CS-6316 Machine Learning

Detecting Statistical Interactions from Neural Network Weights

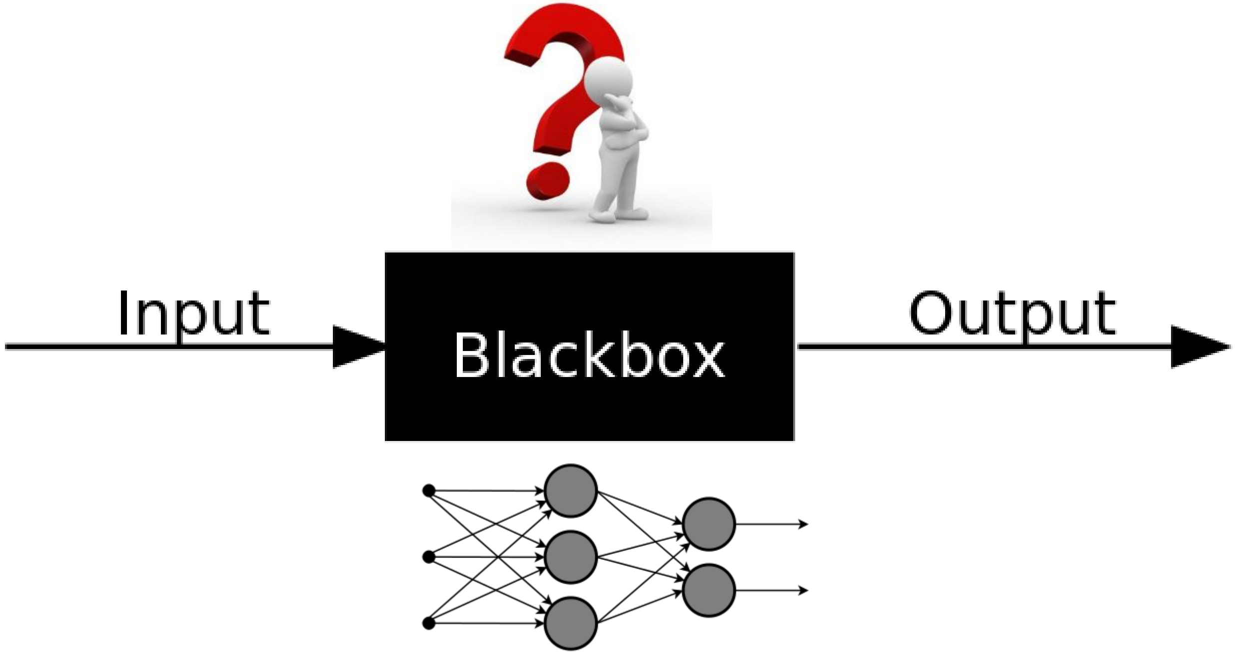
M. Tsang, D. Cheng, Y. Liu – ICLR 2018

Reproduced By:

Magda Amiridi

December 6 , 2019

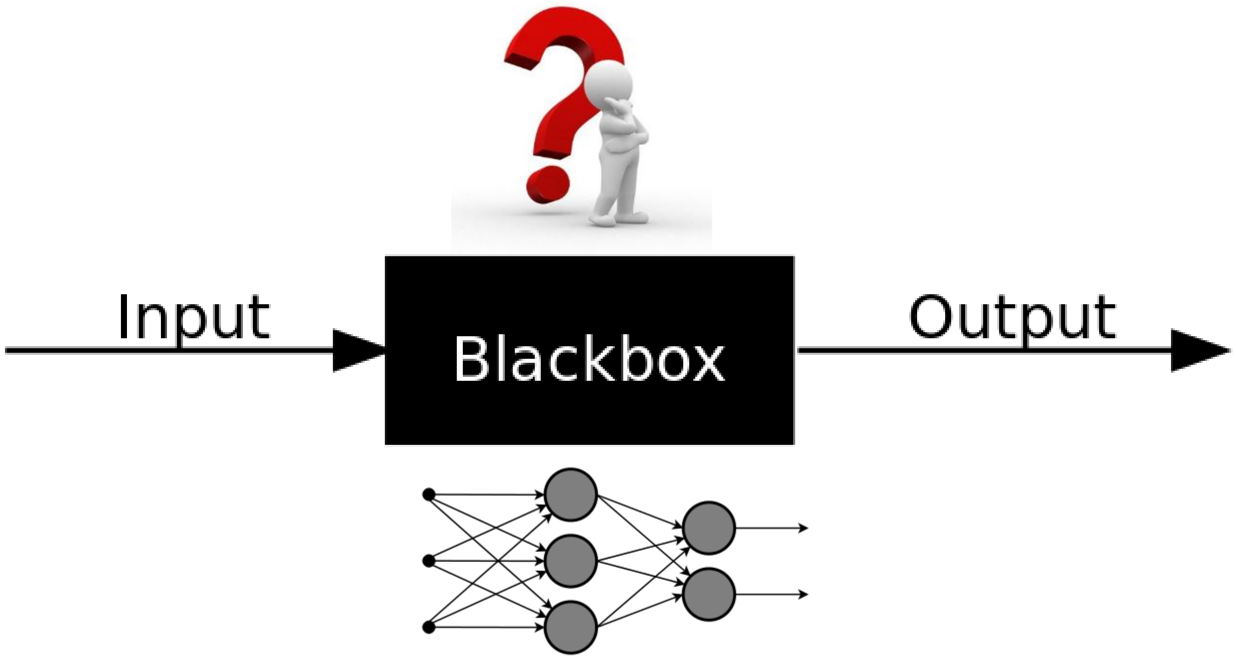
Motivation



Feedforward NNs

- Universal function approximators
- Interpretability

Motivation



Feedforward NNs

- Universal function approximators
- Interpretability

Main goal : ⟶ **detecting** pairwise and high-order **feature interactions** in a dataset by re-interpreting **weights** learned by a **MLP**.

Motivation

➢ **Applications**
- Healthcare: Drug–drug interaction (DDI), co-occurrence of a group of symptoms
- Scientific discoveries, hypothesis validation

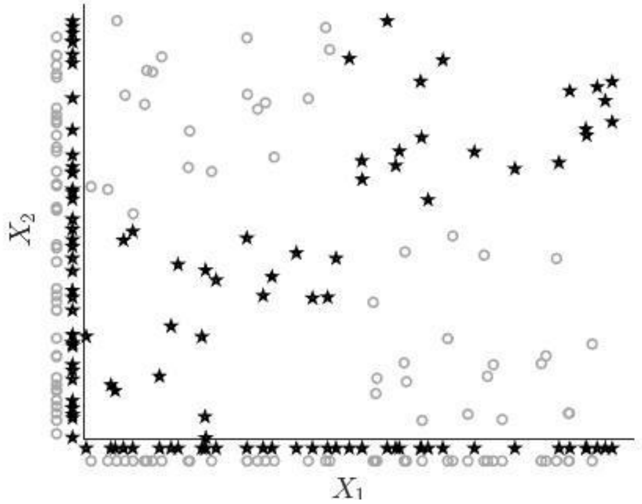
➢ **Challenges**
- p features: Search space size - $O(2^p)$ possible interactions

➢ **Contribution of NID (Neural Interaction Detection)**
- Non-linear feature interactions.
- Invariant of order
- Efficiency

Definition

Interaction: groups of features that have joint effects (non-additive) for predicting an outcome. $\mathcal{I} \subseteq \{1, 2, ..., p\}, |\mathcal{I}| \geq 2$

