Learning Optimal Classification Trees
Robust to Distribution Shifts

Nathan Justin

Center for Artificial Intelligence in Society, University of Southern California, Los Angeles, CA 90089, USA, njustin@usc.edu

Sina Aghaei

Center for Artificial Intelligence in Society, University of Southern California, Los Angeles, CA 90089, USA, saghaei@usc.edu

Andrés Gómez

Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA 90089, USA, gomezand@usc.edu

Phebe Vayanos

Center for Artificial Intelligence in Society, University of Southern California, Los Angeles, CA 90089, USA, phebe.vayanos@usc.edu

We consider the problem of learning classification trees that are robust to distribution shifts between training and testing/deployment data. This problem arises frequently in high stakes settings such as public health and social work where data is often collected using self-reported surveys which are highly sensitive to e.g., the framing of the questions, the time when and place where the survey is conducted, and the level of comfort the interviewee has in sharing information with the interviewer. We propose a method for learning optimal robust classification trees based on mixed-integer robust optimization technology. In particular, we demonstrate that the problem of learning an optimal robust tree can be cast as a single-stage mixed-integer robust optimization problem with a highly nonlinear and discontinuous objective. We reformulate this problem equivalently as a two-stage linear robust optimization problem for which we devise a tailored solution procedure based on constraint generation. We evaluate the performance of our approach on numerous publicly available datasets, and compare the performance to a regularized, non-robust optimal tree. We show an increase of up to 12.48% in worst-case accuracy and of up to 4.85% in average-case accuracy across several datasets and distribution shifts from using our robust solution in comparison to the non-robust one.

Key words: robust optimization, distribution shift, robust machine learning, mixed-integer optimization, decision trees.

1. Introduction

Machine learning techniques are increasingly being used to support decision-making in high-stakes domains, with potentially significant societal impacts. In such settings, the ability to gauge a machine learning model's trustworthiness is necessary to obtain stakeholder buy-in for

deployment. To this end, simple and interpretable models are often preferred over black-box ones (Rudin 2019, Rudin et al. 2022).

Some of the most interpretable machine learning models are classification trees, which are easily visualized and simple to deploy. A classification tree is a model which takes the form of a binary tree. At each branching node, a test is performed, which asks if a feature exceeds a specified threshold value. Given a data sample, if the answer is positive (resp. negative) the sample is directed to the right (resp. left) descendent. Thus, each data sample, based on its features, follows a path from the root of the tree to a leaf node. At each leaf, a label is predicted. A data sample is correctly classified if and only if its label matches the label predicted at the leaf where it lands (Breiman et al. 2017).

Training a decision tree consists in deciding which tests to perform at each branching node and which labels to predict at each leaf node, based on some training dataset. A common criterion to evaluate the performance of a decision tree is accuracy, defined as the percentage of correctly classified datapoints (although other metrics are also often employed). In high-stakes settings (e.g., when training a tree to identify those at risk of suicidal ideation, or the most vulnerable among those experiencing homelessness), training to optimality is often critical to ensure that the best model is learned. As long as the data used to train the decision tree comes from the same distribution as that in deployment, then as the size of the training data grows, a tree trained to optimality in this fashion will likely perform as expected in deployment.

Unfortunately, *distribution shifts*, where the distribution of the training and testing/deployment data are different, are common in real-world applications, falling in two broad categories. The first kind of shift corresponds to a change in the likelihood of each sample from the population and occurs when there is sampling bias in the training and/or testing population (Shimodaira 2000, Bickel et al. 2007), e.g., when using data from Amazon Mechanical Turk[1] to train a model that is then deployed on the entire population. The second kind of shift, which we focus on in the present paper, corresponds to a change in individual entries of the data, often manifesting as a shift in the covariates. This kind of shift occurs when the data collection mechanism changes between training and deployment, or when the environment changes the distribution of data over time, among other reasons

---

[1] https://www.mturk.com

(Quiñonero-Candela et al. 2009). Shifts that change the data entries are increasingly prevalent in modern machine learning applications, as full control of data collection in both training and deployment is rare. For example, in domains where data is collected from self-reported surveys, changing the way a question is phrased, the location where the survey is conducted or even the person collecting the information, may result in a shift in responses.

Classification trees, similarly to most machine learning models, are susceptible to distribution shifts, implying that their performance may deteriorate in deployment compared to what was expected from training (Zadrozny 2004). Thus, methods accounting for potential shifts during the training phase are needed. In this paper, we propose to learn classification trees that are *robust* to distribution shifts, in the sense that they will achieve the best performance under worst-case/adversarial shifts of the training data, resulting in controlled, reliable performance in deployment.

In social sciences applications where interpretability, robustness, and optimality are often required, the available data frequently come from surveys and manually recorded observations, manifesting as integer and/or categorical data (e.g., gender and race). We thus focus our methodology of learning robust classification trees on the case of covariate shifts on integer and categorical data, a setting of special importance in high-stakes domains that also poses significant modeling and computational challenges.

1.1. Problem Statement

We now formally define the problem of training a classification tree robust to distribution shifts, which is the focus of this paper.

Let $\{\mathbf{x}^i, y^i\}_{i \in \mathcal{I}}$ be the training set, where $\mathcal{I}$ is the index set for the training samples and any categorical features are one-hot encoded in each $\mathbf{x}^i$. The vector $\mathbf{x}^i \in \mathbb{Z}^{|\mathcal{F}|}$ collects the covariates of datapoint i, where the elements of $\mathcal{F}$ index the features, and $y^i \in \mathcal{K}$ is a label drawn from a finite set $\mathcal{K}$. We let x_f^i denote the value of covariate $f \in \mathcal{F}$ for sample $i \in \mathcal{I}$ and, with a slight abuse of notation, denote by $\mathbf{x}$ the vector concatenation of $\mathbf{x}^i$ over all samples $i \in \mathcal{I}$. Accordingly, we let $\mathbf{y} := (y^1, y^2, \ldots, y^{|\mathcal{I}|})$.

In the presence of distribution shifts, the training set does not reflect the data at deployment. In most instances, the exact shift in the data at deployment is unknown. Thus, it is necessary to account for the performance on several perturbed training sets that reflect some potential distributions of the testing set. For a sample $i \in \mathcal{I}$, let $\boldsymbol{\xi}^i \in \mathbb{Z}^{|\mathcal{F}|}$ represent a possible

perturbation of the covariates $\mathbf{x}^i$, such that $\{\mathbf{x}^i + \boldsymbol{\xi}^i, y^i\}_{i \in \mathcal{I}}$ is a potential perturbed dataset. With a slight abuse of notation, let $\boldsymbol{\xi}$ be the vector concatenation of $\boldsymbol{\xi}^i$ over all samples $i \in \mathcal{I}$. Since $\boldsymbol{\xi}$ is unknown in training, we say that $\boldsymbol{\xi}$ lies in a set of all possible perturbations Ξ – often termed as an uncertainty set in the robust optimization literature (Ben-Tal et al. 2009).

Let Π_d be the set of all decision trees with depth at most d, where d is a nonnegative integer. In settings where interpretability is desired, the depth d is usually chosen to be small, e.g., less than four or so. Each element $\pi \in \Pi_d$ is a binary decision tree classifier $\pi : \mathbb{Z}^{|\mathcal{F}|} \to \mathcal{K}$. The problem of training an optimal robust classification tree consists in finding a binary tree that correctly classifies the most samples under worst-case realizations of the perturbed data. Mathematically, it is expressible as

$$\max_{\pi \in \Pi_d} \min_{\boldsymbol{\xi} \in \Xi} \sum_{i \in \mathcal{I}} \mathbb{I}\left[\pi(\mathbf{x}^i + \boldsymbol{\xi}^i) = y^i\right]. \tag{$\mathcal{P}_\Xi$}$$

The above problem is a *single-stage* robust optimization problem, as all decisions, i.e., the classifier π, must be decided *before* the uncertain parameters $\boldsymbol{\xi}$ are observed. Note that since the features are integer-valued, Ξ is a discrete set, a setting that is not well studied in the robust optimization literature (see section 1.2.4). Also, the summation of indicator functions in ($\mathcal{P}_\Xi$) is nonlinear and discontinuous. These characteristics make problem ($\mathcal{P}_\Xi$) difficult to solve. In this paper, we work to solve problem ($\mathcal{P}_\Xi$) and address these difficulties.

1.2. Background and Related Work

Our paper relates closely to several literature streams, which we review in the following.

1.2.1. MIO-Based Decision Trees   Traditionally, classification trees are built using heuristic approaches since the problem of building optimal classification trees is $\mathcal{NP}$-hard (Quinlan 1986, 2014, Kass 1980, Breiman et al. 2017). Recently, motivated by the fact that heuristic approaches often return suboptimal solutions, mathematical optimization techniques such as MIO have been developed for training optimal trees. The first such method was proposed by Bertsimas and Dunn (2017). To address the long run times required for learning optimal decision trees on large datasets, Verwer and Zhang (2019) propose a binary linear programming method whose size scales with the logarithm of the number of training samples. Aghaei et al. (2024) propose a different MIO method with better relaxation quality to the aforementioned formulations, resulting in improved computational performance. Hu

et al. (2019) introduce an efficient method for learning optimal decisions trees with sparsity-promoting regularization, giving specialized bounds that can significantly reduce the search space. Boutilier et al. (2022) present a logic-based Benders decomposition method for building optimal trees with multivariate splits. MIO approaches for constructing decision trees allow for several extensions. Building off of the model by Bertsimas and Dunn (2017), Elmachtoub et al. (2020) propose an MIO formulation to build optimal decision trees that minimize a "predict-then-optimize" loss rather than a classification loss to decrease model complexity while improving decision quality. Mišić (2020) examines the setting where the covariates to a given tree ensemble are decision variables, formulating an MIO problem to find the covariates that maximize the ensemble's predicted value. Aghaei et al. (2019) and Jo et al. (2023) use MIO to create optimal and fair decision trees. Kallus (2017) and Jo et al. (2021) solve the problem of learning prescriptive trees via MIO, with the latter approach applying also to the case of observational data.

1.2.2. Machine Learning Robust to Distribution Shifts Under the setting where there is uncertainty in the parameters of an optimization problem, *robust optimization* has been used to generate solutions immunized to uncertainty by hedging against adversarial realizations of the parameters (Ben-Tal et al. 2009). Using robust optimization, several researchers have proposed models and algorithms for machine learning that are robust to distribution shifts in the data entries.

