λ-Guard: Structural & Stability Overfitting Index for Boosting

Fabrizio Di Sciorio, PhD[1]

[1]Department of Economics and Business, University of Almeria, Spain

16/02/2026

Overview

λ-Guard is a framework to detect overfitting **without using a test set**. Traditional overfitting measures rely on a held-out dataset to detect performance drops. λ-Guard instead analyzes:

- **Geometric structure** of the learned representation (how the model partitions the input space)
- **Stability** of predictions under small input perturbations

The model is decomposed into two key conceptual spaces:

1. **Representation Space (Capacity)** – measures how "rich" or complex the model representation is.
2. **Prediction Trajectory Space (Alignment)** – measures how effectively the model's components (trees) contribute to predicting the target.

Each tree in Gradient Boosting partitions the input space into leaf regions. We define a binary matrix Z where each row corresponds to an observation and each column to a leaf region across all trees:

$$Z_{i,j} = \begin{cases} 1 & \text{if observation } i \text{ falls into leaf } j \\ 0 & \text{otherwise} \end{cases}$$

This matrix is analogous to the **hat matrix** H in linear regression: it encodes how the model projects training data into its learned representation.

Mathematical Formulation

1. Leaf Membership Matrix Z

Given a dataset $X \in \mathbb{R}^{n \times d}$ and T trees, each tree t has L_t leaves. Define the total number of leaf regions as:

$$L = \sum_{t=1}^{T} L_t$$

Then $Z \in \mathbb{R}^{n \times L}$ is defined as above. Each row i represents the embedding of observation x_i into leaf space, while each column j represents a specific leaf region. Effectively, Z encodes the **geometric projection of the training data** into the model's functional representation.

2. Capacity C

Capacity quantifies the intrinsic dimensionality of the learned representation:

$$\bar{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i, \quad C = \frac{1}{n} \sum_{i=1}^{n} \|Z_i - \bar{Z}\|_2^2 = \mathrm{Var}(Z)$$

Intuition:

- High $C \rightarrow$ observations spread in many independent directions in leaf space $\rightarrow$ complex partitioning $\rightarrow$ more degrees of freedom $\rightarrow$ higher overfitting risk.
- Low $C \rightarrow$ most observations lie in few effective leaf combinations $\rightarrow$ simpler model.

Equivalently, in functional terms:

$$C = \mathrm{Var}(f(X)) = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - \bar{f})^2$$

3. Alignment A

Alignment measures how well the learned representation predicts the target $y \in \mathbb{R}^n$:

$$A = \mathrm{Corr}(f(X), y) = \frac{\mathrm{Cov}(f(X), y)}{\sigma_{f(X)} \sigma_y}$$

Intuition:

- High $A \rightarrow$ each tree contributes independent information toward predicting the target $\rightarrow$ efficient representation.
- Low $A \rightarrow$ later trees largely redundant $\rightarrow$ model may have wasted capacity.

4. Generalization Index GI

$$GI = \frac{A}{C}, \quad G_{\text{norm}} = \frac{A}{A + C} \in [0, 1]$$

Interpretation:

- $G_{\text{norm}} \to 1 \to$ strong generalization, alignment dominates
- $G_{\text{norm}} \to 0 \to$ high capacity with low alignment $\to$ risk of overfitting

5. Instability Index S

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{|f(x_i) - f(x_i + \epsilon_i)|}{\sigma_f}, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

Interpretation:

- High $S \to$ model is unstable; small changes in input produce large prediction differences $\to$ overfitting risk
- Low $S \to$ model robust

6. Overfitting Index λ

$$\lambda = \frac{C}{A + C} \cdot S, \quad \lambda_{\text{norm}} = \frac{\lambda - \min(\lambda)}{\max(\lambda) - \min(\lambda)} \in [0, 1]$$

Interpretation:

- High $\lambda \to$ many independent leaf regions that do not contribute to prediction + unstable predictions $\to$ strong overfitting signal
- Computable entirely on **training data**, no test set required

Geometric Interpretation

1. Z maps each observation into a high-dimensional leaf space
2. Capacity C measures the "spread" of points in this space
3. Alignment A captures how well this spread correlates with the target
4. Instability S detects whether the representation is sensitive to small input perturbations
5. λ combines both aspects into an overfitting score
6. Essentially, λ-Guard generalizes the hat matrix H concept to Gradient Boosting

geometric_inter.png

Figure 1: Geometric interpretation of λ-Guard. Gray squares: leaf regions, blue points: original observations, red points: instability, green arrows: alignment. High λ occurs when capacity is high, alignment low, and instability high.

References / Inspirations

- Hat matrix H in linear regression
- Gradient Boosting as a functional additive model
- Generalization Index (GI) framework
- λ in H Boosting matrix (pseudo residuals)