Geometric example



Simple examples of explicit functions

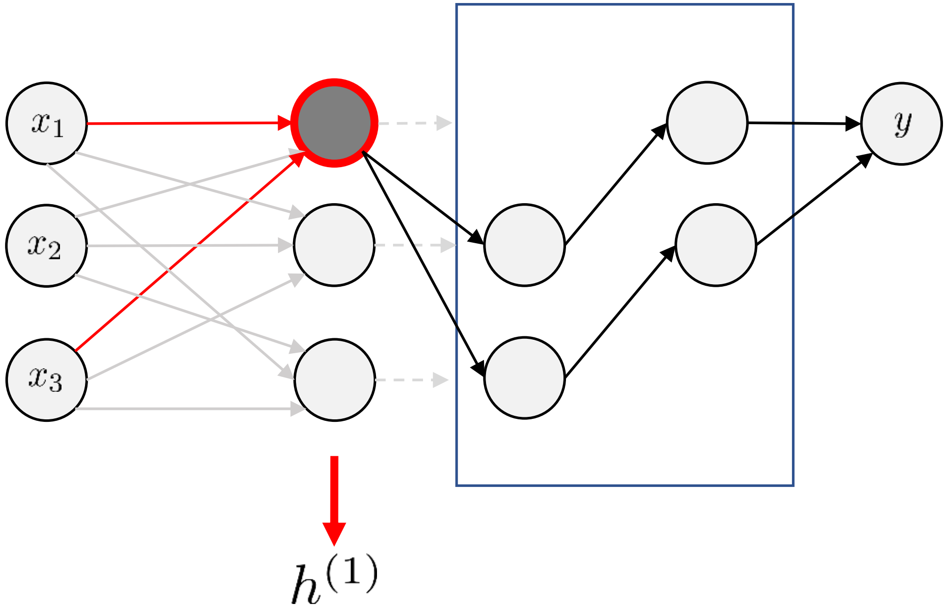
$$f_1(\mathbf{x}) = \sin(x_1 + x_2 + x_3) + x_3 x_4 + x_5$$

$$\{1, 2, 3\} \qquad \{3, 4\}$$

$$f_2(\mathbf{x}) = \log(x_1 x_2) = \log(x_1) + \log(x_2)$$

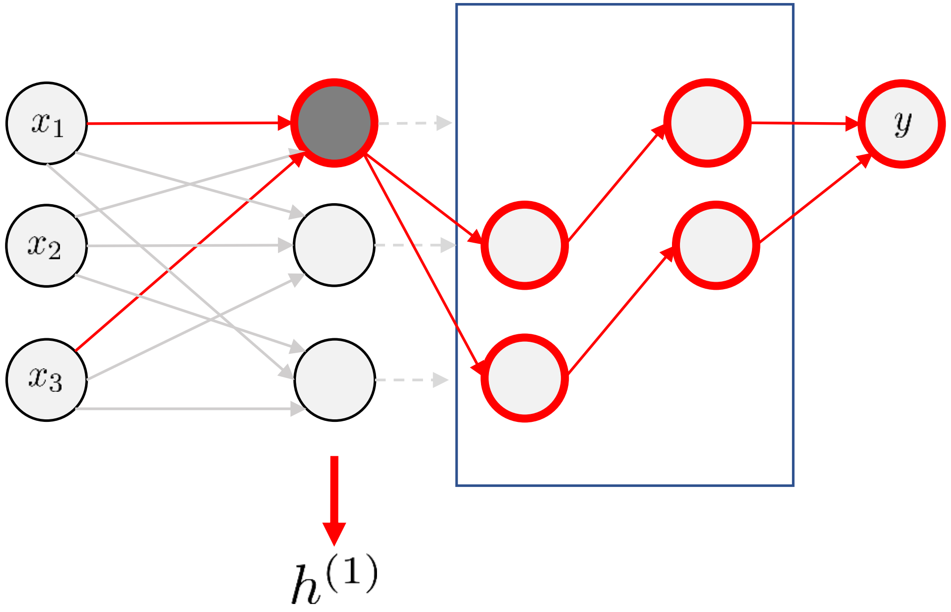
no interaction!

Core Insight Feedforward NNs



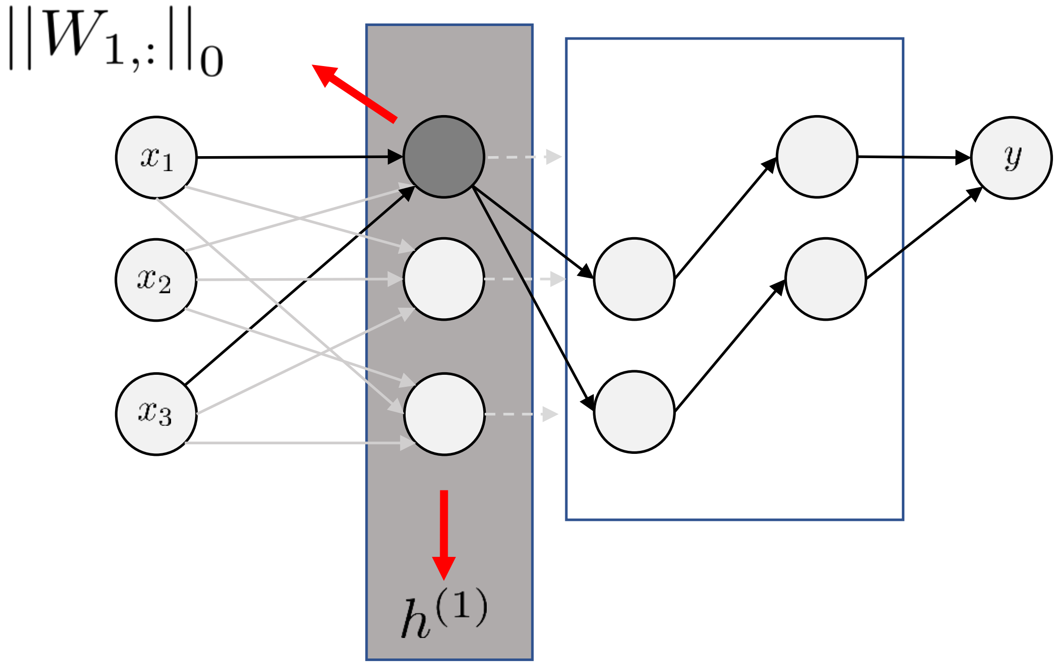
Feature interactions are **created** at hidden units with non-linear activation functions.

Core Insight Feedforward NNs



The influences of the interactions are **propagated** layer-by-layer to the final output.

Core Insight Feedforward NNs



- In general, the weights in a NN are nonzero ⟶ all features are interacting ⟶ large solution space of interactions.
  - ➢ **Assume** *first layer hidden units* are especially good at modeling interactions
  - ➢ Interaction strength.

Interaction strength

Strength $\omega_i(I)$ of an interaction, $\mathrm{I} \subseteq [\mathrm{p}]$ at the i-th unit in the first hidden layer