Chen and Paschalidis (2020) provide a learning framework under distributional perturbations defined by the Wasserstein metric, applying their method to distributionally robust linear regression, semi-supervised learning, and reinforcement learning. More closely related to this paper, there are works on building robust models for a variety of classification tasks. Sinha et al. (2017) provide a distributionally robust framework with proven convergence guarantees for smooth loss functions, and apply their framework to neural networks. Shaham et al. (2018) create a robust optimization framework for non-parametric models that is trained through a stochastic gradient method, and also apply their method to create artificial neural networks that are robust against adversarial perturbations of the data. Both robust and distributionally robust optimization approaches have been used to train robust support vector machines (Shivaswamy et al. 2006, Pant et al. 2011, Bertsimas et al. 2019, Kuhn et al. 2019) and robust logistic regression models (Shafieezadeh Abadeh et al. 2015, Bertsimas et al. 2019, Kuhn et al. 2019).

1.2.3. Robust Classification Trees   Several authors have proposed methods for learning robust classification trees. Chen et al. (2019) and Vos and Verwer (2021) propose a modification of the standard greedy methods to train decision trees that incorporate robustness. More closely related to our work, Bertsimas et al. (2019) utilize MIO to learn robust classification trees. However, their approach does not solve problem ($\mathcal{P}_\Xi$), but instead tackles a simpler (conservative) proxy. We refer to section 5 for comparisons among these methods from the literature and the approach proposed in this paper.

1.2.4. Robust Optimization with Discrete Uncertainty Sets   The approaches discussed in sections 1.2.2 and 1.2.3 assume that the covariates, and therefore the data perturbations, are real-valued. However, real-valued perturbations are usually unrealistic in the high-stakes settings that motivate our work (e.g., in the social sciences and applications using administrative datasets). Additionally, there are very few works in the literature that study problems affected by discrete perturbations. In the case of real-valued perturbations, problem ($\mathcal{P}_\Xi$) is typically handled by dualizing the inner minimization problem. With some exceptions (Bertsimas and Sim 2003, 2004, Mehmanchi et al. 2020), such an approach is usually not possible with non-convex uncertainty sets (such as ours) due to the lack of strong duality. A possible approach is to approximate the discrete set with a convex relaxation (Atamtürk and Gómez 2017) but such methods typically yield poor, conservative solutions (Borrero and Lozano 2021). In general, solving robust optimization problems with discrete uncertainty sets often requires iterations between a worst-case analysis and refinement of a candidate solution. In particular, the worst-case analysis typically requires the solution of an expensive mixed-integer optimization problem (Laporte and Louveaux 1993, Mutapcic and Boyd 2009).

1.3. Contribution and Proposed Approach

We develop the first method for learning optimal robust classification trees.[2] A noteworthy characteristic of our method is that, unlike most approaches in the machine learning literature, it allows uncertainty on discrete and/or categorical covariates. We now summarize the key features of our proposed approach.

---

[2] A preliminary version of this paper was presented at the *2022 AAAI Workshop on Adversarial Machine Learning and Beyond* (Justin et al. 2022). The associated paper was peer-reviewed but is 'non-archival' and therefore the research presented at the workshop can be published in a journal without conflict.

1. We show that the single-stage mixed-integer *nonlinear* robust optimization problem $(\mathcal{P}_\Xi)$ can be formulated equivalently as a *two-stage* mixed-integer *linear* robust problem, where the second-stage (recourse) decisions decide on the path followed by each datapoint in the tree and are allowed to adjust to the realization of the data perturbations. We also detail the connections between this two-stage formulation and previous methods for learning robust trees.

2. We study the unique setting where the data is discrete, a setting that is of particular relevance in high-stakes applications. To address this problem, we devise a discrete uncertainty set through a cost-and-budget framework. We detail a connection between this proposed uncertainty set and hypothesis testing, giving an informed method for calibrating the uncertainty set.

3. We present a cutting plane approach that solves the two-stage formulation to optimality with the proposed uncertainty set, allowing the formulation to be implemented in existing off-the-shelf MIO solvers.

4. We evaluate the performance of the formulation on publicly available datasets for several problem instances and show the effectiveness of our proposed method in mitigating the adverse effects of distribution shifts both on average and in the worst-case. More specifically, in our computations we observe an increase of up to 12.48% in worst-case and 4.85% in average-case accuracy when using a robust tree compared to a non-robust tree in scenarios where the distribution shift is known. The computations also indicate that similar improvements are obtained even if the parameters used to calibrate the uncertainty set are not perfectly known.

The remainder of this paper is organized as follows. Section 2 provides the equivalent reformulation of $(\mathcal{P}_\Xi)$ as a two-stage robust linear MIO and describes the model of uncertainty. In section 3, we describe an algorithm to solve the two-stage formulation to optimality. Section 4 provides a way to calibrate the uncertainty set based on hypothesis testing. We also compare the proposed approach to other robust tree formulations in the literature in section 5. We then present experimental results in section 6. Electronic Companion EC.1 shows the relationship between problem $(\mathcal{P}_\Xi)$ and a Wasserstein distributionally robust version of the problem. Electronic Companions EC.2 and EC.3 contain proofs and empirical results justifying our solution method, respectively. Electronic Companion EC.4 discusses variants and extensions to our model. Finally, we show additional experimental results in
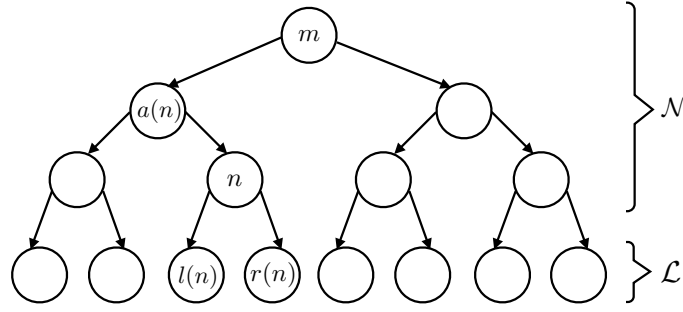
Figure 1   Illustration of our notation on a tree of depth $d = 3$. The nodes in the bottom row are in set $\mathcal{L}$ and all other nodes are in set $\mathcal{N}$. Each node $n \in \mathcal{N}$ has two children: $l(n)$ (on the left) and $r(n)$ (on the right), and each node $n \in \mathcal{N} \cup \mathcal{L} \setminus \{1\}$ has a parent $a(n)$ and ancestors indexed in set $\mathcal{A}(n) = \{m, a(n)\}$.

Electronic Companion EC.5. Note that the approach in this paper has been implemented as part of an open source Python package ODTLearn (Vossler et al. 2023), making it easy to use for researchers and practitioners.

2. Robust Tree Formulation

In problem $(\mathcal{P}_{\Xi})$, the objective function is highly nonlinear and discontinuous, causing significant difficulties in devising computational solution approaches. We thus propose to reformulate $(\mathcal{P}_{\Xi})$ equivalently as a two-stage problem with *linear* objective, which will be possible to solve using a conjunction of MIO solvers and delayed constraint generation, see section 3.

2.1. Defining a Classification Tree

To reformulate problem $(\mathcal{P}_{\Xi})$ as a two-stage problem, we first express Π_d as a set of points satisfying a finite number of linear constraints over a discrete set. For a tree of maximum depth d, let $\mathcal{N}$ be the set of internal nodes and $\mathcal{L}$ be the set of leaf nodes. There are $2^d - 1$ nodes in $\mathcal{N}$ and 2^d nodes in $\mathcal{L}$ and each node is numbered from 1 to $2^{d+1} - 1$ in a breadth-first search pattern. For any internal node $n \in \mathcal{N}$, let the left and right children of n be $l(n)$ and $r(n)$, respectively. For any node $n \in \mathcal{N} \cup \mathcal{L}$, let $a(n)$ be the parent of node n and $\mathcal{A}(n)$ be the set of all ancestor nodes of n. Figure 1 illustrates this notation on a depth 3 classification tree.

Each node in $\mathcal{L}$ can be a *prediction node* where the classification of a sample is made; each node in $\mathcal{N}$ can be a prediction node or a *branching node* where a binary test is performed.

Any node that is not the root node is a *pruned node* (i.e., neither a branching nor prediction node) if any of its ancestors is a prediction node. At each branching node $n \in \mathcal{N}$, a binary test is performed on a chosen feature $f \in \mathcal{F}$ with selected threshold θ, asking whether the perturbed value of the feature of sample i is greater than θ, i.e., if $x_f^i + \xi_f^i \geq \theta + 1$. If the test passes, then the data sample is directed right to $r(n)$; otherwise the sample is directed left to $l(n)$. We let c_f and d_f be the lower and upper bounds of values for feature f within the training data, respectively. The set of possible threshold values we can choose from if branching on feature f is given by

$$\Theta(f) := \{\theta \in \mathbb{Z} \mid c_f \leq \theta < d_f\}.$$

We now encode elements of Π_d as a discrete set of decision variables for the outer maximization of problem ($\mathcal{P}_{\Xi}$). Let $b_{nf\theta}$ indicate whether node $n \in \mathcal{N}$ is a branching node and the binary test is on feature $f \in \mathcal{F}$ with threshold $\theta \in \Theta(f)$. Binary decision variable v_n indicates whether node $n \in \mathcal{N} \cup \mathcal{L}$ is a prediction node. Also, we let w_{nk} be a binary variable that equals one when node $n \in \mathcal{N} \cup \mathcal{L}$ is a prediction node with assignment label $k \in \mathcal{K}$. Let $\mathbf{b}$, $\mathbf{v}$, and $\mathbf{w}$ collect the $b_{nf\theta}$, v_n, and w_{nk} decision variables, respectively. With these variables in hand, we can define a set $\mathcal{S}$ whose elements have a one-to-one mapping to the trees in Π_d:

$$\mathcal{S} := \{ (\mathbf{b}, \mathbf{v}, \mathbf{w}) :$$

$$\sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} b_{nf\theta} + v_n + \sum_{m \in \mathcal{A}(n)} v_m = 1 \qquad \forall n \in \mathcal{N} \qquad (2a)$$

$$v_n + \sum_{m \in \mathcal{A}(n)} v_m = 1 \qquad \forall n \in \mathcal{L} \qquad (2b)$$

$$v_n = \sum_{k \in \mathcal{K}} w_{nk} \qquad \forall n \in \mathcal{N} \cup \mathcal{L} \qquad (2c)$$

$$b_{nf\theta} \in \{0, 1\} \qquad \forall n \in \mathcal{N}, f \in \mathcal{F}, \theta \in \Theta(f) \qquad (2d)$$

$$v_n \in \{0, 1\} \qquad \forall n \in \mathcal{N} \cup \mathcal{L} \qquad (2e)$$

$$w_{nk} \in \{0, 1\} \qquad \forall n \in \mathcal{N} \cup \mathcal{L}, k \in \mathcal{K} \}. \qquad (2f)$$

Constraints (2a) state that at each internal node, either a prediction is made at the node, a prediction is made at an ancestor, or a binary test is performed at the node. Constraints (2b) affirm that at each leaf node, a prediction is made at either the node or one of its ancestors. Constraints (2c) ensures that exactly one label in $\mathcal{K}$ is predicted at each prediction node.

