

STABLE UPDATE OF REGRESSION TREES

Morten Blørstad

Department of Informatics
University of Bergen
Norway
morten.blørstad@uib.no

Berent Å. S. Lunde

Equinor
Norway
berl@equinor.com

Nello Blaser

Department of Informatics
University of Bergen
Norway
nello.blaser@uib.no

ABSTRACT

Updating machine learning models with new information usually improves their predictive performance, yet, in many applications, it is also desirable to avoid changing the model predictions too much. This property is called stability. In most cases when stability matters, so does explainability. We therefore focus on the stability of an inherently explainable machine learning method, namely regression trees. We aim to use the notion of *empirical stability* and design algorithms for updating regression trees that provide a way to balance between predictability and empirical stability. To achieve this, we propose a regularization method, where data points are weighted based on the uncertainty in the initial model. The balance between predictability and empirical stability can be adjusted through hyperparameters. This regularization method is evaluated in terms of loss and stability and assessed on a broad range of data characteristics. The results show that the proposed update method improves stability while achieving similar or better predictive performance. This shows that it is possible to achieve both predictive and stable results when updating regression trees.¹

1 INTRODUCTION

Training an initial machine learning (ML) model is just the start of the ML life cycle. The model can be improved as more data becomes available, leading to an iterative process of continuous improvement. Thus, in industry settings, ML models are frequently updated to incorporate new information. The main objective of such a process has traditionally been to improve the model’s predictive capability. However, many industries, for instance in finance or medicine, have additional requirements such as interpretability and stability. Interpretability is important due to “algorithmic accountability” (Kaminski & Malgieri (2021)), whereas stability is crucial to ensure consistent and reliable predictions on new unseen data. In this manuscript, we focus on the stability aspect.

Achieving stability involves minimizing model variance as stability is connected to generalization and inversely related to variance (Bousquet & Elisseeff (2002)). There have been numerous studies on the stability of learning algorithms (see, e.g., Devroye & Wagner (1979); Kearns & Ron (1999); Bousquet & Elisseeff (2002)) focusing on theoretical bounds of their generalization error. Some previous studies have also presented empirical stability measures for both classification and regression tasks, which can be divided into structural and semantic stability (Turney (1995); Briand et al. (2009); Lin et al. (2002); Lim & Yu (2013); Philipp et al. (2017)). Related to structural stability, there is the notion of predictor stability, which uses invariant risk minimization to find time-invariant predictors (Arjovsky et al., 2020). We focus on semantic stability, (dis-)similarity of the predictions of two models, which we review in Section 2. While *measuring* stability has been studied thoroughly, there are also some studies focusing on improving it for classifiers. The studies by Milani Fard et al. (2016), Goh et al. (2016), Cotter et al. (2018) and Liu et al. (2022) address the issue of unnecessary prediction change when training consecutive classifiers. Milani Fard et al. (2016) propose a Markov Chain Monte Carlo stabilization method, whereas Goh et al. (2016), Cotter et al. (2018) introduces different optimization constraints when training classifiers. Liu et al. (2022) improve model stability of deep-learning classifiers in the context of NLP systems. In this article, we aim to improve the empirical stability of regression trees.

Regression trees are appealing due to their inherently interpretable structure, and their ability to discover interaction effects and select features. However, due to regression trees’ high variance, small changes in data can often lead to very different splits, which can cause inconsistent predictions, i.e., instability (Breiman (1996), Hastie et al. (2001)). This issue of instability in tree models has been addressed in previous research for decision trees. For instance, Last et al. (2002) showed using Info-Fuzzy Networks that semantically stable decision trees could be learned for classification tasks while maintaining a reasonable level of accuracy. Additionally, Dannerger (2000) proposed a bootstrap-based

¹A preliminary version of results in this paper appeared in the Master thesis of the first author (Blørstad, 2023).

node-level stabilizing approach for decision trees. This method was later improved by Briand et al. (2009), demonstrating improved structural stability compared to the classical decision tree algorithm (CART). Furthermore, Moneda & Mauá (2022) propose a method using invariant risk minimization to learn time-robust trees. To the best of our knowledge, no work focuses on improving the empirical semantic stability of regression trees in the context of continuous updates.

Contributions Our contributions to stabilizing regression trees are as follows.

1. This is the first work on *improving* the empirical semantic stability of regression trees.
2. We incorporate stability measures as a regularization term in the loss function, providing a way of ensuring a level of stability by tuning regularization strength. Our novel approach involves dynamically weighting the regularization for each data point based on the model’s uncertainty.
3. We evaluate our approach on several different datasets, showing that this technique in combination with a base stability regularization achieves a better balance of the predictability-stability trade-off.
4. Our python-package for stable regression tree updates is available at <https://github.com/MortenBlorstad/StableTreeUpdates>.

The remainder of the paper unfolds as follows: Section 2 introduces empirical stability and Section 3 introduces regression trees. Section 4 describes the proposed method for stable updates of regression trees. Section 5 provides the experimental setup and results on various datasets. Finally, in Section 6 we discuss the results and conclude.