$$\omega_i(I) = z_i^{(1)} \mu(|W_{i,I}^{(1)}|)$$

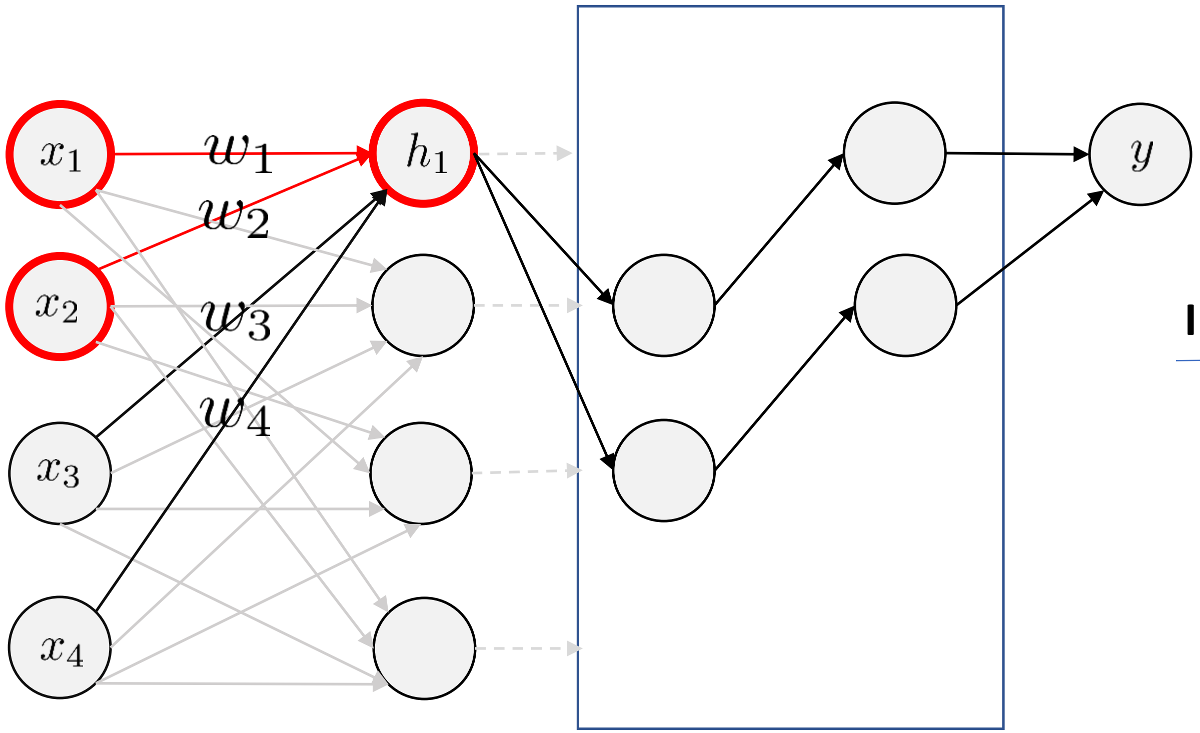
1. **Interactions** created at the **first hidden layer**.
Summarize feature weights between **l = 0** and **l = 1** through function **μ**:

$$\mu(|W_{i,I}^{(1)}|) \longrightarrow \mu(.) = min()$$

2. **Influence of hidden units:** multiplication of the aggregated weight

$$z_i^{(1)} = |w^y|^T |W^{(L)}||W^{(L-1)}|...|W^{(2)}|$$

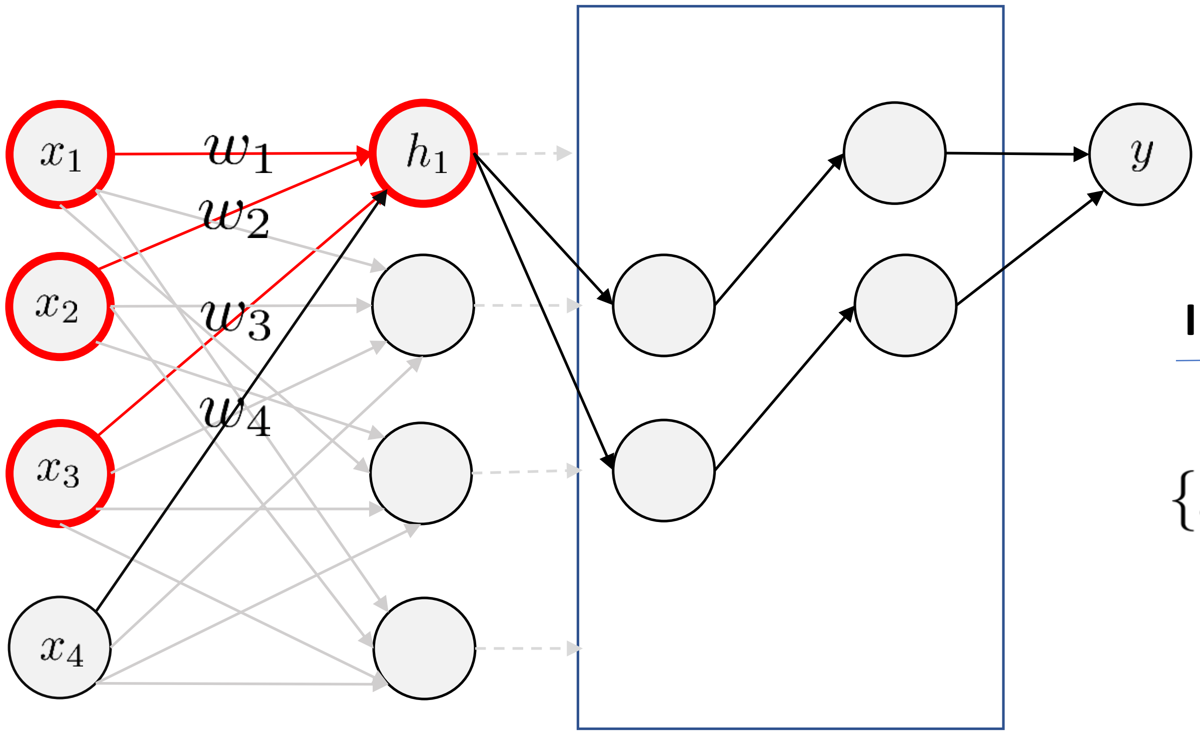
NID example



Interactions	Strength
$\{x_1, x_2\}$	$\min(\lvert w_1\rvert, \lvert w_2\rvert)z_1 = \lvert w_2\rvert z_1$

$$|w_1| > |w_2| > |w_3| > |w_4|$$

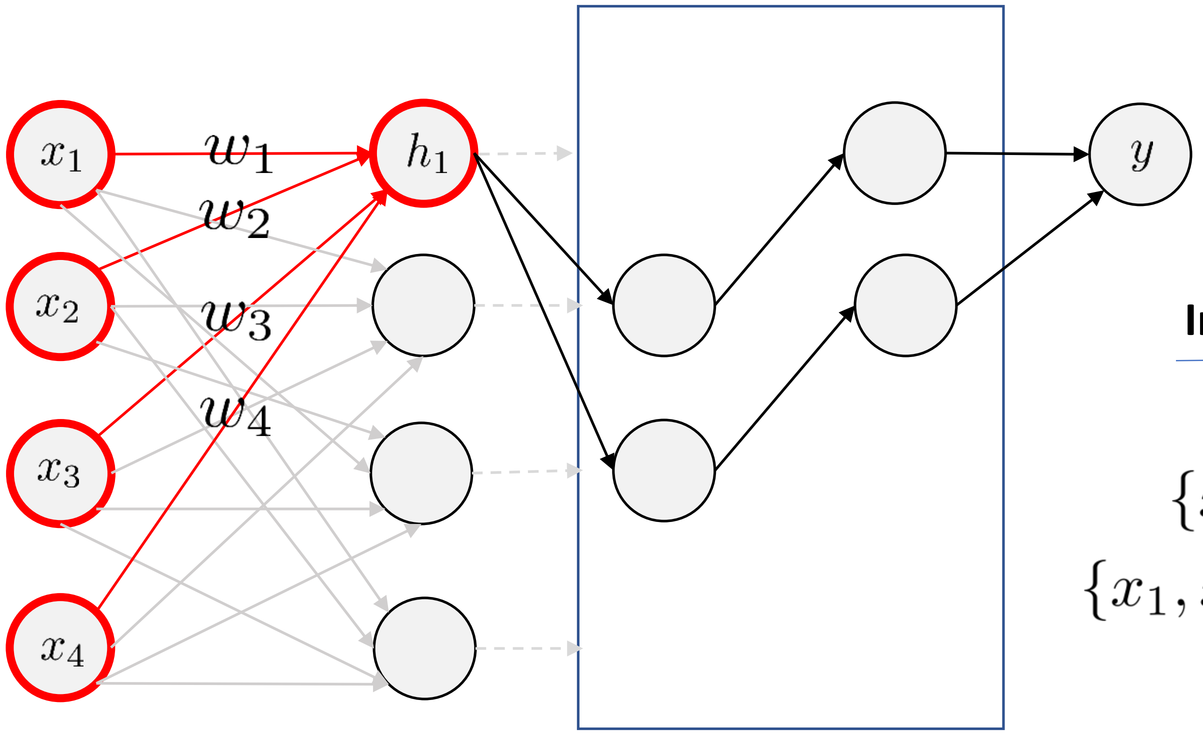
NID example



Interactions	Strength
$\{x_1, x_2\}$	$\min(\lvert w_1\rvert, \lvert w_2\rvert) z_1 = \lvert w_2\rvert z_1$
$\{x_1, x_2, x_3\}$	$\min(\lvert w_1\rvert, \lvert w_2\rvert, \lvert w_3\rvert) z_1 = \lvert w_3\rvert z_1$