2.2. The Uncertainty Set

For the time being, we assume that the covariates of all features are integer-valued. We also assume that all uncertain data may shift in either direction by increasing or decreasing in value. We refer to the Electronic Companion EC.4 for handling categorical features, one-sided shifts that only allow increases (or only decreases) in uncertain data values, and datasets with mixed-feature types and shifts. With these assumptions in place for the inner minimization problem in $(\mathcal{P}_\Xi)$, the uncertainty set Ξ is defined as

$$\Xi := \left\{ \boldsymbol{\xi} \in \mathbb{Z}^{|\mathcal{I}| \times |\mathcal{F}|} \; : \; \sum_{i \in \mathcal{I}} \sum_{f \in \mathcal{F}} \gamma_f^i |\xi_f^i| \le \epsilon \right\}, \tag{3}$$

where $\gamma_f^i \in \mathbb{R}_+$ is the cost of perturbing x_f^i by one (in either direction) and ϵ is the total allowable budget of uncertainty across data samples. Note that γ_f^i can take on different values for different samples $i \in \mathcal{I}$ and features $f \in \mathcal{F}$, which is useful in domains where the likelihood of the covariate shift varies between different groups of samples and/or different features.

As we show in section 4, there exists a connection between (3) and hypothesis testing, which allows for calibration of the values of γ_f^i and ϵ based on domain knowledge about the possible distribution shifts. For details on the connection between (3) and a Wasserstein-based ambiguity set, we refer to Electronic Companion EC.1.

2.3. Counting the Number of Correctly Classified Data Samples

We wish to reformulate the expression for the count of correctly classified data samples in $(\mathcal{P}_\Xi)$ to create a more computationally tractable formulation. To represent the objective, we use the idea outlined by Aghaei et al. (2024) that the number of correctly classified samples can be represented as the optimal value of a sum of maximum flow problems, where we adapt this approach to account for perturbations in the data.

2.3.1. Capacitated Flow Graph for Sample i To determine whether a data sample i is correctly classified for a given tree $(\mathbf{b}, \mathbf{v}, \mathbf{w}) \in \mathcal{S}$ and perturbation $\boldsymbol{\xi} \in \Xi$, we define a *capacitated flow graph for sample* i based on $(\mathbf{b}, \mathbf{v}, \mathbf{w})$, $\mathbf{x}^i$, and $\boldsymbol{\xi}^i$:

DEFINITION 1 (CAPACITATED FLOW GRAPH FOR SAMPLE i). *Let* $\mathcal{V} := \mathcal{N} \cup \mathcal{L} \cup \{s, t\}$ *for s a source node and t a sink node. Given tree* $(\mathbf{b}, \mathbf{v}, \mathbf{w}) \in \mathcal{S}$, *the* capacitated flow graph *for sample $i \in \mathcal{I}$ is a graph* $(\mathcal{V}, \mathcal{E})$ *such that for* all $n \in \mathcal{V} \backslash \{t\}$ *and $m \in \mathcal{V} \backslash \{s\}$, edge (n, m) is *in $\mathcal{E}$ and has capacity 1 if and only if* exactly *one of the following is true:*

1. $n = s$ *and* $m = 1$;

2. $n \in \mathcal{N}$, *$m = l(n)$, and there exists an $f \in \mathcal{F}$ and $\theta \in \Theta(f)$ such that $b_{nf\theta} = 1$ and* $x_f^i + \xi_f^i \leq \theta$;

3. $n \in \mathcal{N}$, *$m = r(n)$, and there exists an $f \in \mathcal{F}$ and $\theta \in \Theta(f)$ such that $b_{nf\theta} = 1$ and* $x_f^i + \xi_f^i \geq \theta + 1$;

4. $n \in \mathcal{N} \cup \mathcal{L}$, *$m = t$, and $w_{ny^i} = 1$.*

Furthermore, edge (n, m) is in $\mathcal{E}$ and has capacity 0 if and only if one of the following is true:

1. $n \in \mathcal{N}$, *$m = l(n)$, and there exists an $f \in \mathcal{F}$ and $\theta \in \Theta(f)$ such that $b_{nf\theta} = 1$ and* $x_f^i + \xi_f^i \geq \theta + 1$;

2. $n \in \mathcal{N}$, *$m = r(n)$, and there exists an $f \in \mathcal{F}$ and $\theta \in \Theta(f)$ such that $b_{nf\theta} = 1$ and* $x_f^i + \xi_f^i \leq \theta$;

3. $n \in \mathcal{N} \cup \mathcal{L}$, *$m = t$, and $w_{ny^i} = 0$.*

For every branching node n with test on threshold θ of feature f, the capacitated flow graph has either capacity 0 or 1 for the edge leading to the sink node t and the edge that is traversed. Specifically, if the test $x_f^i + \xi_f^i \geq \theta + 1$ passes (resp. fails), edge $(n, r(n))$ (resp. $(n, l(n))$) has capacity 1 and edges (n, t) and $(n, l(n))$ (resp. $(n, r(n))$) have capacity 0. For every prediction node, the edge leading to t has capacity 1 only if the assigned class of that node is y^i, and all other edges leaving the prediction node have capacity 0. Lastly, as an entry point for a data sample, the edge from the source s to the root node 1 has capacity 1. This constructed capacitated flow graph for sample i has a maximum flow from s to t of 1 if and only if sample i, perturbed by $\boldsymbol{\xi}^i$, is correctly classified. Figure 2 illustrates the construction of the capacitated flow graph for sample i.

2.3.2. Maximum Flow Problem We now introduce decision variables that represent the flow of point i from source s to t in the capacitated flow graph. Let the binary decision variable $z_{n,m}^i$ indicate whether datapoint i flows down the edge between n and m and is correctly classified by the tree for $m \in \mathcal{N} \cup \mathcal{L} \cup \{t\}$ and $n := a(m) \in \mathcal{N} \cup \mathcal{L} \cup \{s\}$ under perturbation $\boldsymbol{\xi}$. Also let $\mathbf{z}$ collect all the $z_{a(m),m}^i$ for all $i \in \mathcal{I}$ and all graph edges $(a(m), m)$ for $m \in \mathcal{N} \cup \mathcal{L} \cup \{s\}$. Note that $z_{n,m}^i$ are the decision variables of a maximum flow problem, where in the capacitated flow graph for sample i, $z_{n,m}^i$ is 1 if and only if the maximum flow is 1 and the flow traverses arc (n, m). For fixed $(\mathbf{b}, \mathbf{v}, \mathbf{w})$ and $\boldsymbol{\xi}$, maximizing the sum of $z_{n,t}^i$ over all samples $i \in \mathcal{I}$ and all nodes $n \in \mathcal{N} \cup \mathcal{L}$ yields the count of correctly classified samples

$$\max_{\mathbf{z} \in \mathcal{Z}(\mathbf{b}, \mathbf{w}, \boldsymbol{\xi})} \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N} \cup \mathcal{L}} z_{n,t}^i, \tag{4}$$
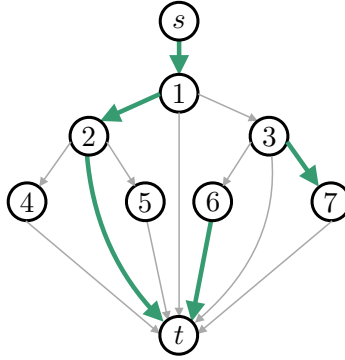
Figure 2   Capacitated flow graph for a given sample i with a classification tree of depth 2. Nodes s and t are the
source and sink nodes, respectively, $\mathcal{N} = \{1,2,3\}$, and $\mathcal{L} = \{4,5,6,7\}$. The green, bold edges are the
edges of capacity 1 in the capacitated flow graph, and the gray edges have capacity 0. Nodes 1 and
3 are branching nodes, where the sample would be routed left and right, respectively. Nodes 2 and 6
assign the correct label, and node 7 assigns an incorrect label. Nodes 4 and 5 are pruned from the tree
as they have an ancestor (i.e., 2) that is a prediction node. The maximum flow from s to t is 1 in this
capacitated flow graph, indicating a correct classification.

where the set $\mathcal{Z}$ defines the maximum flow constraints for each sample's capacitated flow
graph:

$$\mathcal{Z}(\mathbf{b}, \mathbf{w}, \boldsymbol{\xi}) := \{ \ \mathbf{z} \in \{0,1\}^{|\mathcal{I}| \times (2^{d+2}-2)} :$$

$$z_{n,l(n)}^i \le \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} \mathbb{I}\left[x_f^i + \xi_f^i \le \theta\right] b_{nf\theta} \qquad \forall i \in \mathcal{I}, n \in \mathcal{N}, \quad (5a)$$

$$z_{n,r(n)}^i \le \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} \mathbb{I}\left[x_f^i + \xi_f^i \ge \theta+1\right] b_{nf\theta} \qquad \forall i \in \mathcal{I}, n \in \mathcal{N}, \quad (5b)$$

$$z_{a(n),n}^i = z_{n,l(n)}^i + z_{n,r(n)}^i + z_{n,t}^i \qquad \forall i \in \mathcal{I}, n \in \mathcal{N}, \quad (5c)$$

$$z_{a(n),n}^i = z_{n,t}^i \qquad \forall i \in \mathcal{I}, n \in \mathcal{L}, \quad (5d)$$

$$z_{n,t}^i \le w_{n,y^i} \qquad \forall i \in \mathcal{I}, n \in \mathcal{N} \cup \mathcal{L} \ \}. \quad (5e)$$

Problem (4) maximizes the sum of flows over the capacitated flow graphs for all data samples,
counting the number of correctly classified samples after perturbation. Constraints (5a)
and (5b) are capacity constraints that control the flow of samples based on $\mathbf{x} + \boldsymbol{\xi}$ and the tree
structure. Constraints (5c) and (5d) are flow conservation constraints. Lastly, constraint (5e)
blocks any flow to the sink if the node is either not a prediction node or the classification at
that node is incorrect.