2 EMPIRICAL STABILITY

Consider the regression task of learning a function (or model) f that maps the input vector $\mathbf{x} \in \mathbb{R}^m$ to the corresponding response variable $y \in \mathbb{R}$. Learning such a function involves approximating the optimal function f^* from the set of the candidate function \mathcal{F} by minimising the empirical loss given a *loss function* $\mathcal{L}(\cdot, \cdot)$, over a training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$

$$f^* \approx \hat{f} = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n [\mathcal{L}(y_i, f(\mathbf{x}_i))].$$

In this context, stability is defined as how much a model’s predictions change due to changes in the training data, such as the removal or addition of a data point. It reflects a data point’s impact on the learning process. Most studies focus on theoretical stability guarantees, but for many ML algorithms, no such theoretical guarantees exist. We instead take an *empirical approach* and consider empirical stability measures (Turney (1995), Last et al. (2002), Milani Fard et al. (2016), Philipp et al. (2017)) that can be assessed using test data, $\mathcal{D}_{\text{test}}$, which has not been used in the learning process. Similar to how we minimize empirical loss to improve predictability, we introduce an empirical instability measure that we minimize to improve stability.

Consider a trained model f_0 at time $t = 0$ with the training data \mathcal{D}_0 . At $t = 1$ more data, $\mathcal{D}_{\Delta t}$ becomes available. The model is updated using all available data $\mathcal{D}_1 = \mathcal{D}_0 \cup \mathcal{D}_{\Delta t}$ to obtain the updated model f_1 . The instability of the update is given by the *instability function* $\mathcal{S}(\cdot, \cdot)$,

$$\text{empirical instability} = \frac{1}{n} \sum_{i=1}^n [\mathcal{S}(f_0(\mathbf{x}_i), f_1(\mathbf{x}_i))],$$

where $\mathcal{S}(\cdot, \cdot)$ measures the discrepancy between the predictions of f_0 and f_1 over a test dataset. In other words, we are considering the mean pointwise empirical instability. The choice of $\mathcal{S}(\cdot, \cdot)$, similar to the choice of $\mathcal{L}(\cdot, \cdot)$, depends on the use case and domain knowledge. The squared error $\mathcal{S}(a, b) = (a - b)^2$ and absolute error $\mathcal{S}(a, b) = |a - b|$ are examples of empirical instability measures that can be written as a mean pointwise empirical instability. Another example is the negative coverage probability $\mathcal{S}(a, b, k) = -\mathbb{1}_{|a-b| \leq k}$ (Lin et al. (2002)).

3 LEARNING REGRESSION TREES

There are many methods of fitting Regression trees, most notably the Classification And Regression Tree (CART) algorithm by Breiman et al. (1984) (see Figure 1). The CART algorithm uses a top-down greedy approach to partition the feature space $R = \mathbb{R}^m$ into D non-overlapping regions, $R_1, \dots, R_d, \dots, R_D$ and fit a leaf prediction, \hat{w}_d to each

region based on the response observations within the region. Starting from a root node R_{root} containing all the data, the algorithm seeks the splitting variable $j \in 1, \dots, m$ and split value s that creates the half-planes $R_l(j, s) = \{x \in R_{root} | x_j \leq s\}$ and $R_r(j, s) = \{x \in R_{root} | x_j > s\}$ that maximize the loss reduction \mathcal{R} ,

$$\mathcal{R}(j, s) = \sum_{i: x_i \in R_{root}} \mathcal{L}(y_i, \hat{w}_{root}) - \sum_{i: x_i \in R_l(j, s)} \mathcal{L}(y_i, \hat{w}_l) - \sum_{i: x_i \in R_r(j, s)} \mathcal{L}(y_i, \hat{w}_r).$$

Here R_l and R_r are the datasets of left and right nodes after the split and $R_{root} = R_l \cup R_r$ is the dataset before the split. The terms \hat{w}_{root} , \hat{w}_l and \hat{w}_r denote the predictions of the root, left and right node. After finding the partition with the largest loss reduction, the splitting process is repeated on the resulting regions until reaching a stopping criterion. Given a learned tree model f , the feature mapping $q(\mathbf{x})$ is a function that maps inputs to their corresponding region R and its prediction for input x is given as $f(x) = \hat{w}_{q(\mathbf{x})}$.

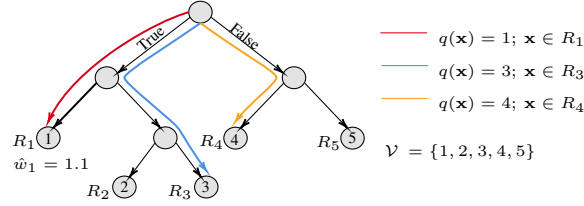


Figure 1: An example of a Classification and Regression Tree (CART) with five leaf nodes \mathcal{V} . The vector $\hat{\mathbf{w}} = (\hat{w}_1, \hat{w}_2, \hat{w}_3, \hat{w}_4, \hat{w}_5)$ is the possible predictions the tree can make.