$$|w_1| > |w_2| > |w_3| > |w_4|$$

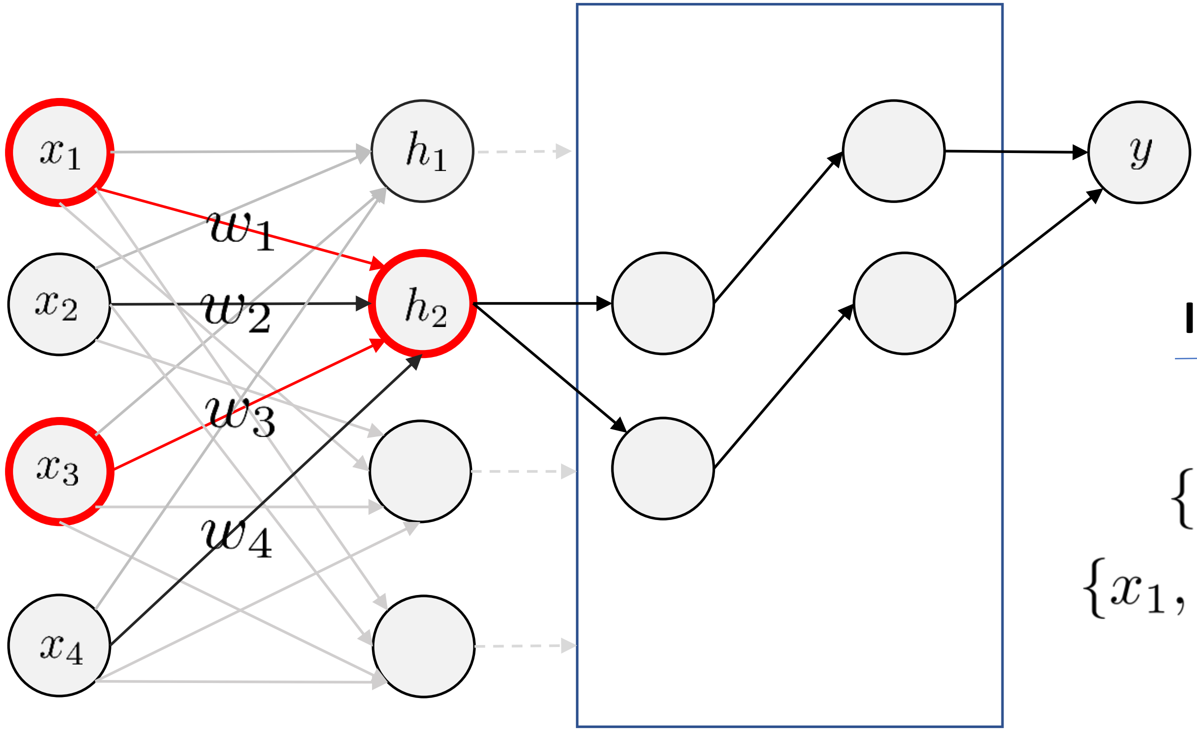
NID example



Interactions	Strength
$\{x_1, x_2\}$	$\min(\lvert w_1\rvert, \lvert w_2\rvert)z_1 = \lvert w_2\rvert z_1$
$\{x_1, x_2, x_3\}$	$\min(\lvert w_1\rvert, \lvert w_2\rvert, \lvert w_3\rvert)z_1 = \lvert w_3\rvert z_1$
$\{x_1, x_2, x_3, x_4\}$	$\min(\lvert w_1\rvert, \lvert w_2\rvert, \lvert w_3\rvert, \lvert w_4\rvert)z_1 = \lvert w_4\rvert z_1$

$$|w_1| > |w_2| > |w_3| > |w_4|$$

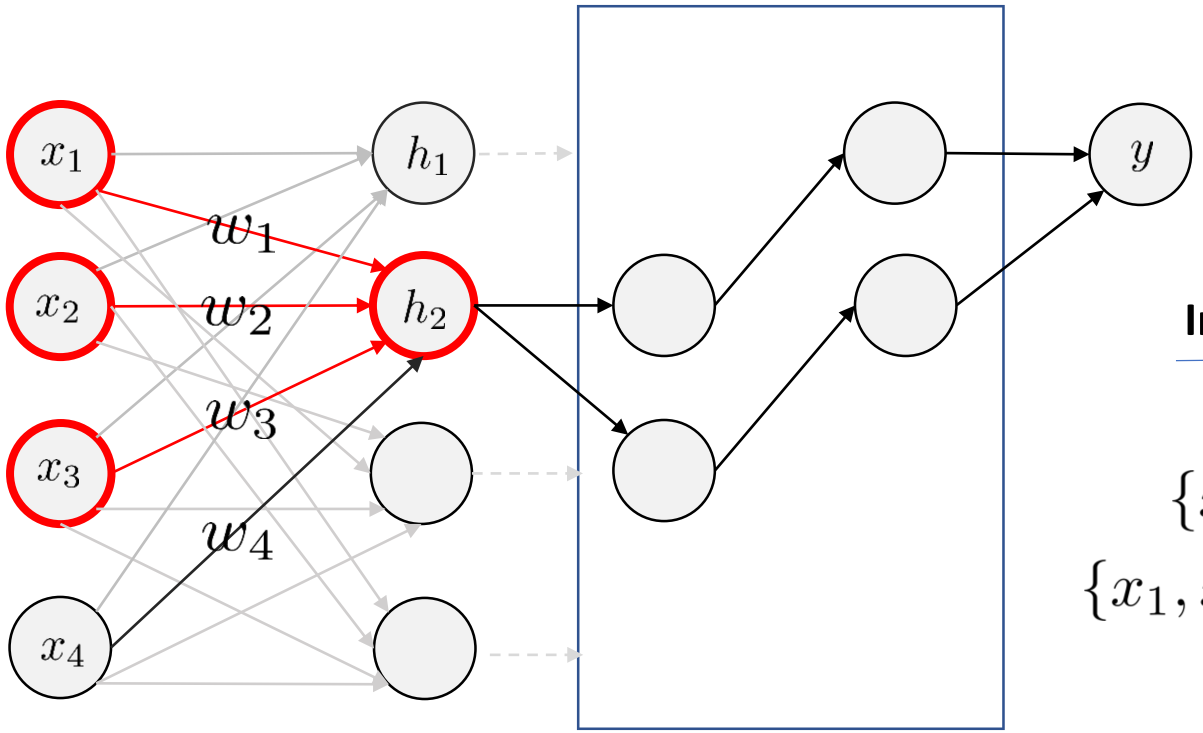
NID example



Interactions	Strength
$\{x_1, x_2\}$	$\lvert w_2\rvert z_1$
$\{x_1, x_2, x_3\}$	$\lvert w_3\rvert z_1$
$\{x_1, x_2, x_3, x_4\}$	$\lvert w_4\rvert z_1$
$\{x_1, x_3\}$	$\min(\lvert w_1\rvert, \lvert w_3\rvert) z_2 = \lvert w_1\rvert z_2$