2.4. Two-Stage Reformulation

With the definition of $\mathcal{S}$ in (2), definition of Ξ in (3), and reformulation of the number of correctly classified datapoints as the optimal value of the maximum flow problem (4) in hand, we can rewrite problem $(\mathcal{P}_\Xi)$ equivalently as a two-stage *linear* robust optimization problem. In the *first-stage*, the variables $(\mathbf{b}, \mathbf{v}, \mathbf{w}) \in \mathcal{S}$ that encode the tree are selected, corresponding to the outer maximization in problem $(\mathcal{P}_\Xi)$. Once the tree is selected, an adversarial perturbation of the covariates $\boldsymbol{\xi}$ from the set Ξ is chosen, corresponding to the inner minimization in $(\mathcal{P}_\Xi)$. In the *second-stage* problem, given $(\mathbf{b}, \mathbf{v}, \mathbf{w})$ and $\boldsymbol{\xi}$, the number of correctly classified perturbed samples in the data $\{\mathbf{x}^i + \boldsymbol{\xi}^i, y^i\}_{i \in \mathcal{I}}$ is calculated as in (4). This idea leads to the following equivalent reformulation of problem $(\mathcal{P}_\Xi)$ for learning optimal robust classification trees:

$$\max_{(\mathbf{b},\mathbf{v},\mathbf{w}) \in \mathcal{S}} \ \min_{\boldsymbol{\xi} \in \Xi} \ \max_{\mathbf{z} \in \mathcal{Z}(\mathbf{b},\mathbf{w},\boldsymbol{\xi})} \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N} \cup \mathcal{L}} z_{n,t}^i. \tag{6}$$

Problem (6) is equivalent to problem $(\mathcal{P}_\Xi)$, but unlike formulation $(\mathcal{P}_\Xi)$, it has a linear objective with a linear set of constraints for each stage of the formulation. However, (6) complicates the problem in that it introduces a second stage maximization. Yet in spite of problem (6) being a two-stage problem, this equivalent reformulation of $(\mathcal{P}_\Xi)$ can be solved to optimality with the help of MIO solvers, as we show in section 3.

REMARK 1. *Note that the set $\mathcal{S}$ does not involve any constraints that prevent tests that do not improve the tree's accuracy from being performed. To prevent trees that involve such tests from being returned when solving problem* (6), *we can add a small penalty term to our objective, ensuring that shallow trees are preferred over deeper trees with the same accuracy. We include such a term in all our experiments (see section 6.1.2 for details).*

3. Solution Method

We now present a method for solving problem (6) (and therefore problem $(\mathcal{P}_\Xi)$) through a reformulation that can leverage existing off-the-shelf MIO solvers. There are several exact methods for solving two-stage robust optimization problems. If strong duality holds for the inner maximization problem (as it does in problem (6)), replacing the inner maximization problem with its dual results in an equivalent reformulation with the same optimal objective value and same set of optimal solutions. This dual reformulation is a single-stage robust optimization problem, which can be solved using a delayed constraint generation approach that

adds violated constraints during branch-and-cut. We can also use a column-and-constraint generation approach as in Zeng and Zhao (2013), which iteratively adds both constraints and the inner maximization problem's primal variables to a main problem and resolves the main problem. However, our main problem, represented by the outer maximization of (6), is difficult to solve as shown in Aghaei et al. (2024). Column-and-constraint generation therefore would necessitate iteratively solving many difficult main problems in succession. Thus, we utilize a tailored delayed constraint generation approach that leverages the structure of the problem, implementable with common off-the-shelf MIO solvers such as Gurobi (Gurobi Optimization, LLC 2024).

3.1. Reformulating the Two-Stage Problem

To solve the two-stage optimization problem (6), we reformulate it equivalently as a single-stage robust MIO. The first part of the reformulation is the dualization of the inner maximization problem (4). Recall that the inner maximization problem is a maximum flow problem; therefore, the dual of the inner maximization problem is a minimum cut problem.

To write the dual, we define the dual variables corresponding to the minimum cut problem. Let $q_{n,l(n)}^i$, $q_{n,r(n)}^i$, and $q_{n,t}^i$ be the dual variables corresponding to the capacity constraints (5a), (5b), and (5e), respectively. Each $q_{n,m}^i$ equals 1 if and only if the edge that connects nodes n and m in the capacitated flow graph for sample i is in the minimum cut-set. We let $\mathbf{q}^i$ be the collection of $q_{n,m}^i$ over tree edges (n,m) so that $\mathbf{q}^i$ represents a cut-set on the capacitated flow graph for data sample i. We also define p_n^i to be a dual variable corresponding to the flow conservation constraints (5c) and (5d) for $n \in \mathcal{N} \cup \{s\}$ and $n \in \mathcal{L}$, respectively. Each p_n^i equals 1 if and only if node n is in the source set corresponding to the capacitated flow graph of data sample $i \in \mathcal{I}$.

Let $\mathcal{Q}$ be the set of all possible cut-sets in a classification tree described in section 2.1, defined through

$$\mathcal{Q} := \{\, \mathbf{q} \in \{0,1\}^{(2^{d+2}-2)} \;:\; \exists p_n \in \{0,1\}, \forall n \in \mathcal{N} \cup \mathcal{L} \cup \{s\} \text{ s.t.}$$

$$q_{n,l(n)} - p_n + p_{l(n)} \geq 0 \qquad\qquad \forall n \in \mathcal{N} \qquad (7a)$$

$$q_{n,r(n)} - p_n + p_{r(n)} \geq 0 \qquad\qquad \forall n \in \mathcal{N} \qquad (7b)$$

$$q_{s,1} + p_1 \geq 1 \qquad\qquad (7c)$$

$$-p_n + q_{n,t} \geq 0 \qquad\qquad \forall n \in \mathcal{N} \cup \mathcal{L} \,\}. \qquad (7d)$$

Constraints (7a)-(7d) ensure that if a given node n is in the source set and one of its children is in the sink set, then the arc connecting the nodes is in the cut-set. Moreover, let $\mathbf{q} \in \mathcal{Q}^{|\mathcal{I}|}$ be a collection of cut-sets $\mathbf{q}^i \in \mathcal{Q}$ across data samples $i \in \mathcal{I}$. Then, taking the dual of the inner maximization problem in (6) gives the following single-stage formulation:

$$
\max_{(\mathbf{b},\mathbf{v},\mathbf{w})\in\mathcal{S}} \min_{\mathbf{q}\in\mathcal{Q}^{|\mathcal{I}|},\boldsymbol{\xi}\in\Xi} \sum_{i\in\mathcal{I}}\sum_{n\in\mathcal{N}\cup\mathcal{L}} q^i_{n,t}w_{n,y^i} + \sum_{i\in\mathcal{I}} q^i_{s,1} + \sum_{i\in\mathcal{I}}\sum_{n\in\mathcal{N}}\sum_{f\in\mathcal{F}}\sum_{\theta\in\Theta(f)} \mathbb{I}\left[x^i_f + \xi^i_f \leq \theta\right] q^i_{n,l(n)}b_{nf\theta}
$$
$$
+ \sum_{i\in\mathcal{I}}\sum_{n\in\mathcal{N}}\sum_{f\in\mathcal{F}}\sum_{\theta\in\Theta(f)} \mathbb{I}\left[x^i_f + \xi^i_f \geq \theta+1\right] q^i_{n,r(n)}b_{nf\theta}.
\tag{8}
$$

Equivalently, we can formulate (8) using the hypograph reformulation

$$
\max_{\mathbf{b},\mathbf{v},\mathbf{w},t} t \tag{9a}
$$
$$
\text{s.t.} \ \ t \leq \sum_{i\in\mathcal{I}}\sum_{n\in\mathcal{N}\cup\mathcal{L}} q^i_{n,t}w_{n,y^i} + \sum_{i\in\mathcal{I}} q^i_{s,1} + \sum_{i\in\mathcal{I}}\sum_{n\in\mathcal{N}}\sum_{f\in\mathcal{F}}\sum_{\theta\in\Theta(f)} \mathbb{I}\left[x^i_f + \xi^i_f \leq \theta\right] q^i_{n,l(n)}b_{nf\theta}
$$
$$
+ \sum_{i\in\mathcal{I}}\sum_{n\in\mathcal{N}}\sum_{f\in\mathcal{F}}\sum_{\theta\in\Theta(f)} \mathbb{I}\left[x^i_f + \xi^i_f \geq \theta+1\right] q^i_{n,r(n)}b_{nf\theta} \qquad \forall \mathbf{q}\in\mathcal{Q}^{|\mathcal{I}|}, \boldsymbol{\xi}\in\Xi \tag{9b}
$$
$$
(\mathbf{b},\mathbf{v},\mathbf{w}) \in \mathcal{S} \tag{9c}
$$
$$
t \in \mathbb{R}, \tag{9d}
$$

where decision variable $t \in \mathbb{R}$ and constraints (9b) represent the hypograph of the objective function of (8). Note that formulation (9) is a linear MIO formulation for solving problem $(\mathcal{P}_\Xi)$, where the optimal value of t represents the optimal objective value of $(\mathcal{P}_\Xi)$ – that is, the number of correctly classified datapoints in training under a worst-case realization of the perturbed data. However, (9) introduces an extremely large number of constraints (9b), one for each combination of cut-set and perturbation in $\mathcal{Q}^{|\mathcal{I}|} \times \Xi$, and is impractical to solve directly using MIO solvers.

3.2. Solving the Single-Stage Reformulation

A common method for solving problems with a large number of constraints is to use a cutting plane approach: solving a simpler relaxation of the original problem as an initial main problem, then iteratively adding constraints as needed. This method typically avoids solving a problem with a prohibitive number of constraints. We propose such a delayed constraint generation approach to solve formulation (9).