4 STABLE REGRESSION TREE UPDATE

We propose an update method that incorporates a stability regularization term into the loss function. The key lies in the determination of the regularization strength, which we propose to be a composite of base regularization and uncertainty-weighted regularization, each adjusted through a hyperparameter. This section first describes the modification made to the CART algorithm followed by a description of our proposed update method.

We modify the CART by using the second-order approximation of the loss reduction,

$$\tilde{\mathcal{R}}(j, s) = \frac{1}{2} \left[\frac{(\sum_{i: x_i \in R_l} g_i)^2}{\sum_{i: x_i \in R_l} h_i} + \frac{(\sum_{i: x_i \in R_r} g_i)^2}{\sum_{i: x_i \in R_r} h_i} - \frac{(\sum_{i: x_i \in R_{root}} g_i)^2}{\sum_{i: x_i \in R_{root}} h_i} \right],$$

where $g_i = \partial_{\hat{y}^{k-1}} \mathcal{L}(y_i, 0)$ and $h_i = \partial_{\hat{y}^{k-1}}^2 \mathcal{L}(y_i, 0)$. This is equivalent to a gradient boosting iteration from a model $f(\mathbf{x})^{k-1} = \hat{y}^{k-1} = 0$ predicting zero and a learning rate of one (Chen & Guestrin (2016)). For squared error loss, this tree will be identical to a CART tree built using exact loss reduction \mathcal{R} .

The second-order approximation allows the use of the *adaptive tree complexity* method by Lunde et al. (2020) which adaptively stops the splitting process of a branch based on an information criterion, alleviating the need for hyperparameter tuning to optimize the tree complexity. Additionally, it makes it easy to incorporate additional objectives to the loss function as long as they are twice differentiable. Note that as the information criterion is rooted in the loss function, any additional objectives added to the loss are factored in when determining the tree complexity.

4.1 STABLE LOSS

In our approach to improve stability, we introduce a stable loss function that incorporates a regularization term. Similar to *explicit regularization* using a penalty term in the loss function to encourage simpler solutions and address the high-variance problem (i.e., overfitting), the *Stable Loss* (SL) adds a penalty term to encourage small change in the predictions (i.e., stability). SL uses the predictions of the current tree f_0 on the combined dataset $\mathcal{D}_1 = \{\mathbf{x}^{(1)}, y^{(1)}\}$ to regularize the loss function to prevent the predictions of the updated tree f_1 from deviating too much from the predictions of f_0 . In other word, when calculating loss reduction $\tilde{\mathcal{R}}$, it replaces the loss \mathcal{L} with the stability loss

$$\mathcal{L}_{SL}(y_i, f_1(\mathbf{x}_i); f_0(\mathbf{x}_i)) = \underbrace{\mathcal{L}(y_i, f_1(\mathbf{x}_i))}_{\text{loss}} + \gamma_i \underbrace{\mathcal{S}(f_0(\mathbf{x}_i), f_1(\mathbf{x}_i))}_{\text{instability}}, \quad (1)$$

where γ_i determines the strength of regularization (or belief in f_0). The choice of γ_i is key, and we explore two distinct approaches and a combination of these to determine its value.

4.1.1 CONSTANT REGULARIZATION

The simplest approach employs a uniform regularization factor α , such that $\gamma_i = \alpha$ for every data point. Here α is a hyperparameter that is tuned to balance the trade-off between predictability and stability.

4.1.2 UNCERTAINTY-WEIGHTED REGULARIZATION (UWR)

Intuitively, to preserve predictability, we want to be more stable for confident regions of f_0 compared to uncertain ones. The UWR dynamically adjusts γ_i based on each data point contribution to the model’s uncertainty. Abusing the notations $\hat{w}(\mathbf{x}) = \hat{w}_q(\mathbf{x})$, we denote $\hat{\sigma}_{\hat{w}_0(\mathbf{x})}^2$ as the variance estimate of the predictions $\hat{w}_0(\mathbf{x})$ under the model f_0 . We divide the precision $\frac{1}{\hat{\sigma}_{\hat{w}_0(\mathbf{x})}^2}$ by the number of data points $n(\mathbf{x})$ contributing to this variance in order to get the data point’s contribution to the precision. We then use the uncertainty factor

$$\phi(\mathbf{x}_i) = \frac{c}{n(\mathbf{x}_i)\hat{\sigma}_{\hat{w}_0(\mathbf{x}_i)}^2 + \epsilon},$$

for a scaling constant $c > 0$ and a small $\epsilon > 0$ to ensure numerical stability. We introduce β , a hyperparameter akin to α , and set $\gamma_i = \beta\phi(\mathbf{x}_i)$. This is reminiscent of relative weights according to a Gaussian prior on $\hat{w}_0(\mathbf{x}_i)$. The scaling constant c is chosen as the average response variance across leaf nodes, $c = n_0^{-1} \sum_{i=1} \hat{\sigma}_{y_i|\mathbf{x}_i}^2$. Assuming the variance is $O(n(\mathbf{x})^{-1})$, this places β and α on a comparable scale.