$$|w_3| > |w_1| > |w_2| > |w_4|$$

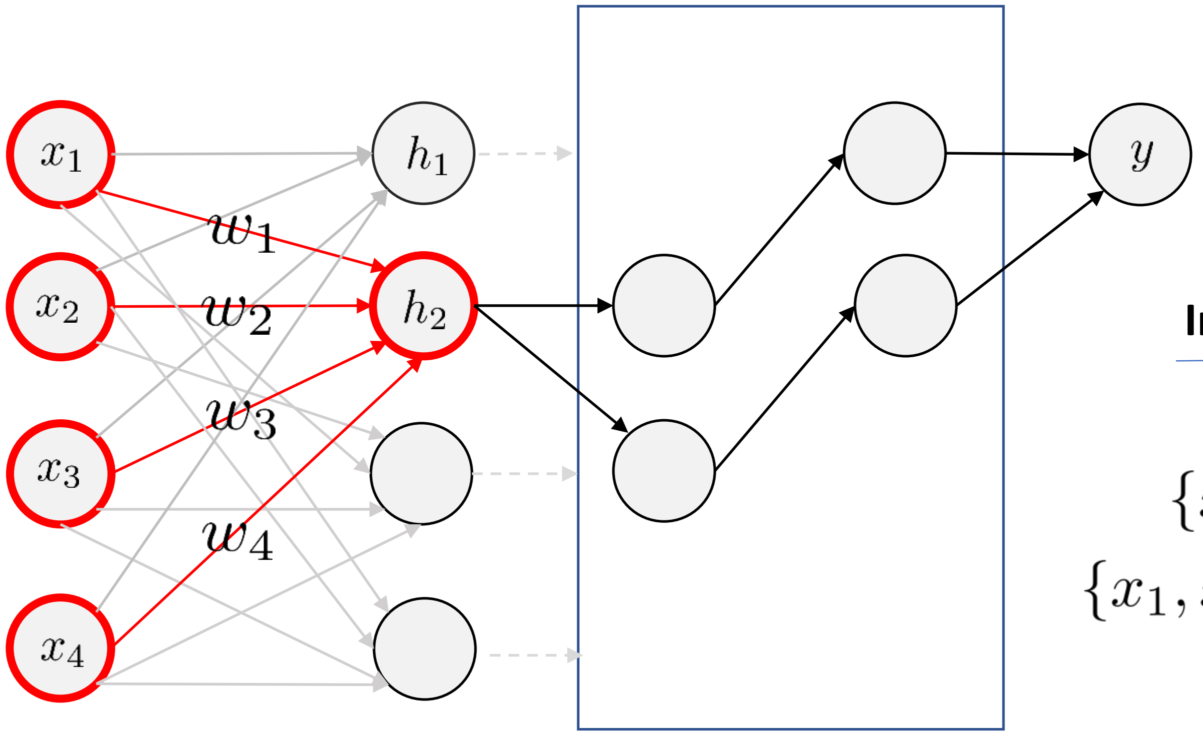
NID example



Interactions	Strength
$\{x_1, x_2\}$	$\lvert w_2\rvert z_1$
$\{x_1, x_2, x_3\}$	$\lvert w_3\rvert z_1 + \min(\lvert w_1\rvert, \lvert w_2\rvert, \lvert w_3\rvert) z_2$
$\{x_1, x_2, x_3, x_4\}$	$\lvert w_4\rvert z_1$
$\{x_1, x_3\}$	$\lvert w_1\rvert z_2$

$$|w_3| > |w_1| > |w_2| > |w_4|$$

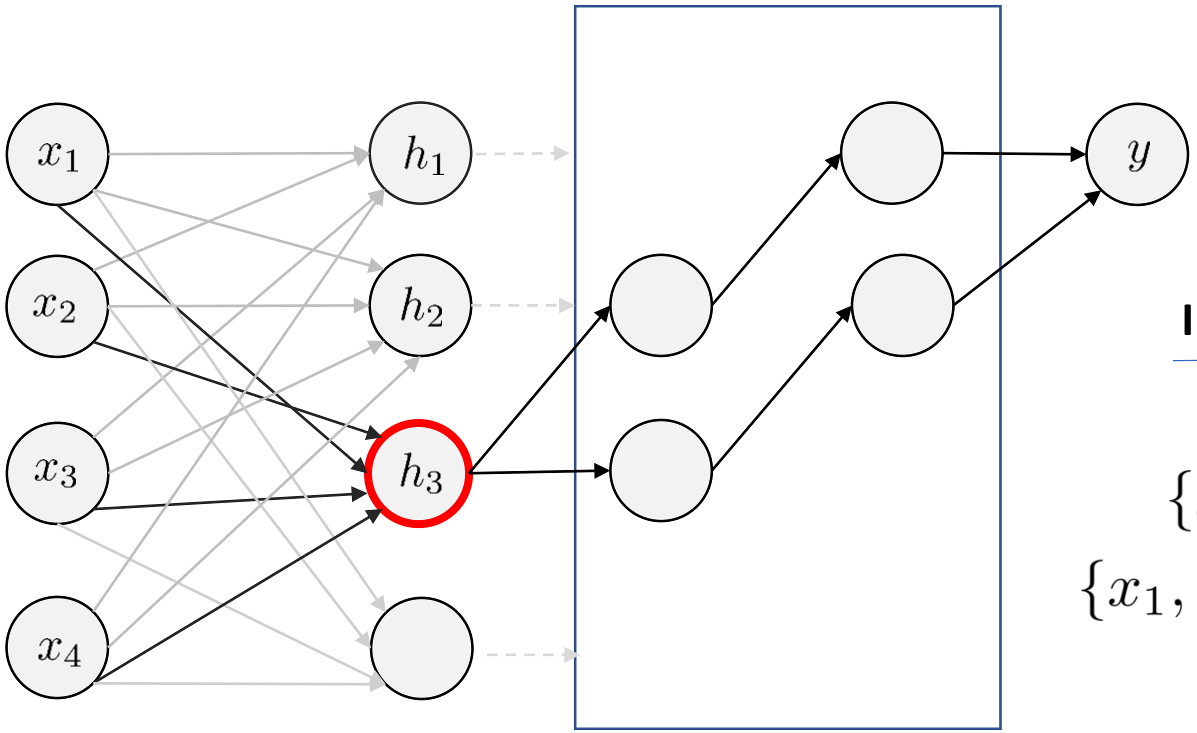
NID example



Interactions	Strength
$\{x_1, x_2\}$	$\lvert w_2\rvert z_1$
$\{x_1, x_2, x_3\}$	$\lvert w_3\rvert z_1 + \min(\lvert w_1\rvert, \lvert w_2\rvert, \lvert w_3\rvert) z_2$
$\{x_1, x_2, x_3, x_4\}$	$\lvert w_4\rvert z_1 + \min(\lvert w_1\rvert, \lvert w_2\rvert, \lvert w_3\rvert, \lvert w_4\rvert) z_2$
$\{x_1, x_3\}$	$\lvert w_1\rvert z_2$

