Alternating optimization of bivariate decision trees

Rasul Kairgeldin and **Miguel Á. Carreira-Perpiñán**
EECS, UC Merced

BayLearn
Bay Area Machine Learning Symposium

1 Introduction

Univariate decision trees, while being interpretable, they often lack competitive predictive accuracy due to their inability to model feature correlations, use a limited number of input features, and rely on heuristic algorithms for training. In contrast, multivariate (oblique) trees can use multiple features in each node, capturing high-dimensional correlations better. However, they can be difficult to interpret. Bivariate decision trees offer a practical compromise by using pairs of features in each node, striking a balance between interpretability and accuracy. By adapting the Tree Alternating Optimization (TAO) algorithm, bivariate trees can be trained more effectively, resulting in smaller and more accurate trees. The TAO algorithm updates node parameters iteratively, optimizing a well-defined objective function over the entire tree. While slower than traditional algorithms, it scales well to large datasets. Our experiments demonstrate that bivariate trees outperform univariate trees in terms of interpretability and accuracy. We believe bivariate trees offer a practical and scalable solution for data analysis tasks.

Work partially supported by NSF award IIS–2007147.

2 Learning bivariate trees with TAO

We establish the following objective function over all parameters of a tree:

$$\min_{\boldsymbol{\Theta}} E(\boldsymbol{\Theta}) = \sum_{n=1}^{N} L(y_n, T(\mathbf{x}_n; \boldsymbol{\Theta})) + \lambda \sum_{i \in \mathcal{N}_{\text{dec}}} \phi(\mathbf{w}_i), \text{s.t. } \|\mathbf{w}_i\|_0 \leq 2,\ b_i \in \mathbb{R},\ i \in \mathcal{N}_{\text{dec}}; \quad c_j \in \{1, \dots, K\},\ j \in \mathcal{N}_{\text{leaf}} \tag{1}$$

where $L(\cdot,\cdot)$ is $0/1$ loss function. Furthermore, we introduce the following regularization:

$$\phi(\mathbf{w}_i) = \begin{cases} C, & \text{if } \|\mathbf{w}_i\|_0 = 2 \\ \|\mathbf{w}_i\|_0, & \text{if } \|\mathbf{w}_i\|_0 < 2 \end{cases}$$

Separability condition implies that equation 1 can be separated and optimized over parameters of any non-descendant nodes (located on the same depth) independently and in parallel. **Reduced problem over a node** (RP) states that optimizing equation 1 over parameters of the given node $i \in \mathcal{N}$ reduces to simpler, well-defined problem involving its reduced set $\mathcal{R}_i$.

For leaf $i \in \mathcal{N}_{\text{leaf}}$ the exact solution of RP is a majority class of samples in $\mathcal{R}_i$.

For decision node $i \in \mathcal{N}_{\text{dec}}$ RP is $0/1$ *loss binary classification problem*:

$$E_i(\mathbf{w}_i, b_i) = \sum_{n \in \mathcal{R}_i} L(\bar{y}_n, f_i(\mathbf{x}_n; \mathbf{w}_i, b_i)) + \lambda\phi(\mathbf{w}_i),\ \text{s.t. } \|\mathbf{w}_i\|_0 \leq 2,\ b_i \in \mathbb{R} \tag{2}$$

where L is a $0/1$ loss and $\bar{y}_n \in \{\texttt{left}, \texttt{right}\}$ corresponds to a pseudolabel assigned to a training instance x_n, signifying the child that yields a lower loss value. The loss is computed by propagating a sample through the corresponding child.

```
input training set $\{\mathbf{x}_n, \bar{y}_n\}_{n \in \mathcal{R}_i}$ of
decision node $i \in \mathcal{N}_{\text{dec}}$,
  matrix of orientations $\mathbf{W} \in \mathbb{R}^{2 \times H}$
for each pair of features $j, k \in D$
  for $\mathbf{w}_l \in \mathbf{W}$
    $\mathbf{x}_l^{j,k} \leftarrow$ project selected features onto $w_l$
    $b_l^{j,k} \leftarrow$ optimal thresholding over $\mathbf{x}_l^{j,k}$
    if $j, k, \mathbf{w}_l, b_l^{j,k}$ produce lowest value of eq. 2
      $\theta_i^{biv} \leftarrow \{\mathbf{w}^*, b_l^{j,k}\}$, where $\mathbf{w}^*$ is a sparse vector
      of all zeros with corresponding value of $\mathbf{w}_l$
      at $j, k$
    end if
  end for
end for
return $\theta_i^{biv}$
```

Figure: Pseudocode of bivariate solution.