3.2.1. The Main Problem The approach begins with a main problem that initially relaxes all constraints (9b) within formulation (9):

$$\max_{\mathbf{b},\mathbf{v},\mathbf{w},t} \ t \tag{10a}$$

$$\text{s.t.} \quad t \leq |\mathcal{I}| \tag{10b}$$

$$(\mathbf{b},\mathbf{v},\mathbf{w}) \in \mathcal{S} \tag{10c}$$

$$t \in \mathbb{R}, \tag{10d}$$

where constraint (10b) bounds t by its maximum value, i.e., the size of the training set.

Formulation (10) can be solved with a linear programming-based branch-and-bound algorithm: a standard approach used to solve MIOs in most solvers. At any integer solution in the branch-and-bound tree, we find and add a violated constraint of the form (9b) to (10) if one exists. We describe such a process to find a violated constraint in the following section. By adding violated constraints to the main problem along the branch-and-bound process, we converge to the optimal solution of (9), which occurs when no violated constraints can be found for the candidate solution and no better integer solutions exist in the branch-and-bound tree. Note that most MIO solvers allow adding constraints as indicated via callbacks.

3.2.2. The Subproblem We now describe how to check whether a given solution to the relaxed main problem is feasible for (9), and how to find a violated constraint (9b) when infeasible. Given an integer solution of the main problem $(\bar{\mathbf{b}}, \bar{\mathbf{v}}, \bar{\mathbf{w}}, \bar{t})$, we consider the following subproblem:

$$\min_{\mathbf{q} \in \mathcal{Q}^{|\mathcal{I}|}, \boldsymbol{\xi} \in \Xi} \left\{ \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N} \cup \mathcal{L}} q_{n,t}^i \bar{w}_{n,y^i} + \sum_{i \in \mathcal{I}} q_{s,1}^i + \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} \mathbb{I}\left[x_f^i + \xi_f^i \leq \theta\right] q_{n,l(n)}^i \bar{b}_{nf\theta} \right.$$
$$\left. + \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} \mathbb{I}\left[x_f^i + \xi_f^i \geq \theta + 1\right] q_{n,r(n)}^i \bar{b}_{nf\theta} \right\}. \tag{11}$$

Note that the objective function of problem (11) is the right-hand side of constraints (9b). Hence, if $\bar{t}$ is greater than the optimal value of (11), then the cut (9b) where $(\mathbf{q}, \boldsymbol{\xi})$ are fixed to their values in an optimal solution $(\tilde{\mathbf{q}}, \tilde{\boldsymbol{\xi}})$ of (11) yields a violated constraint that can be added to the main problem.

Problem (11) contains products of terms that jointly optimize over $\mathbf{q}$ and $\boldsymbol{\xi}$, presenting its own challenges to solve. The key observations that enable the effective solution of problem (11) given a tree defined by $(\bar{\mathbf{b}}, \bar{\mathbf{v}}, \bar{\mathbf{w}})$ are that: *1)* Under a given perturbation $\boldsymbol{\xi}$, each datapoint is

either correctly classified or misclassified, contributing 1 or 0 to the objective, respectively; *2)* It is never beneficial to perturb (i.e., to have $\boldsymbol{\xi}^i \neq \mathbf{0}$ for) a datapoint i that is already misclassified; *3)* For any given datapoint that is correctly classified, if it is beneficial to misclassify it, we should do so in the minimum cost way; and, *4)* Identifying the minimum cost way of perturbing any given datapoint can be done separately for each datapoint. These observations imply that, for any given tree $(\bar{\mathbf{b}}, \bar{\mathbf{v}}, \bar{\mathbf{w}})$, we can find the worst-case pair $(\tilde{\mathbf{q}}, \tilde{\boldsymbol{\xi}})$ that solves problem (11) through a two-step process. First, solve a collection of auxiliary subproblems each of which identifies a pair $(\bar{\mathbf{q}}^i, \bar{\boldsymbol{\xi}}^i)$ encoding a worst-case path and associated worst-case perturbation that misclassify each datapoint i at minimum cost (ignoring the perturbation budget ϵ). Second, rank datapoints by minimum cost of misclassification from lowest to highest and perturb as many datapoints as possible according to $\bar{\boldsymbol{\xi}}^i$, starting from the lowest cost ones and without exceeding the budget ϵ. In the following, we detail and justify this procedure and propose, in particular, an efficient way to identify $(\bar{\mathbf{q}}^i, \bar{\boldsymbol{\xi}}^i)$ for each datapoint i, resulting in a procedure that solves (11) in $\mathcal{O}(|\mathcal{L}||\mathcal{I}| + |\mathcal{I}| \log |\mathcal{I}|)$.

To help formulate the procedure of finding an optimal $\mathbf{q}$ and $\boldsymbol{\xi}$ in (11), we now define a *decision path for sample i* and its associated *decision cut-set* $\mathbf{q}^i$.

DEFINITION 2 (DECISION PATH FOR SAMPLE i). *Given a tree $(\bar{\mathbf{b}}, \bar{\mathbf{v}}, \bar{\mathbf{w}})$ and sample $i \in \mathcal{I}$, a decision path is a sequence of nodes $(s, n_1, n_2, \ldots, n_k, t) \subseteq \mathcal{N} \cup \mathcal{L} \cup \{s, t\}$ for $k \in \{1, 2, \ldots, d+1\}$ such that, in the capacitated flow graph for sample i, $n_j = a(n_{j+1})$ for all $j \in \{2, \ldots, k-1\}$, n_1 is the root node 1, and n_k is a prediction node (i.e., $v_{n_k} = 1$).*

We encode a decision path for sample i in our subproblem (11) through a decision cut-set $\mathbf{q}^i$, defined as follows.

DEFINITION 3 (DECISION CUT-SET). *Given a decision path $(s, n_1, n_2, \ldots, n_k, t) \subseteq \mathcal{N} \cup \mathcal{L} \cup \{s, t\}$ and sample i, a decision cut-set $\mathbf{q}^i$ is the element of $\mathcal{Q}$ such that $q^i_{n,m} = 1$ if and only if $n \in \{n_1, \ldots, n_k\}$ and $m \notin \{n_2, \ldots, n_k\}$.*

Accordingly, the value of a decision cut-set for sample i is defined as the sum of edge capacities of the cut-set in the capacitated flow graph for sample i. So, sample i is misclassified on a decision path if and only if the associated decision cut-set has a value of 0. We provide examples of decision paths and associated decision cut-sets in Figure 3.

To find the minimum cost perturbation that misclassifies a sample i, we enumerate over all decision paths that misclassify sample i and identify the one with the lowest cost perturbation.
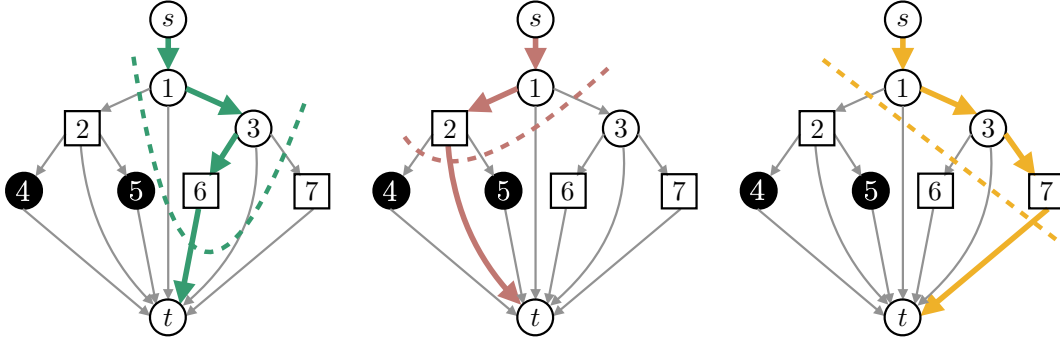
Figure 3   Visual example of all the decision paths and their associated decision cut-sets in the tree described in
Figure 2. Branching nodes are numbered circles (nodes 1 and 3), prediction nodes are numbered squares
(nodes 2, 6, and 7), and pruned nodes are filled-in circles (nodes 4 and 5). There are three decision paths
in this tree, denoted by the colored, thick edges. For each decision path, its associated decision cut-set
$\mathbf{q}^i$ is denoted by the dotted line, where $q_{n,m}^i = 1$ if and only if edge (n,m) crosses the dotted line. For
example, the decision path on the left is $(s, 1, 3, 6, t)$ and the associated decision cut-set has $q_{1,2}^i = 1$,
$q_{1,t}^i = 1$, $q_{3,7}^i = 1$, $q_{3,t}^i = 1$, $q_{6,t}^i = 1$, and all other $q_{n,m}^i = 0$. Using the example capacitated flow graph in
Figure 2, the left and middle decision cut-sets have value 1, and the right decision cut-set has value 0.

For each of these decision paths for sample i, we identify the minimum cost perturbation
that ensures the sample flows through the decision path by solving the following problem
for $\boldsymbol{\xi}^i$ given the decision path's associated decision cut-set $\bar{\mathbf{q}}^i$:

$$
\min_{\boldsymbol{\xi}^i \in \mathbb{Z}^{|\mathcal{F}|}} \sum_{f \in \mathcal{F}} \gamma_f^i |\xi_f^i|
$$
$$
\text{s.t.} \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} \mathbb{I}\left[x_f^i + \xi_f^i \le \theta\right] \bar{b}_{nf\theta} \bar{q}_{n,l(n)}^i + \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} \mathbb{I}\left[x_f^i + \xi_f^i \ge \theta + 1\right] \bar{b}_{nf\theta} \bar{q}_{n,r(n)}^i = 0.
$$
$$(12)$$

In this formulation, the objective minimizes the cost of perturbation. The left-hand side
of the constraint corresponds to the objective of (11). The requirement that it equals zero
ensures that the choice of $\boldsymbol{\xi}^i$ follows a decision path associated with decision cut-set $\bar{\mathbf{q}}^i$ that
misclassifies i (i.e., a path ending at prediction node n such that $\bar{w}_{n,y^i} = 0$). For a formal
justification of why it is sufficient to solve (12) at only decision cut-sets $\bar{\mathbf{q}}^i$, we refer to the
Electronic Companion EC.2.