$$|w_3| > |w_1| > |w_2| > |w_4|$$

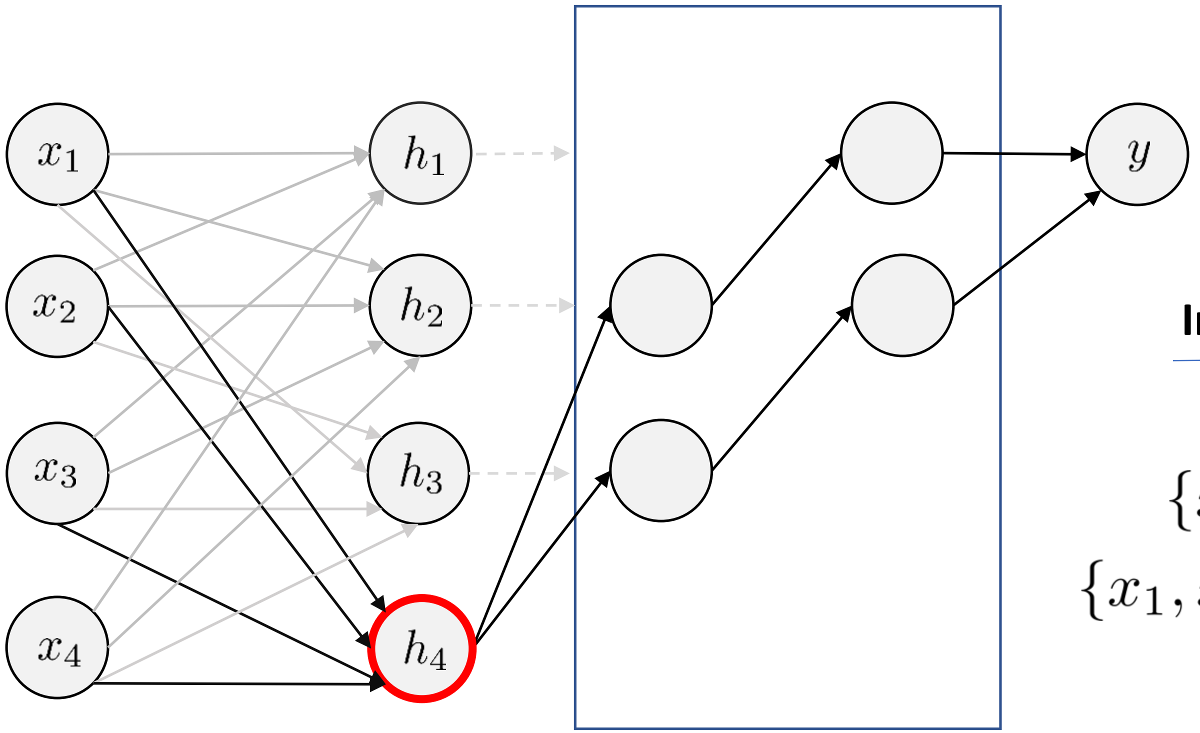
NID example



Interactions	Strength
$\{x_1, x_2\}$	$\lvert w_2\rvert z_1$
$\{x_1, x_2, x_3\}$	$\lvert w_3\rvert z_1 + \lvert w_2\rvert z_2$
$\{x_1, x_2, x_3, x_4\}$	$\lvert w_4\rvert z_1 + \lvert w_4\rvert z_2$
$\{x_1, x_3\}$	$\lvert w_1\rvert z_2$

$\ldots$

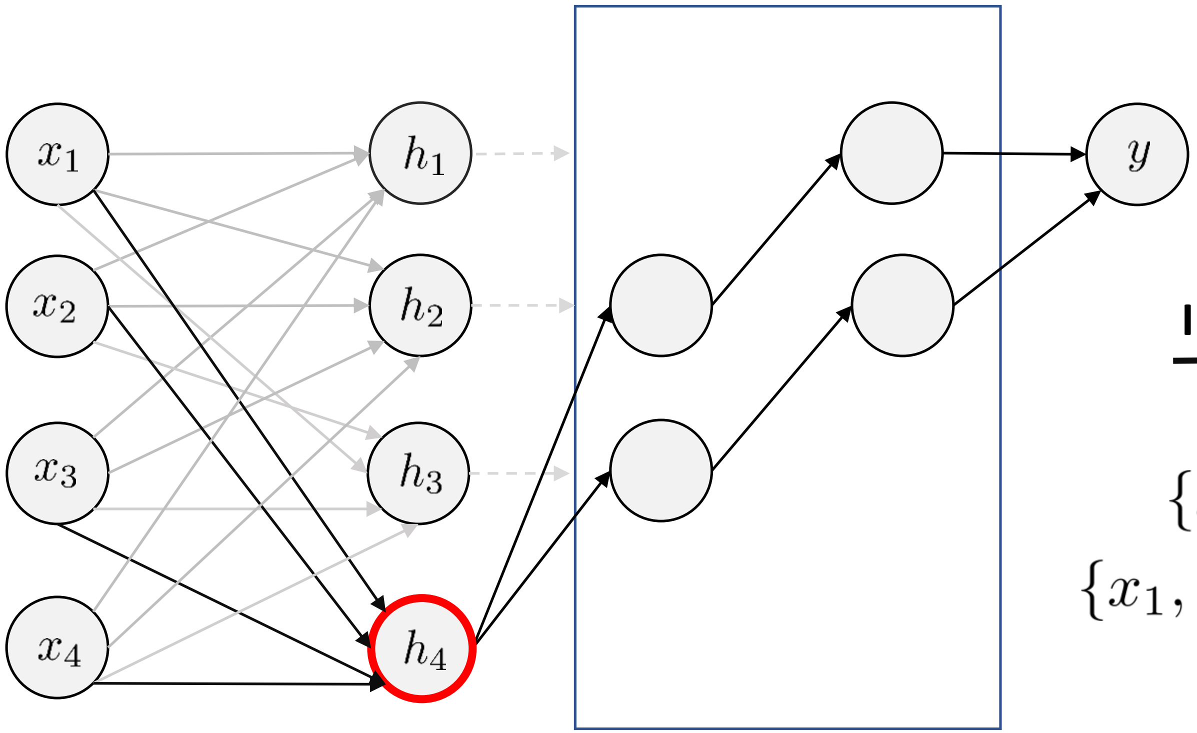
NID example



Interactions	Strength
$\{x_1, x_2\}$	$\|w_2\|z_1$
$\{x_1, x_2, x_3\}$	$\|w_3\|z_1 + \|w_2\|z_2$
$\{x_1, x_2, x_3, x_4\}$	$\|w_4\|z_1 + \|w_4\|z_2$
$\{x_1, x_3\}$	$\|w_1\|z_2$

$\vdots$

NID example



Interactions	Strength
$\{x_1, x_2\}$	1.3421
$\{x_1, x_2, x_3\}$	0.8241
$\{x_1, x_2, x_3, x_4\}$	0.3415
$\{x_1, x_3\}$	0.2310
⋮	⋮

top-K

NID: **Neural Interaction detection**

1. Train a Lasso-regularized MLP.
2. Interpret learned weights to obtain a ranking of interaction candidates.
3. Determine a cutoff for the top-K interactions.

Data often contains both

- ➢ statistical interactions.
- ➢ main effects: univariate influences of variables on an outcome variable.

- Model separately 2 simple networks: (**MLP**, **MLP-M**)
- Learn jointly with L1-regularization only on the interaction part to cancel out the main effect as much as possible

NID: **Neural Interaction detection**

1. Train a Lasso-regularized MLP.
2. Interpret learned weights to obtain a ranking of interaction candidates.
3. Determine a cutoff for the top-K interactions.

A **greedy algorithm** generates a ranking of interaction candidates

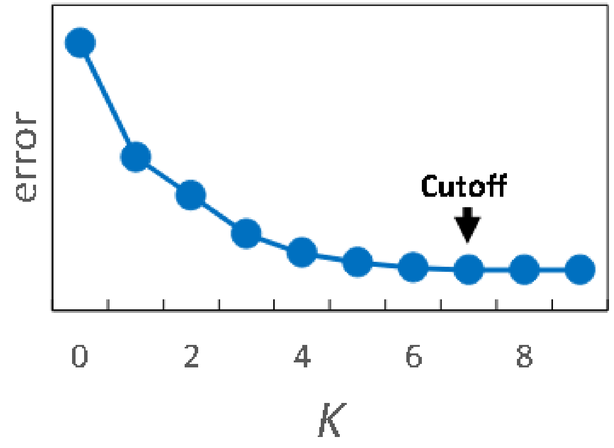
- at each hidden unit, it only considers the top-ranked interactions of every order based on their interaction strengths (set μ=min(.)).
  - drastically reduces the search space of potential interactions (O(hp) tests)
  - but still considers all orders.