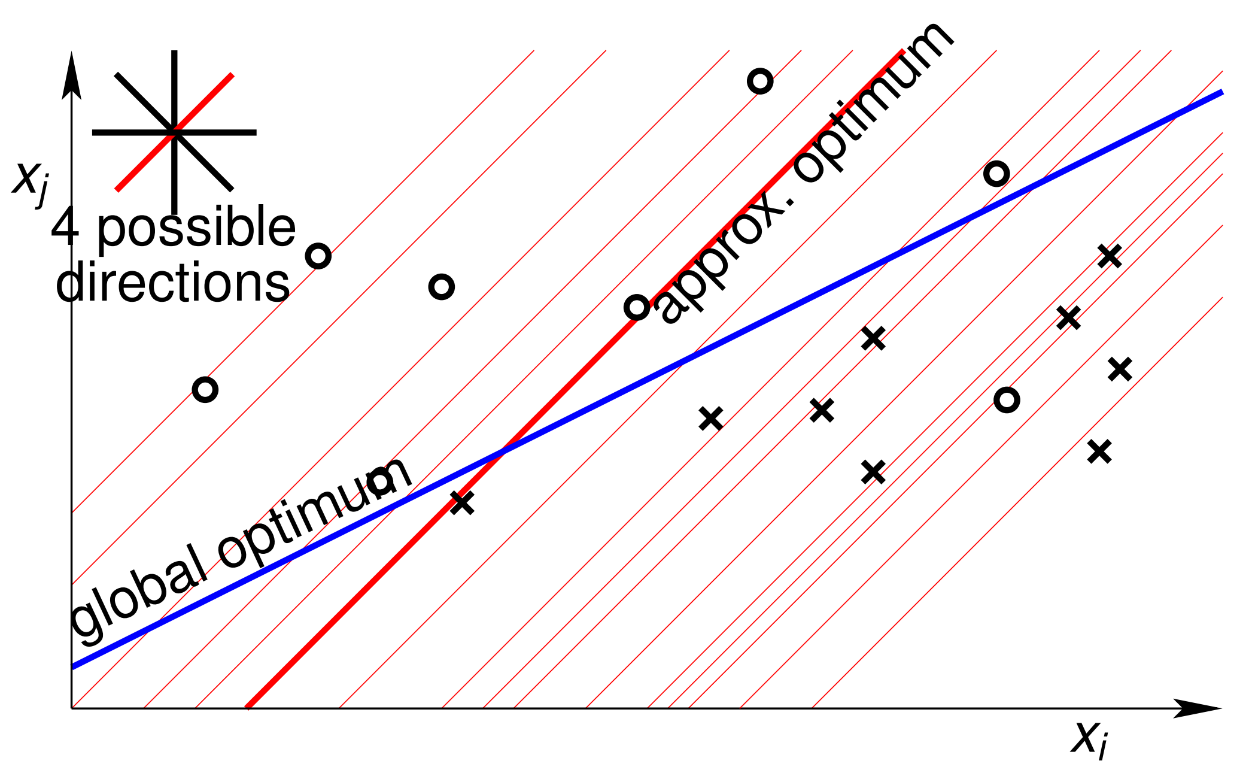


Figure: Illustration of our approximate solution of the RP at a decision node. The instances in the reduced set of the node are labeled according to their pseudolabels (preferred child, left ○ or right +). The thin red lines are all the possible thresholds (passing through midpoints between projected instances) for the red orientation.