Problem (12) can be decomposed and solved easily. Indeed, the constraint in (12)
implies that if $\bar{b}_{nf\theta} \bar{q}_{n,l(n)}^i > 0$, then $\mathbb{I}\left[x_f^i + \xi_f^i \le \theta\right] = 0$. Similarly, if $\bar{b}_{nf\theta} \bar{q}_{n,r(n)}^i > 0$,

then $\mathbb{I}\left[x_f^i + \xi_f^i \geq \theta + 1\right] = 0$. Thus, we can reformulate (12) as

$$
\begin{aligned}
\min_{\boldsymbol{\xi}^i \in \mathbb{Z}^{|\mathcal{F}|}} \quad & \sum_{f \in \mathcal{F}} \gamma_f^i |\xi_f^i| \\
\text{s.t.} \quad & \xi_f^i \geq \theta + 1 - x_f^i \quad \forall f \in \mathcal{F},\, \theta \in \Theta(f) : \bar{b}_{nf\theta} \vec{q}_{n,l(n)}^i > 0 \text{ for some } n \in \mathcal{N} \\
& \xi_f^i \leq \theta - x_f^i \qquad \forall f \in \mathcal{F},\, \theta \in \Theta(f) : \bar{b}_{nf\theta} \vec{q}_{n,r(n)}^i > 0 \text{ for some } n \in \mathcal{N}.
\end{aligned}
\tag{13}
$$

The constraints in this problem admit an intuitive interpretation. The first set of constraints states that, if at some branching node $n \in \mathcal{N}$ we branch on feature f at level θ ($\bar{b}_{nf\theta} = 1$) and the perturbed datapoint i is routed right ($\vec{q}_{n,l(n)}^i = 1$), then it must be that $\xi_f^i + x_f^i \geq \theta + 1$. The second set of constraints states that, if at some branching node $n \in \mathcal{N}$ we branch on feature f at level θ ($\bar{b}_{nf\theta} = 1$) and the perturbed datapoint i is routed left ($\vec{q}_{n,r(n)}^i = 1$), then it must be that $\xi_f^i + x_f^i \leq \theta$. Thus, the path followed by each datapoint (as determined by decision cut-set $\bar{\mathbf{q}}^i$) places constraints on the amount of perturbation for each feature of each datapoint, and the goal of problem (13) is to find the smallest cost perturbation that ensures that the perturbed datapoint i is routed through this path. Note that problem (13) may be infeasible for certain values of $\bar{\mathbf{b}}$ and $\bar{\mathbf{q}}^i$ if there exists a decision path in tree $(\bar{\mathbf{b}}, \bar{\mathbf{v}}, \bar{\mathbf{w}})$ that cannot be reached by any value of $\mathbf{x}$. In such cases, we work with the convention that the optimal value of (13) is ∞.

Solving problem (13) can be done efficiently, noting that the problem decomposes by datapoint and feature. For any fixed sample i and feature f, determine upper and lower bounds on ξ_f^i by iterating on the nodes of the decision path associated with decision cut-set $\bar{\mathbf{q}}^i$ and imposing either the first or second constraint on ξ_f^i depending on whether the datapoint is routed left or right; subsequently, determine the value of ξ_f^i from the resulting interval that minimizes $|\xi_f^i|$. Problem (13) is infeasible if no ξ_f^i satisfies the upper and lower bounds.

Let ψ^i be the minimum cost of (13) across all possible decision cut-sets that have value 0 (i.e., decision cut-sets associated with decision paths that misclassify i). Once $\{\psi^i\}_{i \in \mathcal{I}}$ has been obtained, we can impose the constraint in Ξ, see (3), which caps uncertainty to a budget ϵ. Recall that ψ^i denotes the smallest cost of perturbation to misclassify sample i. Thus, the worst-case set of samples to perturb can be obtained by sorting all non-zero ψ^i in non-decreasing order, performing the perturbations in this order until the budget ϵ is saturated. These perturbations define $(\tilde{\mathbf{q}}, \tilde{\boldsymbol{\xi}})$, where $\tilde{\mathbf{q}}$ is the collection of decision cut-sets for the decision paths of every sample after perturbation, and $\tilde{\boldsymbol{\xi}}$ is the collection of perturbations made (with

any unperturbed samples i after the budget is saturated having perturbation $\tilde{\boldsymbol{\xi}}^i = \mathbf{0}$). The solution $(\tilde{\mathbf{q}}, \tilde{\boldsymbol{\xi}})$ then defines the constraint

$$
\begin{aligned}
t \leq \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N} \cup \mathcal{L}} \tilde{q}_{n,t}^i w_{n,y^i} + \sum_{i \in \mathcal{I}} \tilde{q}_{s,1}^i + \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} \mathbb{I}\left[x_f^i + \tilde{\xi}_f^i \leq \theta \right] \tilde{q}_{n,l(n)}^i b_{nf\theta} \\
+ \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} \mathbb{I}\left[x_f^i + \tilde{\xi}_f^i \geq \theta + 1 \right] \tilde{q}_{n,r(n)}^i b_{nf\theta}
\end{aligned}
\tag{14}
$$

which is of the form (9b). The violated constraint (14) is then added to the main problem. In summary, solving problem (13) requires iterating over all decision paths and data samples, which takes time $\mathcal{O}(|\mathcal{L}||\mathcal{I}|)$, and sorting all ψ^i values requires $\mathcal{O}(|\mathcal{I}| \log |\mathcal{I}|)$ time. This results in an overall runtime of $\mathcal{O}(|\mathcal{L}||\mathcal{I}| + |\mathcal{I}| \log |\mathcal{I}|)$ for the subproblem. We summarize the procedure for the subproblem in algorithm 1.

3.3. Improving the Algorithm

A direct implementation of the aforementioned delayed constraint generation algorithm may be slow in practice: a prohibitive number of constraints (9b) may need to be added before convergence. To improve the performance of the method, we represent the number of correctly classified points, corresponding to the inner mimimization in (8), in an extended formulation. Such representations of nonlinear functions using additional variables often yield stronger relaxations (Tawarmalani and Sahinidis 2005, Vielma et al. 2017), since each linear cut in this lifted space translates to a nonlinear and more powerful cut in the original space.

In particular, let $t^i \in \{0, 1\}$ be a decision variable that indicates whether sample i is correctly classified for a given tree $(\mathbf{b}, \mathbf{v}, \mathbf{w})$ and worst-case perturbation $\tilde{\boldsymbol{\xi}}$ with corresponding minimum cut $\tilde{\mathbf{q}}$, and let $\mathbf{t}$ be the collection of t^i over $i \in \mathcal{I}$. Since t is the total number of correctly classified points, t is equal to $\sum_{i \in \mathcal{I}} t^i$. So, we reformulate problem (9) equivalently as

$$
\max_{\mathbf{b}, \mathbf{v}, \mathbf{w}, \mathbf{t}} \sum_{i \in \mathcal{I}} t^i \tag{15a}
$$

$$
\text{s.t.} \ \sum_{i \in \mathcal{I}} t^i \leq \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N} \cup \mathcal{L}} q_{n,t}^i w_{n,y^i} + \sum_{i \in \mathcal{I}} q_{s,1}^i + \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} \mathbb{I}\left[x_f^i + \xi_f^i \leq \theta \right] q_{n,l(n)}^i b_{nf\theta}
$$

$$
+ \sum_{i \in \mathcal{I}} \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} \mathbb{I}\left[x_f^i + \xi_f^i \geq \theta + 1 \right] q_{n,r(n)}^i b_{nf\theta} \qquad \forall \mathbf{q} \in \mathcal{Q}^{|\mathcal{I}|}, \boldsymbol{\xi} \in \Xi \tag{15b}
$$

$$
(\mathbf{b}, \mathbf{v}, \mathbf{w}) \in \mathcal{S} \tag{15c}
$$

$$
t^i \in \mathbb{R} \qquad \forall i \in \mathcal{I}. \tag{15d}
$$

Algorithm 1 Subproblem Procedure

Input: training set indexed by $\mathcal{I}$ with features $\mathcal{F}$ and labels $\mathcal{K}$, uncertainty set parameters $\boldsymbol{\gamma}$ and ϵ, tree $\pi := (\mathbf{b}, \mathbf{v}, \mathbf{w})$

Output: cut-sets $\tilde{\mathbf{q}}$ and perturbation $\tilde{\boldsymbol{\xi}}$

1: **for** $i \in \mathcal{I}$ **do**
2:    **Initialize:** $\boldsymbol{\xi}^i \leftarrow \mathbf{0}$ ▷ Store optimal perturbation
3:    **Initialize:** $\psi^i \leftarrow \infty$ ▷ Store minimum cost of perturbation
4:    **Initialize:** $\mathbf{q}^i \leftarrow \mathbf{0}$ ▷ Store optimal cut-set after perturbation
5:    **for** each decision path $\mathcal{P}$ **do**   ▷ $\mathcal{P}$ is a path from root node 1 to a prediction node
6:       $\hat{\mathbf{q}}^i \leftarrow$ decision cut-set associated with $\mathcal{P}$
7:       $\hat{\psi}^i \leftarrow$ optimal value of $(12_{\hat{\mathbf{q}}^i, \mathbf{b}})$, $\hat{\boldsymbol{\xi}}^i \leftarrow \arg\min\{(12_{\hat{\mathbf{q}}^i, \mathbf{b}})\}$ ▷ Find cost.
8:       **if** $(12_{\hat{\mathbf{q}}^i, \mathbf{b}})$ feasible and $\hat{\psi}^i \leq \psi^i$ **then**   ▷ If infeasible, choose another path
9:          $\boldsymbol{\xi}^i \leftarrow \hat{\boldsymbol{\xi}}^i$, $\psi^i \leftarrow \hat{\psi}^i$, $\mathbf{q}^i \leftarrow \hat{\mathbf{q}}^i$ ▷ Store current minimum cost perturbation
10:      **end if**
11:    **end for**
12: **end for**
13: **Initialize:** $C \leftarrow \epsilon$ ▷ C is the budget allotted for perturbation
14: **Initialize:** $\tilde{\boldsymbol{\xi}}^i \leftarrow \mathbf{0}$ ▷ Default perturbation is no perturbation.
15: **Initialize:** $\tilde{\mathbf{q}}^i \leftarrow \mathbf{q}^i$ ▷ Default cut-set is associated with the nominal path $\mathbf{q}^i$
16: **for** $i \in \mathcal{I}$ in non-decreasing order of ψ^i **do**
17:    **if** $C - \psi^i > 0$ **then**
18:       $\tilde{\boldsymbol{\xi}}^{\boldsymbol{i}} \leftarrow \boldsymbol{\xi}^i$, $\tilde{\mathbf{q}}^i \leftarrow \mathbf{q}^i$ ▷ Add perturbations if there is still budget left
19:       $C \leftarrow C - \psi^i$ ▷ Reduce cost from budget if perturbation of i included.
20:    **end if**
21: **end for**
22: **return** $(\tilde{\mathbf{q}}, \tilde{\boldsymbol{\xi}})$

Since formulation (15) is obtained from (9) through the substitution $t = \sum_{i \in \mathcal{I}} t^i$, both formulations have identical continuous relaxations. Our approach to further strengthen formulation (15) relies on one observation: for any datapoint $i \in \mathcal{I}$ and fixed tree $\pi \in \Pi_d$ such that $\pi(\mathbf{x}^i) \neq y^i$, there exists an optimal solution to the inner minimization problem in $(\mathcal{P}_\Xi)$ such that $\pi(\mathbf{x}^i + \boldsymbol{\xi}^i) \neq y^i$. This is because the inner minimization problem aims to misclassify the

most points, so no perturbation is needed on i to misclassify it, i.e., $\boldsymbol{\xi}^i = 0$ is part of the optimal solution.