NID: **Neural Interaction detection**

1. Train a Lasso-regularized MLP.
2. Interpret learned weights to obtain a ranking of interaction candidates.
3. Determine a cutoff for the top-K interactions.

$$c_K(x) = \sum_{i=1}^{p} g_i(x_i) + \sum_{i=1}^{K} g_i^{'}(x_I)$$

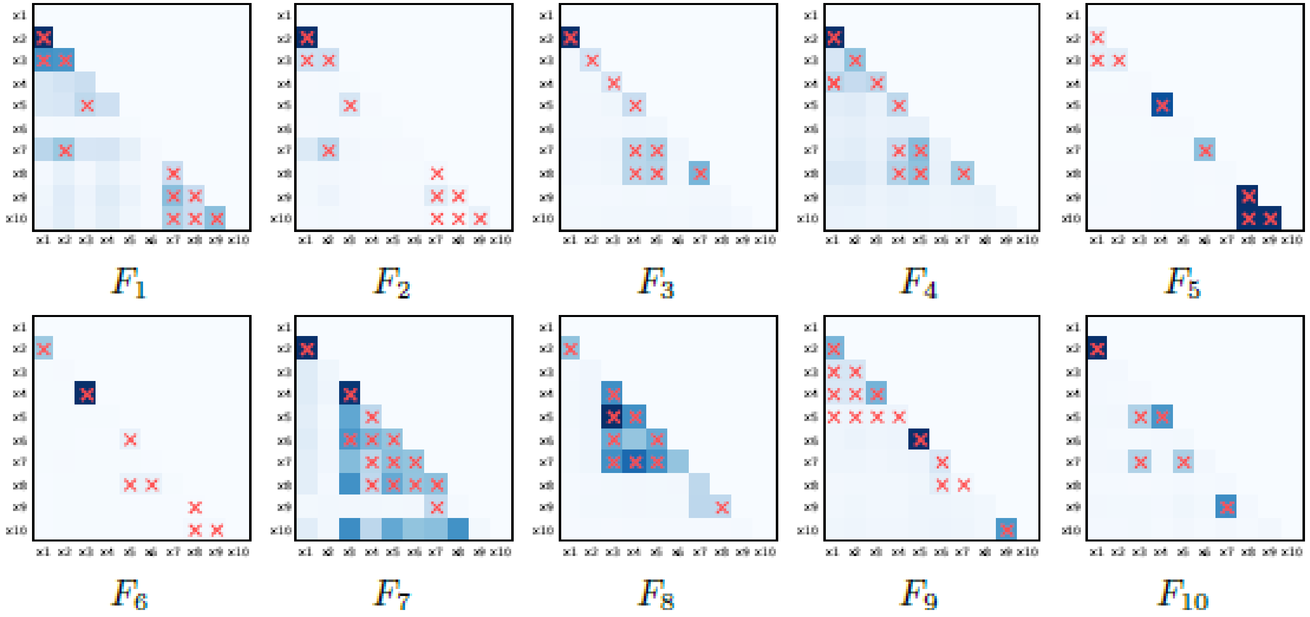
captures main effects **captures the interactions**



Gradually **add top-ranked interactions** to the GAM, increasing K, until GAM performance on a validation set plateaus.

Experiments

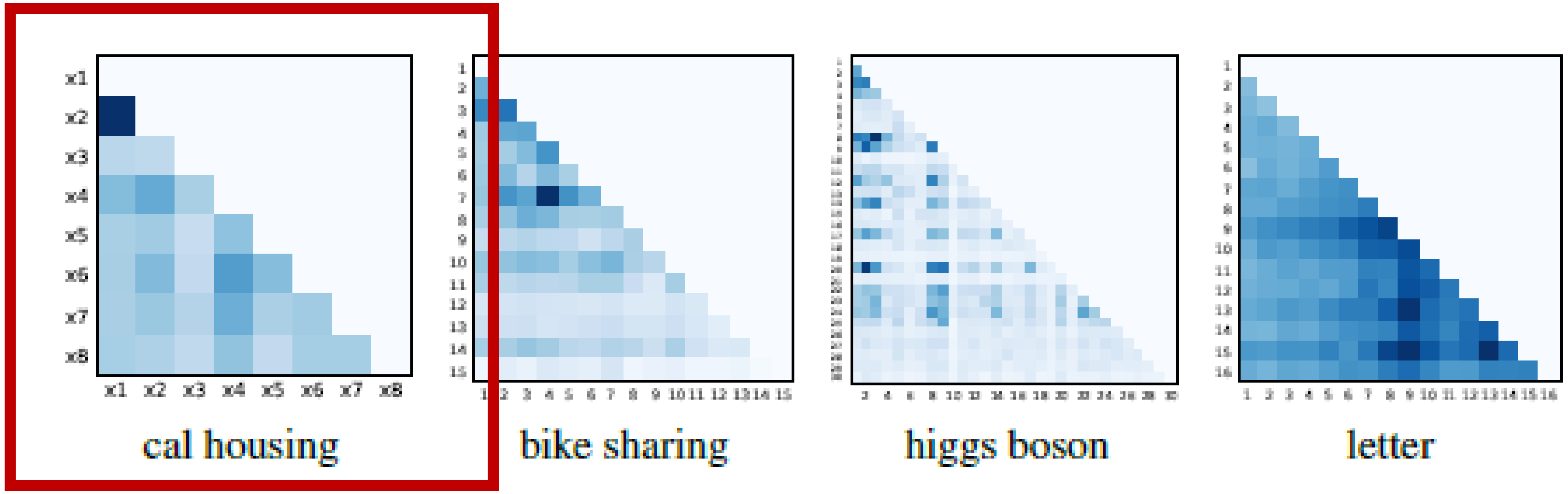
❖ **Tasks**:
- **Pairwise interaction detection -** Synthetic functions



The interaction strengths shown are normally high at the cross-marks!