3 Effect of regularization and Interpretability

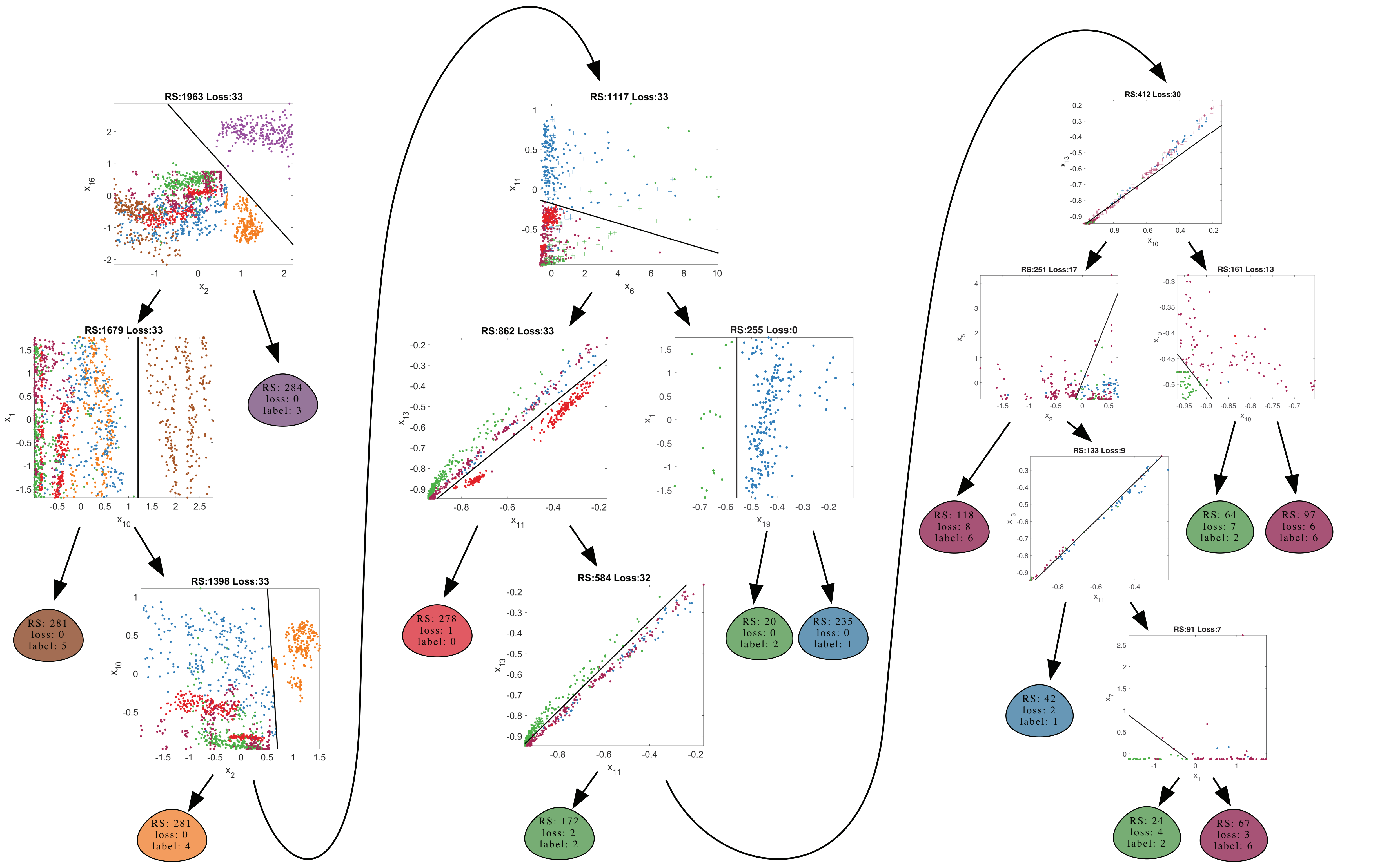


Figure: Figure 1 shows resulting bivariate tree trained on Segment dataset. We show 0/1 loss of each node on its reduced set along with number of samples in it. In decision nodes we visualize the best univariate or bivariate split.