It follows that we can enforce the condition that any misclassified point i in the nominal case cannot be correctly classified under a worst-case perturbation via the constraint

$$
\begin{aligned}
t^i \leq &\sum_{n \in \mathcal{N} \cup \mathcal{L}} q^i_{n,t} w_{n,y^i} + q^i_{s,1} + \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} \mathbb{I}\left[x^i_f \leq \theta\right] q^i_{n,l(n)} b_{nf\theta} \\
&+ \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} \mathbb{I}\left[x^i_f \geq \theta + 1\right] q^i_{n,r(n)} b_{nf\theta} \qquad \forall \mathbf{q}^i \in \mathcal{Q}.
\end{aligned}
\tag{16}
$$

A derivation of constraints (16) can be found in Aghaei et al. (2024). Intuitively, the constraints are analogous to constraints (14), but without summing over all datapoints in $\mathcal{I}$ and fixing the perturbation $\tilde{\boldsymbol{\xi}}$ to zero. Moreover, separation of inequalities (16) is similar to the separation of (14). Adding constraints (16) for each decision cut-set $\tilde{\mathbf{q}}^i$ associated with the decision path of each sample i without perturbation improves the solving times by an order of magnitude, which we detail further in Electronic Companion EC.3.

4. Calibration of Uncertainty Set Parameters

In this section, we propose a method for calibrating the parameters of the uncertainty set (3). For simplicity, we examine the cases where all features are unbounded integers (or where no bound is known). Datasets with bounded integer features, categorical features, and mixed feature types are discussed in the Electronic Companion EC.4.

We calibrate the parameters of (3) with an assumed probability of certainty in the nominal value of feature f and training sample i, denoted $\rho^i_f \in (0, 1]$. Without any other knowledge of the distribution shift, we follow the principle of maximum entropy (Jaynes 1957), which chooses the distribution of the perturbations with greatest entropy and thus highest uncertainty subject to our assumption of the probability of certainty ρ^i_f. To this end, we select the geometric distribution with parameter ρ^i_f as the distribution of the magnitude of perturbation $|\xi^i_f|$, and select the direction of perturbation $\text{sgn}(\xi^i_f)$ uniformly. That is, for $\zeta^i_f \in \mathbb{Z}$ the realization of the perturbation of x^i_f, the probability that x^i_f is perturbed by ξ^i_f is given by the symmetric geometric distribution

$$
\mathbb{P}(\xi^i_f = \zeta^i_f) := (0.5)^{\mathbb{I}[|\zeta^i_f| > 0]} \rho^i_f (1 - \rho^i_f)^{|\zeta^i_f|}
\tag{17}
$$

and the magnitude of perturbation follows a geometric distribution such that

$$
\mathbb{P}(|\xi^i_f| = |\zeta^i_f|) = \rho^i_f (1 - \rho^i_f)^{|\zeta^i_f|}.
\tag{18}
$$

We now follow the idea of building uncertainty sets using hypothesis testing as in Bertsimas et al. (2018). We set up a likelihood ratio test on the magnitude of the perturbation with threshold $\lambda^{|\mathcal{I}|}$ for $\lambda \in (0, 1]$, where we add the exponent $|\mathcal{I}|$ to normalize across different datasets with varying number of training samples. Our null hypothesis is that the magnitude of a given perturbation $\boldsymbol{\xi}$ of our dataset comes from the distribution described by (18). If this null hypothesis fails to be rejected, then $\boldsymbol{\xi}$ lies within our uncertainty set. Hence, $\boldsymbol{\xi}$ lies within our uncertainty set if it satisfies the constraint

$$\frac{\prod_{i \in \mathcal{I}} \prod_{f \in \mathcal{F}} \rho_f^i (1 - \rho_f^i)^{|\xi_f^i|}}{\prod_{i \in \mathcal{I}} \prod_{f \in \mathcal{F}} \rho_f^i} \geq \lambda^{|\mathcal{I}|}. \tag{19}$$

The numerator of the left hand side of (19) is the likelihood under distribution (17) of a perturbation of magnitude $|\xi_f^i|$, and thus the likelihood under the null hypothesis. The denominator of the left hand side is the likelihood of the most probable realization under (18) (i.e., the likelihood of no perturbation where all $\xi_f^i = 0$). The test (19) can be equivalently represented as

$$\sum_{i \in \mathcal{I}} \sum_{f \in \mathcal{F}} |\xi_f^i| \log \left(\frac{1}{1 - \rho_f^i} \right) \leq -|\mathcal{I}| \log \lambda. \tag{20}$$

Note that (20) is of the form of the constraint in uncertainty set (3) with parameters $\gamma_f^i = \log \left(\frac{1}{1 - \rho_f^i} \right)$ and $\epsilon = -|\mathcal{I}| \log \lambda$.

Therefore, a method of tuning the parameters of (3) is to use the probability of certainty ρ_f^i for each feature of each sample, which can be derived from domain knowledge. The value of λ is used to tune the size of the uncertainty set, and hence the level of robustness. For $\lambda = 1$, the budget of uncertainty ϵ is 0, meaning that no distribution shift occurs and our formulation is equivalent to the formulation from Aghaei et al. (2024). As λ decreases, more and more perturbations of the data are considered, and thus the model is robust to more scenarios. Hence, λ is a parameter that can be tuned to adjust the level of robustness, which can be chosen by methods such as cross validation. Lastly, we note that other distributions can be used based on domain knowledge of the possible distribution shift, and that the geometric distribution (18) is a generic choice given limited domain knowledge.

5. Comparison to State-of-the-Art Methods

In this section, we compare the uncertainty sets and notions of robustness used in state-of-the-art models from the literature and in our model for learning robust classification trees.

In particular, we examine the approach of Bertsimas et al. (2019), which also utilizes an MIO-based model but with a different uncertainty set and concept of robustness. We also compare and contrast to Chen et al. (2019) and Vos and Verwer (2021) who also use a different uncertainty set and employ a heuristic approach, but who use a similar notion of robustness to ours.

5.1. Uncertainty Sets

The model of uncertainty (3) that we employ differs in several regards from that in previous works on robust decision trees. Indeed, Bertsimas et al. (2019), Chen et al. (2019), and Vos and Verwer (2021) employ a *row-wise uncertainty set* based on an p-norm, of the form

$$\Xi_p = \left\{ \boldsymbol{\xi} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{F}|} \; : \; \|\boldsymbol{\xi}^i\|_p \le \epsilon \quad \forall i \in \mathcal{I} \right\}, \tag{21}$$

where $\epsilon \ge 0$ is a user selected budget of uncertainty parameter. Chen et al. (2019) and Vos and Verwer (2021) model uncertainty with Ξ_∞ specifically. On the other hand, Bertsimas et al. (2019) uses the uncertainty set (21) with any choice of p-norm. We note that in Bertsimas et al. (2019), due to the notion of robustness employed, the robust counterpart ends up taking the same form independently of the choice of p, resulting in the same sets of optimal trees independently of the norm used in the uncertainty set.

We now argue that the uncertainty set in (21) does not model distribution shifts in the datasets and applications that motivate us. First, the perturbation $\boldsymbol{\xi}$ is not integral, and thus the realization of the covariates may not be realistic if the covariates are integer or categorical. In addition, the set Ξ_p imposes the strong assumption that the distribution shift is rectangular across samples, resulting in an overly conservative model where the perturbations associated with all datapoints can all simultaneously take on their worst-case values. Lastly, the set (21) assumes the same cost of shift (represented by γ_f^i in our model) across all datapoints $i \in \mathcal{I}$ and features $f \in \mathcal{F}$, implying that the magnitude and direction of distribution shifts is constant for all samples and features.

In contrast, our proposed uncertainty set (3) fixes the aforementioned issues of uncertainty set (21) by restricting the perturbations to be integer, having a single budget of uncertainty shared among the data samples, and introducing costs of perturbation that can differ for each feature and sample. Thus, our proposed model of uncertainty is more flexible, being able to capture shifts in integer covariates and, as discussed in section EC.4, extending to the cases of one-sided shifts and categorical features.

5.2. Notions of Robustness

In our problem, the tree structure must be decided before the perturbation of covariates is realized, and only *after* this realization is observed can we decide if a given datapoint is correctly classified or not. This is similar to the frameworks of Chen et al. (2019) and Vos and Verwer (2021), who calculate their objective based on an adversarial perturbation of the data. But unlike our approach of optimizing accuracy over the whole tree, these methods use either information gain or Gini impurity as an objective at each node where a test is performed. These approaches, therefore, cannot guarantee optimal worst-case accuracy.

On the other hand, Bertsimas et al. (2019) does have accuracy as an objective, but uses a different notion of robustness. Their model postulates that the robust trees created must maintain the *same* predictions for each data sample across *all possible perturbations of the data*. Mathematically, the problem solved by Bertsimas et al. (2019) is equivalent to

$$\max_{\pi \in \Pi_d} \ \sum_{i \in \mathcal{I}} \mathbb{I}\left[\pi(\mathbf{x}^i) = y^i \right] \tag{22_Ξ}$$
$$\text{s.t.} \ \ \pi(\mathbf{x}^i) = \pi(\mathbf{x}^i + \boldsymbol{\xi}^i) \quad \forall i \in \mathcal{I}, \ \boldsymbol{\xi} \in \Xi.$$

It follows from the constraints of (22_Ξ) that it is equivalent to

$$\max_{\pi \in \Pi_d} \ \min_{\boldsymbol{\xi} \in \Xi} \ \sum_{i \in \mathcal{I}} \mathbb{I}\left[\pi(\mathbf{x}^i + \boldsymbol{\xi}^i) = y^i \right] \tag{23_Ξ}$$
$$\text{s.t.} \ \ \pi(\mathbf{x}^i) = \pi(\mathbf{x}^i + \boldsymbol{\xi}^i) \quad \forall i \in \mathcal{I}, \ \boldsymbol{\xi} \in \Xi.$$

Note that (23_Ξ) has the same objective function as ($\mathcal{P}_\Xi$). We formalize the relationship between problems ($\mathcal{P}_\Xi$) and (23_Ξ) in the following proposition.