The greedy splitting in CART complicates estimating $\sigma_{\hat{w}_0(\mathbf{x}_i)}^2$. However, the relationship $C = \mathbb{E}[h_i]\sigma_{\hat{w}_d}^2$ between estimator variance and loss-based information criteria, C (see [Takeuchi \(1976\)](#) or [Burnham & Anderson \(2004\)](#) for an English version), aids in this regard. Specifically, [Lunde et al. \(2020, Appendix A\)](#) provides a conservative estimate for C correcting for greedy split points. We use this relationship to construct the estimator for the prediction variance in node d :

$$\hat{\sigma}_{\hat{w}_d}^2 = \frac{\sum_{i \in I_d} (g_i + h_i \hat{w}_d)^2}{(\sum_{i \in I_d} h_i)^2} \cdot \frac{1 + \mathbb{E}[\max_j B_{j,root}]}{2}.$$

Here $B_{j,root} = \max_{\pi} S_j(\pi) \geq 1$ represents the maximum of a specific Cox-Ingersoll-Ross process $S_j(\pi)$ ([Cox et al. \(1985\)](#)), with π indicating the data fraction passed to node d from its parent with splitting variable j . The initial term is the Huber Sandwich Estimator ([Huber, 1967](#)), suitable if the tree structure q wasn’t learned from \mathcal{D}_0 . The latter adjustment term grows with the number of possible splits.

4.1.3 COMBINED REGULARIZATION

We can also combine the constant regularization and UWR and define $\gamma_i = \alpha + \beta\phi(\mathbf{x}_i)$. This combination allows for a base regularization level (α) while still adapting to the uncertainty connected to individual predictions (β). The hyperparameter pair (α, β) dictates the regularization strategy, $(0,0)$ corresponds to the baseline model (i.e., minimizing the squared error with no regularization), whereas $(\alpha > 0, 0)$ and $(0, \beta > 0)$ corresponds to constant regularization and UWR strategy, respectively. The stable loss given a hyperparameter pair (α, β) results in the following estimate of a leaf prediction

$$\hat{w}_d = -\frac{\sum_{i \in I_d} g_i}{\sum_{i \in I_d} h_i}.$$

For squared error loss, this is equivalent to

$$\hat{w}_d = \frac{\sum_{i \in I_d} (y_i + (\alpha + \beta\phi(\mathbf{x}_i)) \hat{w}_0(\mathbf{x}_i))}{\sum_{i \in I_d} (1 + \alpha + \beta\phi(\mathbf{x}_i))}.$$

5 EXPERIMENTS

In this section, we evaluate the Pareto efficiency of the proposed update method with different (α, β) settings on different regression tasks. Note that because we try to achieve two objectives (loss and instability minimization) simultaneously, there is a trade-off between the two objectives. The Pareto front represents all the solutions that have the best-achieved loss for a given level of stability and vice versa. For our experiments, we use squared error loss and squared error instability. We first describe the experimental setups and then present the results.

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate our proposed method on various benchmark regression datasets to get a comprehensive predictability-stability assessment across a broad range of data characteristics, detailed in Table 1.

Dataset	$n \times m$	Description
California	$20,640 \times 8$	Aggregated housing data in the California district using one row per census block group.
Boston	506×13	Housing data for the suburbs of Boston.
Carseats	400×11	Simulated data of child car seat sales.
College	777×18	Statistics for a large number of US Colleges.
Hitters	263×20	Major League Baseball Data from the 1986 and 1987 seasons.
Wage	3000×16	Wage data for male workers in the Mid-Atlantic region.

Table 1: Datasets Description. The California dataset is adapted from the book Sparse Spatial Autoregressions, Statistics and Probability Letters (Kelley Pace & Barry (1997)), whereas the other datasets are from the book "Introduction to Statistical Learning" (ISLR) by James et al. (2022). The dimensions are after data preprocessing, i.e., one-hot encoding and removal of any missing values.

We perform a comprehensive grid search across 121 configurations of the hyperparameters α and β , where both range from 0 to 2 in increments of 0.2. For all experiments we set $\epsilon = 0.01$. **Baseline.** Squared error loss, corresponds to hyperparameter setting ($\alpha = 0, \beta = 0$). **Constant.** Corresponds to hyperparameter setting ($\alpha > 0, \beta = 0$). **UWR.** Corresponds to hyperparameter setting ($\alpha = 0, \beta > 0$); **Combined.** Corresponds to hyperparameter setting ($\alpha > 0, \beta > 0$).