Experiments

- Tasks:
  - **Pairwise interaction detection -** Real data

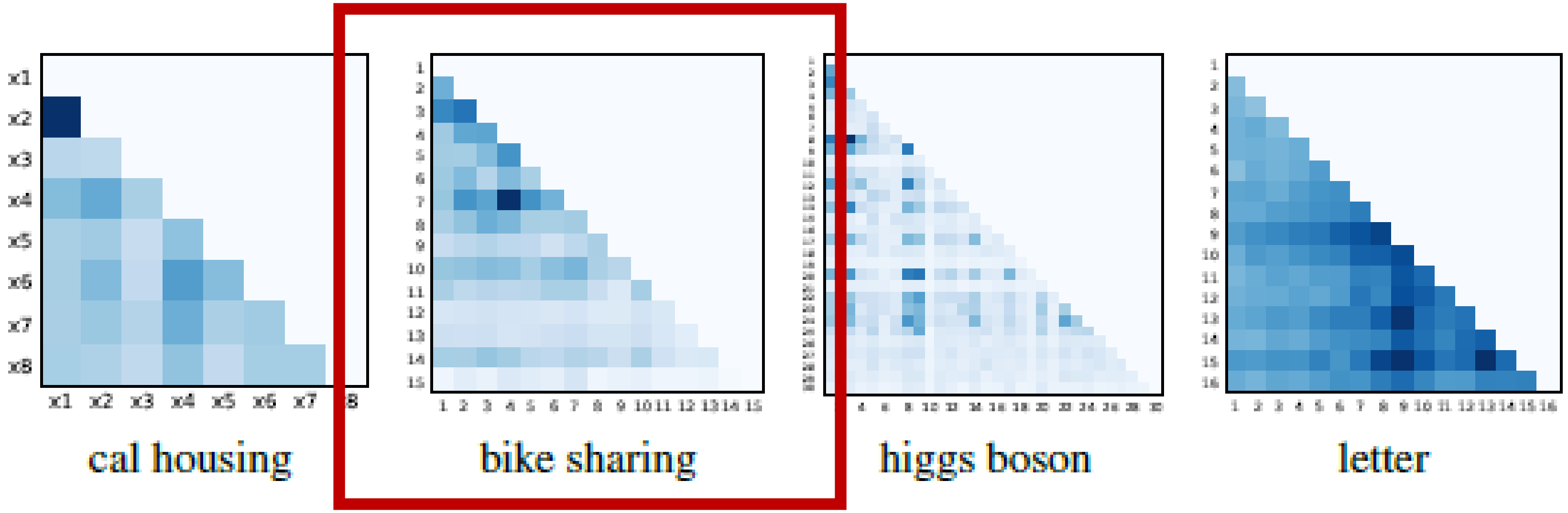


California Housing Prices
{1,2}: longitude and latitude!

Experiments

❖ Tasks:

- **Pairwise interaction detection -** Real data

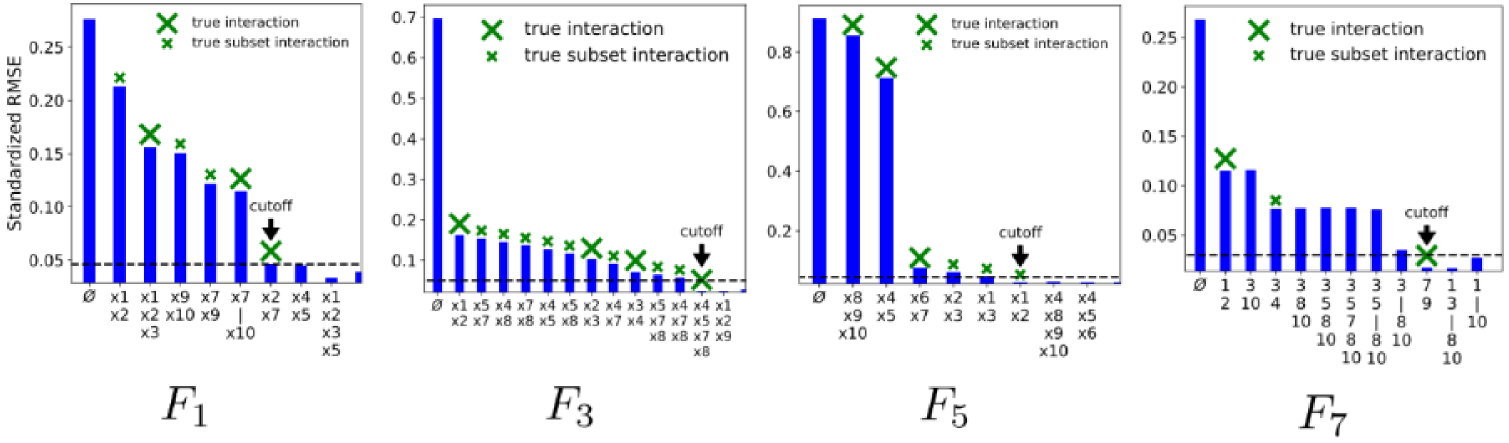


Number of Bike-share Users

{4,7}: hour and working day!

Experiments

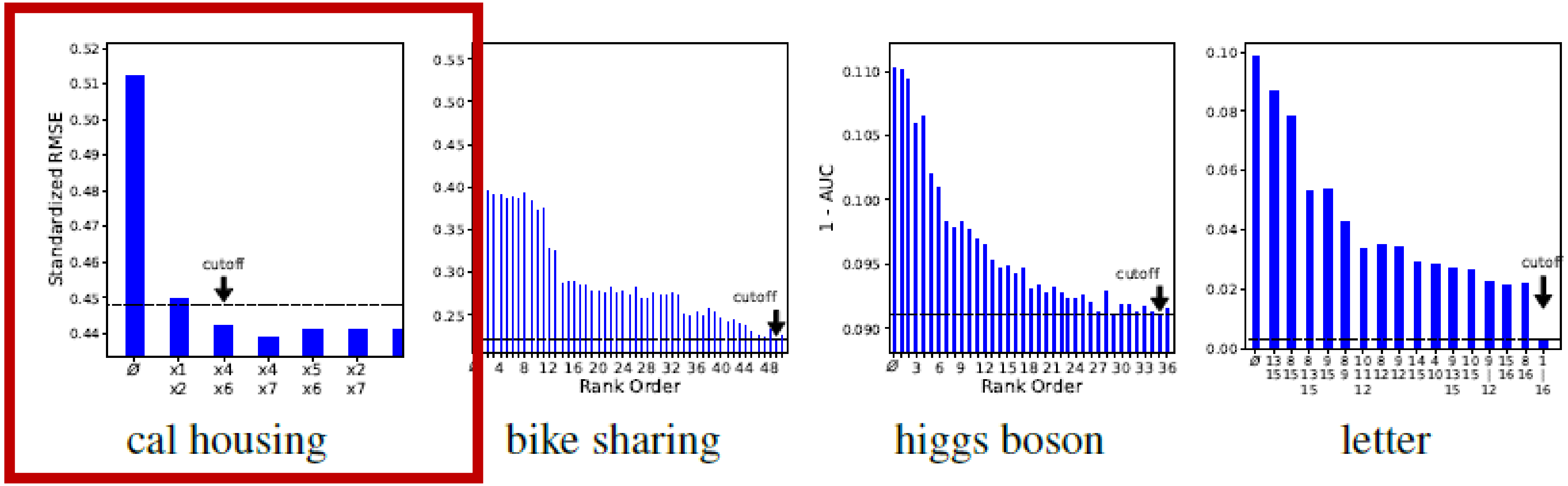
- **Higher order interaction detection -** Synthetic functions



$F_1(\mathbf{x})$	$\pi^{x_1 x_2}\sqrt{2x_3} - \sin^{-1}(x_4) + \log(x_3 + x_5) - \frac{x_9}{x_{10}}\sqrt{\frac{x_7}{x_8}} - x_2 x_7$
$F_3(\mathbf{x})$	$\exp\lvert x_1 - x_2\rvert + \lvert x_2 x_3\rvert - x_3^{2\lvert x_4\rvert} + \log(x_4^2 + x_5^2 + x_7^2 + x_8^2) + x_9 + \frac{1}{1 + x_{10}^2}$
$F_5(\mathbf{x})$	$\frac{1}{1 + x_1^2 + x_2^2 + x_3^2} + \sqrt{\exp(x_4 + x_5)} + \lvert x_6 + x_7\rvert + x_8 x_9 x_{10}$
$F_7(\mathbf{x})$	$(\arctan(x_1) + \arctan(x_2))^2 + \max(x_3 x_4 + x_6, 0) - \frac{1}{1 + (x_4 x_5 x_6 x_7 x_8)^2} + \left(\frac{\lvert x_7\rvert}{1 + \lvert x_9\rvert}\right)^5 + \sum_{i=1}^{10} x_i$

Experiments

❖ **Tasks**:
- **Higher order interaction detection -** Synthetic functions



Adding the first interaction significantly reduces RMSE.

Take - home points

- **Neural networks for a traditional statistical problem!**
- **Accurately detect general types of interactions**
- Without assuming any explicit interaction **order**
- Without searching an exponential solution space of interaction candidates.

Thank you!