \mathbb{1}_{x_1 > 0 \text{ and } x_3 = 1}$



(c) $f_2(x_2) \mathbb{1}_{x_1 \leq 0 \text{ or } x_3 \neq 1}$

Figure: (Left) GAM, (Middle and Right) RAM

How RAM works

3-step approach:

How RAM works

3-step approach:

- Fit a black-box model to learn complex feature interactions
 - ▶ it should be differentiable
 - ▶ neural network is a good option

How RAM works

3-step approach:

- Fit a black-box model to learn complex feature interactions
 - ▶ it should be differentiable
 - ▶ neural network is a good option
- Use a Regional Effect method to isolate the important interactions
 - ▶ [RHALE - Gkolemis et. al](#)
 - ▶ [Feature Interactions - Herbinger et. al](#)
 - ▶ finds which features $f(x_i)$ should be split into subregions $f(x_i | x_j \leq \tau)$

How RAM works

3-step approach:

- Fit a black-box model to learn complex feature interactions
 - ▶ it should be differentiable
 - ▶ neural network is a good option
- Use a Regional Effect method to isolate the important interactions
 - ▶ [RHALE - Gkolemis et. al](#)
 - ▶ [Feature Interactions - Herbinger et. al](#)
 - ▶ finds which features $f(x_i)$ should be split into subregions $f(x_i|x_j \leq \tau)$
- Fit a univariate function on each detected subregion
 - ▶ learn all $f(x_i|x_j \leq \tau)$

Step 1

- Fit a black-box model to capture all complex structures
 - ▶ it should be differentiable
 - ▶ A neural network is a good option

Step 2

- Regional Effect method to find important interactions
 - ▶ [RHALE - Gkolemis et. al](#)
 - ▶ [Feature Interactions - Herbinger et. al](#)
- Idea:
 - ▶ Feature effect is the average effect of each feature x_s on the output y
 - ▶ It is computed by averaging the instance-level effects
 - ▶ Heterogeneity \mathcal{H} (or uncertainty) measures the deviation of the instance-level effects from the average effect
 - ▶ we want to split the dataset in subgroups in order to minimize the heterogeneity
- In mathematical terms:

$$\underbrace{\mathcal{H}(f_i(x_i))}_{\mathcal{H} \text{ before split}} \gg \underbrace{\mathcal{H}(f_i(x_i|x_j > \tau)) + \mathcal{H}(f_i(x_i|x_j \leq \tau))}_{\text{sum of } \mathcal{H} \text{ after split}}$$

Step 3

- Step 2 defines a new feature space \mathcal{X}^{RAM}
- Every feature is split to T_s subregions which are defined by \mathcal{R}_{st} :

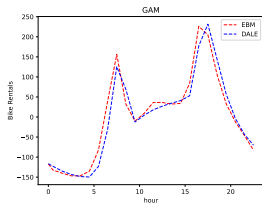
$$\begin{aligned}\mathcal{X}^{\text{RAM}} &= \{x_{st} | s \in \{1, \dots, D\}, t \in \{1, \dots, T_s\}\} \\ x_{st} &= \begin{cases} x_s, & \text{if } \mathbf{x}_{/s} \in \mathcal{R}_{st} \\ 0, & \text{otherwise} \end{cases}\end{aligned}\tag{1}$$

- Fit a univariate function on each subregion:

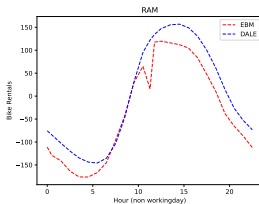
$$f^{\text{RAM}}(\mathbf{x}) = c + \sum_{s,t} f_{st}(x_{st}) \quad \mathbf{x} \in \mathcal{X}^{\text{RAM}}\tag{2}$$

Bike Sharing dataset

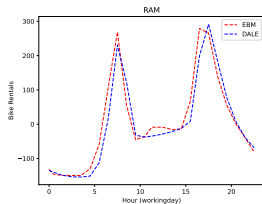
Predict bike-rentals per hour



(a) $f(X_{\text{hour}})$



(b) $f(X_{\text{hour}})\mathbb{1}_{X_{\text{workingday}} \neq 1}$



(c) $f(X_{\text{hour}})\mathbb{1}_{X_{\text{workingday}} = 1}$

Experimental Results

Tested on [Bike Sharing](#) and [California Housing](#) Datasets.

	Black-box	x-by-design			
	all orders	1 st order		2 nd order	
	DNN	GAM	RAM	GA²M	RA²M
Bike (MAE)	0.254	0.549	0.430	0.298	0.278
Bike (RMSE)	0.389	0.734	0.563	0.438	0.412
Housing (MAE)	0.373	0.600	0.553	0.554	0.533
Housing (RMSE)	0.533	0.819	0.754	0.774	0.739

What is next?

- Results are preliminary
 - ▶ Compare RAM vs GAM and RA^2M vs GA^2M in more datasets
 - ▶ Check robustness on edge cases:
 - ★ highly correlated features
 - ★ limited training examples
- Can we model uncertainty?
 - ▶ Uncertain because we do not model higher-order interactions
 - ▶ Uncertain about the conditionals, i.e., detected subregions
 - ▶ Uncertain about the univariate functions we learn
- Could we make it a 1-step process?
 - ▶ a network that automatically learns both the univariate functions and the conditions

Thank you for your attention

- For more discussion or future ideas on RAM, please, contact me:
 - ▶ vgkolemis@athenarc.gr
 - ▶ gkolemis@hua.gr
- More info about the paper: <https://arxiv.org/abs/2309.12215>



- Questions?