The setups. The setups aim to reflect how a model needs to adapt when new information becomes available. **The main setup** consists of a single update iteration $t = 1$. First, a model is initialized and trained on the currently available data \mathcal{D}_0 . At $t = 1$ new information is available that is added to \mathcal{D}_0 to obtain \mathcal{D}_1 and the model is updated using SL. The loss and stability are computed using repeated cross-validation. For each dataset, the dataset is divided into five folds, where each fold serves as test data once. The remaining four folds represent \mathcal{D}_1 and half of \mathcal{D}_1 represents \mathcal{D}_0 . This process is repeated ten times and the average of the loss and stability estimates from test data is used to obtain the final loss and stability measures. Standard deviations are reported in Table 2 in Appendix A.

Secondary setups. Two additional experiments are conducted using the California dataset. We select five different hyperparameter configurations from the Pareto-efficient configurations identified in the main setup, each representing a different trade-off. **The varied sample sizes setup** uses 5000 data points as test data to compute loss and stability. The dataset \mathcal{D}_1 is obtained by samples n samples without replacement from the remaining data. Half of \mathcal{D}_1 represents \mathcal{D}_0 . An initial model is trained on \mathcal{D}_0 and updated using \mathcal{D}_1 . We run the experiment for different values of $n \in [1000, 5000, 10000, 15000]$ and repeat it 50 times and average loss and stability estimates on test data. **The iterative updates setup** divides the dataset into seven folds, where one fold serves as \mathcal{D}_0 and one serves as test data. The remaining five folds represent additional information, \mathcal{D}_{Δ_t} , obtained at different update iterations. An initial model is trained on \mathcal{D}_0 . For each update iteration a fold is added to \mathcal{D}_0 to obtain \mathcal{D}_t and the model is updated and evaluated on the test data. The experiment is repeated 50 times and averages loss and stability estimates on test data.

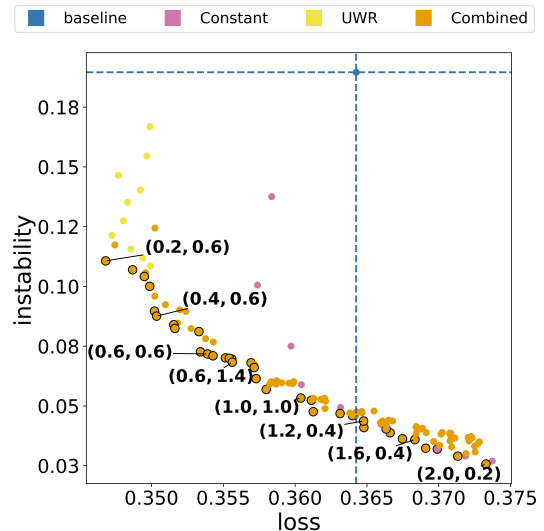


Figure 2: The loss and stability of the model update strategies, highlighting the models that form the Pareto frontier with black borders.

5.2 EXPERIMENTAL RESULTS

Figure 2 shows the results of the main experiment for the *California* dataset. The baseline is the least stable model and is not on the Pareto frontier. As we introduce regularization, more predictive and/or stable solutions can be achieved by tuning α and β . The constant approach shows a consistent reduction in instability as α increases and up to a certain α value, it also reduces loss. The UWR approach results in more stable models than the baseline with lower loss compared to both the baseline and constant approach. The combination of the constant and the UWR approach leads to better solutions than the pure constant and UWR solutions, showcased by the Pareto frontier.

Figure 3 shows and compares the results of the main experiment for all six datasets. Notably, the baseline does not appear on the Pareto frontier for any of the datasets, with all (α, β) configurations being more stable. Furthermore, many configurations also achieve lower loss than the baseline. Comparing the constant and UWR approaches, the constant approach tends to lead to more stable solutions, whereas the UWR approach tends to achieve lower loss, although there is an exception for *Wage* datasets. The combined approach frequently finds Pareto-efficient solutions that are unattainable by either the constant or UWR approaches in isolation. The statistical uncertainty, measured by standard deviation, ranged from 0.04-0.05 and 0.04-0.1 for loss and instability on the *California* dataset. For the other datasets, the uncertainty was between 0.28-0.38 for loss and 0.17-0.47 for instability (see Table 2 in Appendix A). The standard deviations for instability, seem to relate to the regularization strength, with higher values of α and/or β resulting in lower standard deviations.