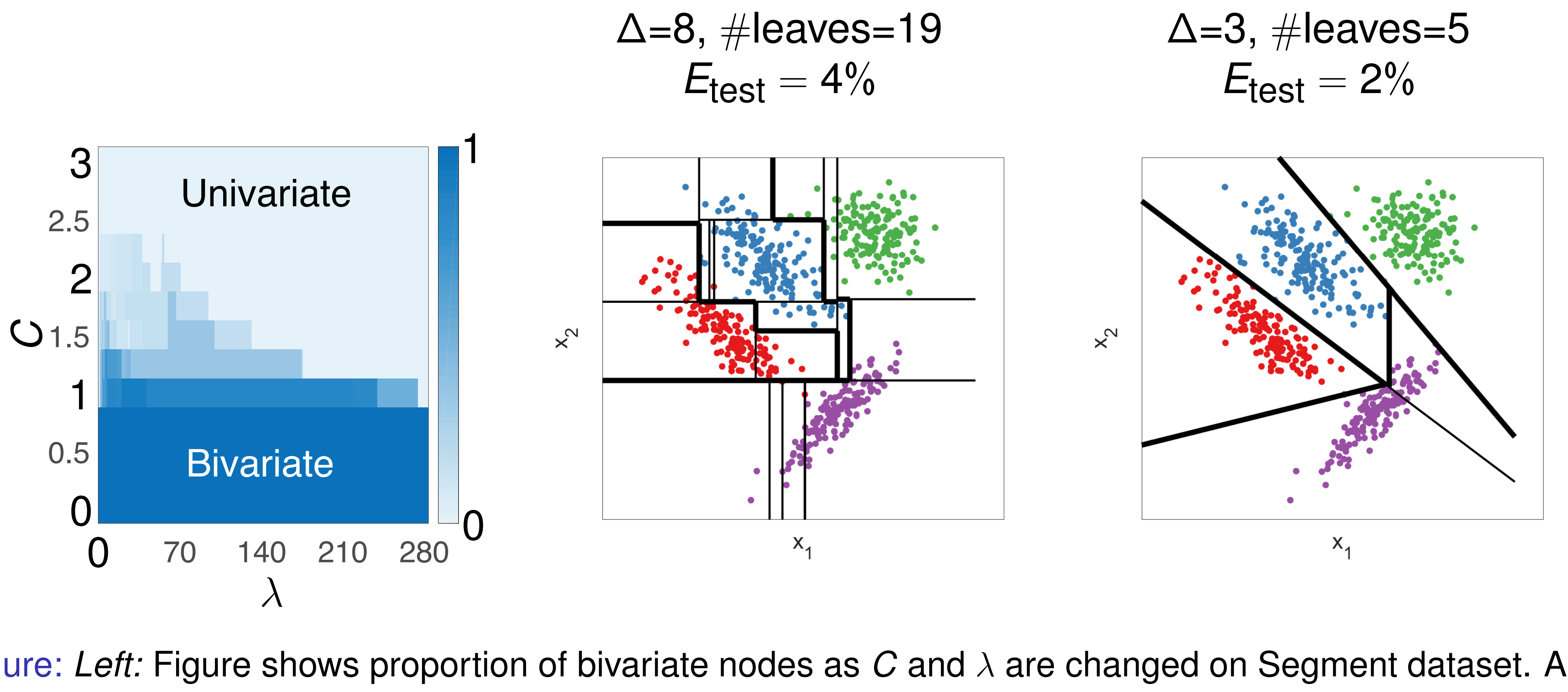
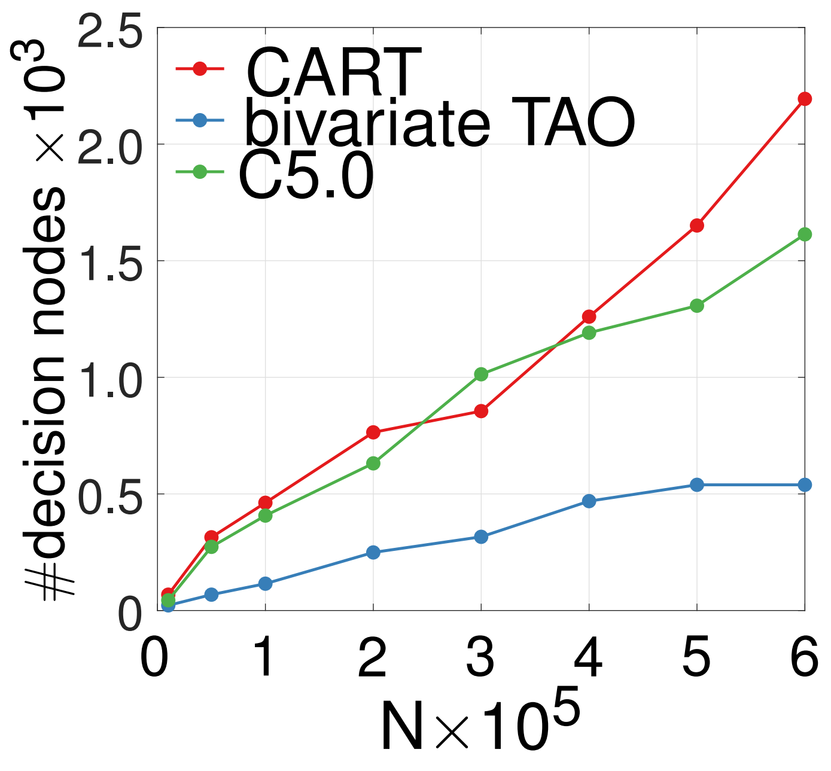
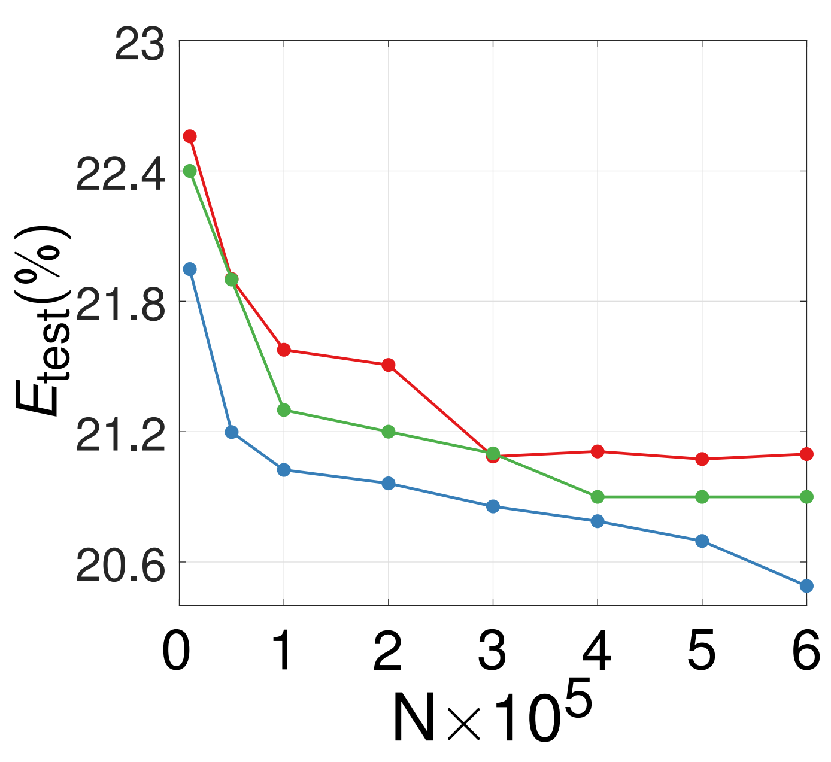


Figure: *Left:* Figure shows proportion of bivariate nodes as C and λ are changed on Segment dataset. As λ increases beyond 280 tree collapses into a root. We indicate regions where tree is fully bivariate or fully univariate. *Right:* Partitioning of the space provided by univariate and bivariate trees.

4 The size of the pruned tree


CART
bivariate TAO
C5.0
#decision nodes $\times 10^3$
N×10^5


$E_{\text{test}}(\%)$
N×10^5