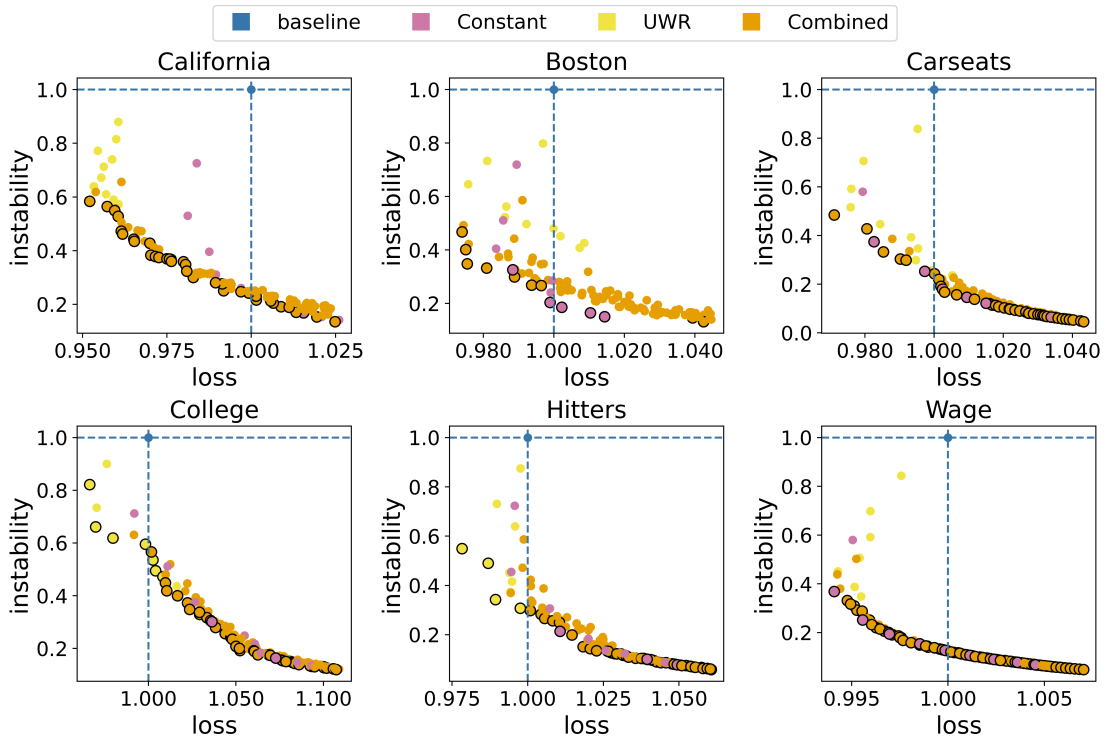


Figure 3: The figure plots the loss and stability of the model update strategies and highlights the models that form the Pareto frontier with black borders. The loss and stability values are relative to the baseline.

The result of the varied sample size experiment (Figure 4a) shows that loss and stability decrease with larger sample sizes for baseline. For the five selected (α, β) configurations, loss decreases while instability remains around the same level. Among the six models, all five configurations are always on the Pareto frontier for all sample sizes. The configurations are consistently more stable than the baseline at corresponding sample sizes. The configurations (0.2, 0.6) and (0.4, 0.6) also consistently achieve lower losses than the baseline.

Figure 4b presents the result of five selected (α, β) configurations over five update iterations. Compared to the baseline, each configuration achieves improved stability at all update iterations. At the first update iteration, the configurations (0.2, 0.6) and (0.4, 0.6) achieve a lower loss than the baseline and at the second update iteration, also configurations (1, 1) and (1.2, 0.4) achieve a lower loss. The next iteration, (1, 1) becomes more stable than (1.2, 0.4) but with a higher

loss than the baseline. By the final update iteration, all configurations except (1, 1) result in a more predictive solution than the baseline. At the first iteration, all configurations are Pareto-efficient solutions, where the configurations (0.2, 0.6), (0.4, 0.6), (1.2, 0.4) and (2, 0.2) are Pareto-efficient models across all update iterations. Note that in practice one would not stick to one configuration across all updates, but rather do a hyperparameter search for each update iteration.

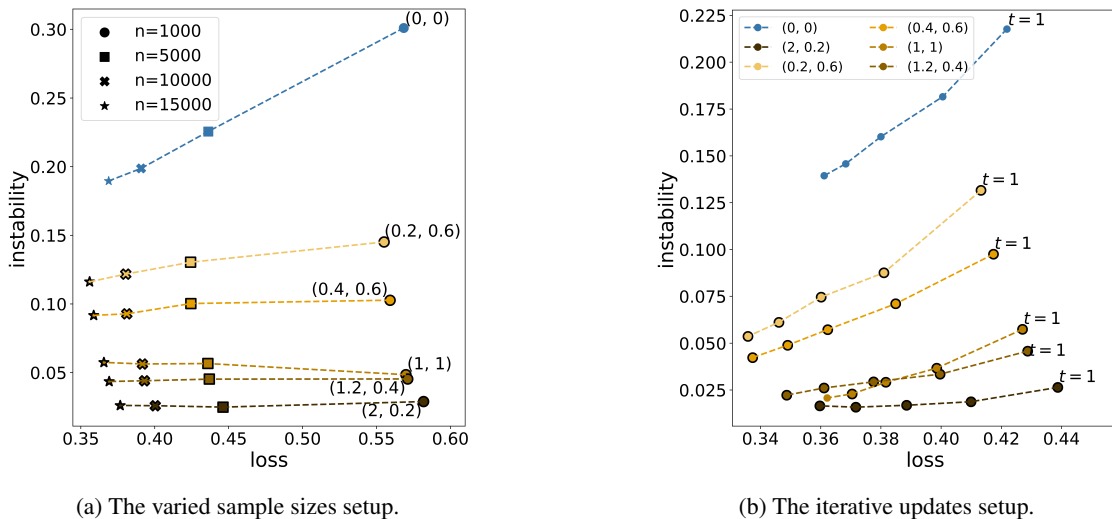


Figure 4: The loss-instability trade-off of model updates with different hyperparameter configurations (α, β) using the California Housing dataset, highlighting Pareto efficient solutions with black borders: (a) Effect of sample size on model updates. (b) Evolution of the loss-instability trade-off over five update iterations, where each point indicates a model’s performance at a specific iteration t .

6 DISCUSSION

The objective of this paper was to provide stable updates for regression trees that better balance the predictability-stability trade-off. We propose a regularization method with two tunable hyperparameters. Through a series of experiments conducted across various datasets and conditions, we demonstrated the limitations of the baseline approach. Compared to our method, the baseline consistently fails to be on the Pareto frontier, suggesting the necessity of regularization to achieve a better trade-off. As we explore the regularization landscape by adjusting α and β , we observe a diverse set of Pareto efficient solutions, each representing a unique balance between predictability and stability.

The main experiment on the California dataset revealed that the constant approach tends to be more stability-focused as stability is equally important for each data point. Incrementally increasing the constant regularization factor α consistently improves model stability and up to a certain value of α also improves predictability. Conversely, the UWR approach tends to be more predictability-focused as stability for predictions in uncertain regions of f_0 receives less weight. The fact that UWR often improves predictability demonstrates the benefit of utilizing prediction uncertainty for regularization. Combining the two approaches leads us to a spectrum of Pareto-efficient solutions, where we can strategically choose (α, β) configurations based on the desired predictability-stability trade-off. These findings showcase that both regularization approaches are relevant. Furthermore, these results also extend to the other datasets, demonstrating the generalizability of our approach. Across all datasets, the baseline is never one of the Pareto-efficient models. All (α, β) configurations led to more stable models with many configurations also achieving lower loss.

Our secondary experiments showed that our regularization method gives consistent results regardless of sample size and that the benefits of our method extend over multiple updates. When updating multiple times, our method continues to outperform the baseline in terms of predictability and/or stability. The models with lower α and β values are the most predictive with steady improvements in stability, while the models with large α and small β are the most stable and remain at the same stability level across updates, with steady improvements in predictability. Having large α and β values seems to quickly become too stability-focused, which comes at the cost of predictability.

A good choice of hyperparameters depends on the user’s preference. Conservative users might prefer stability and therefore opt for larger α and/or β values, while users primarily concerned with performance would rather opt for lower values. In practice, users should specify a stability threshold before model selection.

A notable constraint of our approach is the requirement for batch training, needing all available data to update the model. This is necessary as the CART algorithm, on which our updating strategy relies, is a batch algorithm. Future work could explore ways to remove the reliance on previous data to make the method fit in an incremental learning setting.

The focus of our study was on regression tasks using squared error loss. It would be interesting to apply stable updates of tree models to other tasks, such as classification or survival analysis and use other types of loss and instability functions, such as log loss and Poisson loss. This would also expand the potential for industry applications.

While we show that our method improves the empirical stability of regression trees in terms of semantic stability, we do not consider structural stability in this paper. Structural stability is more directly related to explainability and therefore of substantial interest. Investigating how our method influences the structural stability of trees is a valuable direction for future research.

Our results show that iterative updates improve predictability. This leads us to the following untested hypothesis: Learning a regression tree by starting with a subset of the data and iteratively incorporating more data using our update method improves upon the CART algorithm. This raises the questions of how large the subset should be and how many updates to perform.

Lastly, our focus has been solely on individual regression trees. An interesting extension would be to investigate how our method works in an ensemble setting like random forests.

In conclusion, stable regression trees provide a good way of balancing the predictability-stability trade-off, highlighted by its ability to outperform the baseline approach across various datasets, data sizes and multiple update iterations.

ACKNOWLEDGMENTS

NB was supported in part by the Research Council of Norway grant “Parameterized Complexity for Practical Computing (PCPC)” (NFR, no. 274526).

REFERENCES

- Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.
- Morten Blørstad. Improving Stability of Tree-Based Models. Master’s thesis, University of Bergen, Bergen, 2023.
- Olivier Bousquet and André Elisseeff. Stability and Generalization. *J. Mach. Learn. Res.*, 2:499–526, 3 2002. ISSN 1532-4435. doi: 10.1162/153244302760200704. URL <https://doi.org/10.1162/153244302760200704>.
- L Breiman, Jerome H Friedman, Richard A Olshen, and C J Stone. *Classification and Regression Trees*. New York, 1st edition edition, 1984. doi: <https://doi.org/10.1201/9781315139470>.
- Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. ISSN 1573-0565. doi: 10.1007/BF00058655. URL <https://doi.org/10.1007/BF00058655>.
- Bénédicte Briand, Gilles R Ducharme, Vanessa Parache, and Catherine Mercat-Rommens. A similarity measure to assess the stability of classification trees. *Computational Statistics & Data Analysis*, 53(4):1208–1217, 2009. ISSN 0167-9473. doi: <https://doi.org/10.1016/j.csda.2008.10.033>. URL <https://www.sciencedirect.com/science/article/pii/S0167947308004970>.
- Kenneth P. Burnham and David R. Anderson. *Model Selection and Multimodel Inference*. Springer New York, New York, NY, 2004. ISBN 978-0-387-95364-9. doi: 10.1007/b97636.
- Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. *arXiv*, 3 2016. doi: 10.1145/2939672.2939785. URL <https://arxiv.org/abs/1603.02754>.
- Andrew Cotter, Heinrich Jiang, Serena Wang, Taman Narayan, Maya Gupta, Seungil You, and Karthik Sridharan. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals, 2018.
- John C Cox, Jonathan E Ingersoll, and Stephen A Ross. A Theory of the Term Structure of Interest Rates. *Econometrica*, 53(2):385–407, 1985. ISSN 00129682, 14680262. doi: 10.2307/1911242. URL <http://www.jstor.org/stable/1911242>.

- Felix Dannegger. Tree stability diagnostics and some remedies for instability. *Statistics in medicine*, 19:475–491, 1 2000. doi: 10.1002/(SICI)1097-0258(20000229)19:43.O.CO;2-V.
- L Devroye and T Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979. doi: 10.1109/TIT.1979.1056087.
- Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying Real-world Goals with Dataset Constraints. In D D Lee, M Sugiyama, U V Luxburg, I. Guyon, and R Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2415–2423. 2016.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York, New York, NY, 2001. ISBN 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7.
- Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. 1967. URL <https://api.semanticscholar.org/CorpusID:123642812>.
- Gareth James, Daniela Witten, Trevor Hastie, Rob Tibshirani, and Balasubramanian Narasimhan. ISLR2. Technical report, packaged 2022-11-20, 2022.
- Margot E Kaminski and Gianclaudio Malgieri. Algorithmic impact assessments under the GDPR: producing multi-layered explanations. *International Data Privacy Law*, 11(2):125–144, 4 2021. ISSN 2044-3994. doi: 10.1093/idpl/ipaa020. URL <https://doi.org/10.1093/idpl/ipaa020>.
- Michael Kearns and Dana Ron. Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation. *Neural Computation*, 11(6):1427–1453, 7 1999. ISSN 0899-7667. doi: 10.1162/089976699300016304. URL <https://doi.org/10.1162/089976699300016304>.
- R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. ISSN 0167-7152. doi: [https://doi.org/10.1016/S0167-7152\(96\)00140-X](https://doi.org/10.1016/S0167-7152(96)00140-X). URL <https://www.sciencedirect.com/science/article/pii/S016771529600140X>.
- Mark Last, Oded Maimon, and Einat Minkov. Improving Stability of Decision Trees. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(02):145–159, 2002. doi: 10.1142/S0218001402001599. URL <https://doi.org/10.1142/S0218001402001599>.
- Chingway Lim and B Yu. Estimation Stability with Cross Validation (ESCV). *Journal of Computational and Graphical Statistics*, 25, 1 2013. doi: 10.1080/10618600.2015.1020159.
- Lawrence Lin, A. S Hedayat, Bikas Sinha, and Min Yang. Statistical Methods in Assessing Agreement. *Journal of the American Statistical Association*, 97(457):257–270, 2002. doi: 10.1198/016214502753479392. URL <https://doi.org/10.1198/016214502753479392>.
- Huiting Liu, Avinesh S., Siddharth Patwardhan, Peter Gräsch, and Sachin Agarwal. Model Stability with Continuous Data Updates, 2 2022.
- Berent Ånund Strømnes Lunde, Tore Selland Kleppe, and Hans Julius Skaug. An information criterion for automatic gradient tree boosting. *arXiv*, 8 2020. doi: 10.48550/arxiv.2008.05926. URL <https://arxiv.org/abs/2008.05926>.
- Mahdi Milani Fard, Quentin Cormier, Kevin Canini, and Maya Gupta. Launch and Iterate: Reducing Prediction Churn. In D Lee, M Sugiyama, U Luxburg, I Guyon, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/dc5c768b5dc76a084531934b34601977-Paper.pdf.
- Luis Moneda and Denis Mauá. *Time Robust Trees: Using Temporal Invariance to Improve Generalization*, pp. 385–397. 4 2022. ISBN 978-3-031-21685-5. doi: 10.1007/978-3-031-21686-2_27.
- Michel Philipp, Thomas Rusch, Kurt Hornik, and Carolin Strobl. Measuring the Stability of Results From Supervised Statistical Learning. *Journal of Computational and Graphical Statistics*, 131, 1 2017. doi: 10.1080/10618600.2018.1473779.
- K Takeuchi. Distribution of information statistics and validity criteria of models. *Mathematical Science*, 153:12–18, 5 1976.
- Peter Turney. Technical Note: Bias and the Quantification of Stability. *Machine Learning*, 20:23–33, 7 1995. doi: 10.1007/BF00993473.