PROPOSITION 1. *Given an uncertainty set Ξ, all optimal solutions of (23_Ξ) are feasible in ($\mathcal{P}_\Xi$), but all optimal solutions to ($\mathcal{P}_\Xi$) may be infeasible in (23_Ξ).*

Proof. Note that problem (23_Ξ) is problem ($\mathcal{P}_\Xi$) with the additional constraints in (23_Ξ). Therefore, problem (23_Ξ) has a feasible region that is a subset of the feasible region of problem ($\mathcal{P}_\Xi$). Hence, any optimal solution of (23_Ξ) is a feasible solution of ($\mathcal{P}_\Xi$).

We now present an example where all optimal solutions to ($\mathcal{P}_\Xi$) are infeasible in (23_Ξ). Consider the dataset with a single feature and nine datapoints given in table 1, and assume that the uncertainty set is

$$\Xi = \left\{ \boldsymbol{\xi} \in \mathbb{R}^{|\mathcal{I}| \times |\mathcal{F}|} \ : \ \|\boldsymbol{\xi}^i\|_\infty \leq 1 \quad \forall i \in \mathcal{I} \right\}.$$

x	1	2	3	4	5	6	7	8	9
y	0	0	0	0	1	1	1	1	1

Table 1   Example data.

The only optimal solution of $(\mathcal{P}_\Xi)$ is given by the function $\pi^\star(x) = \mathbb{I}(x \geq 6)$, which in the worst case misclassifies the two datapoints with feature values $x = 5$ and $x = 6$ However, $\pi^\star(x)$ is not a feasible solution to (23_Ξ), as the fifth sample has $\pi^\star(5) = 0$, but with perturbation $\xi^5 = 1$, $\pi^\star(5 + 1) = 1$. This violates the constraints in (23_Ξ), and thus $\pi^\star(x) = \mathbb{I}(x \geq 6)$ is not a feasible solution to (23_Ξ). Thus, this example shows that all optimal solutions to $(\mathcal{P}_\Xi)$ may be infeasible in (23_Ξ).   □

In the example from proposition (1), the optimal solution of (23_Ξ), and thus (22_Ξ), is given by $\bar{\pi}(x) = \mathbb{I}(x \geq 2)$, resulting in the missclassification of the three datapoints with $x \in \{3, 4, 5\}$. We see that the two problems yield different optimal solutions, and that $\bar{\pi}(x)$ is a more conservative solution that misclassifies more samples than $\pi^\star(x)$.

Because problems (22_Ξ) and (23_Ξ) are equivalent, by proposition 1, problem $(\mathcal{P}_\Xi)$ considers no less than the set of robust solutions that (22_Ξ) considers. Thus, problem $(\mathcal{P}_\Xi)$ may lead to potentially less conservative solutions in comparison to (22_Ξ), resulting in better performance as the example in the proof of proposition 1 shows: $(\mathcal{P}_\Xi)$ considers additional trees with more branching decisions and fewer misclassifications in both the nominal and worst cases. Indeed as uncertainty ϵ grows, the only feasible tree for (22_Ξ) that can ensure that data samples always land at the same leaf no matter how the data is perturbed is a tree of depth zero (with no branching node).

6. Experiments

In this section, we evaluate our method on various datasets and across uncertainty set parameters. We assess the effect of robustness on accuracy under shifts, accuracy under no shifts (a.k.a. price of robustness) (Bertsimas and Sim 2004), sparsity, and computation times of our approach.

We refer to the Electronic Companion EC.5 for an empirical comparison of our method to the method of Bertsimas et al. (2019) for learning robust trees, where we show that our method overall outperforms the method of Bertsimas et al. (2019) in terms of worst-case and average-case accuracy and confirm empirically the theoretical observations made in section 5.2.

6.1. Setup and Instances

We conduct all our experiments in Python 3.6 with Gurobi 9.0.1 as our MIO solver. All problems are solved on a single core of an Intel Xeon Processor 2640v4 and 4GB of memory with a time limit of 7200 seconds. For instances that do not solve to optimality within the time limit, we return the best feasible tree found within the time limit and record the corresponding optimality gap reported by the solver.

6.1.1. Instances Each instance of our experiments consists of a choice of uncertainty set parameters, a maximum depth of tree, and a dataset. For each instance's uncertainty set, we utilize the hypothesis testing framework as described in section (4): for each $f \in \mathcal{F}$, we choose the probability of certainty ρ^i_f by sampling from a normal distribution with a standard deviation of 0.2, keeping ρ^i_f the same across all $i \in \mathcal{I}$ for each f, and assuming no bounds on any non-categorical features. Note that if the number sampled from the normal distribution is greater than 1 (resp. less than 0), then the value of ρ^i_f is set to 1 (resp. 0). For the means of this normal distribution, we create instances with means of 0.6, 0.7, 0.8, and 0.9. An instance is also created for different budgets of uncertainty ϵ by setting λ to be 0.5, 0.75, 0.85, 0.9, 0.95, 0.97, and 0.99. For every dataset and uncertainty set, we test with tree depths of 2, 3, 4, and 5.

We evaluate on 12 publicly available datasets as listed in Table 2. Each dataset contains either integer-valued data, categorical data, or a mixture of both, as detailed in the table. We preprocess datasets with categorical data by one-hot encoding each categorical feature. The number of samples in the datasets range from 124 to 3196 and the number of features from 4 to 36. For each dataset, we randomly split it into 80% training data and 20% testing data once. In total, we experiment on 1536 instances (128 per dataset and 192 per value of λ) each with different uncertainty sets and maximum depths of tree.

6.1.2. Generating Shifted Test Sets To test our method's robustness against distribution shifts, we generate 5,000 different perturbed test sets for each instance. To create each perturbed test set, we independently perturb the original test data based on expected perturbations. That is, for the collection of q^i_f values for every $f \in \mathcal{F}$ used to construct an uncertainty set based off of (20), we perturb each test set based on the symmetric geometric distribution described in (17).

In order to measure the robustness of our method against unexpected shifts of the data, we repeat the same process of generating shifted test sets for each instance but with *unexpected*

Name	Samples	Ints.	Cats.	$|\mathcal{F}|$	Range of Ints.	Num. Cat. Values
balance-scale	625	4	0	4	5	N/A
breast-cancer	277	0	9	38	N/A	1-11
car-evaluation	1728	5	1	6	4-55	1
hayes-roth	132	4	0	4	3-4	N/A
house-votes-84	232	0	16	16	N/A	1
kr-vs-kp	3196	0	36	38	N/A	1-3
monk1	124	6	0	6	2-4	N/A
monk2	169	6	0	6	2-4	N/A
monk3	122	6	0	6	2-4	N/A
soybean-small	47	35	0	35	2-7	N/A
spect	267	0	23	22	N/A	1
tic-tac-toe	958	0	9	27	N/A	3

Table 2   Datasets used from the UCI Machine Learning Repository (Dua and Graff 2017). Columns from left to right: name of dataset, number of samples, number of integer features, number of categorical features, value of $|\mathcal{F}|$, range of integer features (if applicable), and number of categorical values (if applicable).

perturbations: perturbing the test data using values of ρ_f^i different than what we gave our model. We shift each ρ_f^i value down 0.2, then perturb our test data in 5,000 different ways based on these new values of ρ_f^i. We do the same procedure but with ρ_f^i shifted down by 0.1 and up by 0.1. In a similar fashion, we also uniformly sample a new ρ_f^i value for each feature in a neighborhood of radius 0.05 of the original expected ρ_f^i value, and perturb the test data in 5,000 different ways with the new ρ_f^i values. We do the same procedure for the radii of 0.1, 0.15, and 0.2 for the neighborhoods.

Note that we also add a small regularization parameter to our robust model to discourage redundant branching on features and thresholds. Specifically, the objective function in problem (15) is slightly modified to:

$$\sum_{i \in \mathcal{I}} t_i - \frac{1}{2^d} \sum_{n \in \mathcal{N}} \sum_{f \in \mathcal{F}} \sum_{\theta \in \Theta(f)} b_{nf\theta}.$$

As the regularization term is always smaller than 1, the accuracy under worst-case distribution shift is the same between the above problem and problem (15), so a tree optimal in the above problem is also optimal in (15). We do not add any more regularization to our robust

model otherwise, as we will empirically show that adding robustness itself has a regularizing (sparsity-promoting) effect in section 6.2.3.

6.1.3. Learning Non-Robust Trees For comparison, we use our method to create a non-robust optimal tree by setting the budget of uncertainty to 0 (i.e., $\lambda = 1$), and tune a regularization parameter for the non-robust tree using a validation set from the training set. The regularization term penalizes the number of branching nodes to yield the modified objective in problem (15):

$$\max_{\mathbf{b},\mathbf{w},\mathbf{t}} R\sum_{i\in\mathcal{I}} t_i - (1-R)\sum_{i\in\mathcal{I}}\sum_{n\in\mathcal{N}}\sum_{\Theta(f)} b_{nf\theta},$$

where $R \in [0,1]$ is the tuned regularization parameter. To tune the value R, we first randomly hold out 20% of the training set into a validation set. Then, we select various values of R from the set $\{0.6, 0.7, 0.8, 0.9, 0.95, 1\}$ and learn a tree using our method with the uncertainty set $\{\mathbf{0}\}$ and the same specifications as learning our robust tree but instead with a time limit of 1 hour. We find the accuracy of the learned tree on the held-out validation set, and select R with the best accuracy. We then create the non-robust optimal tree using the entire training set and tuned R value.

6.2. Effect of Robustness

We now examine our method's robustness to distribution shifts across different uncertainty sets in comparison to a non-robust model.

6.2.1. Accuracy Under Distribution Shifts For each instance, we measure the worst-case (lowest) and average accuracy across all perturbed test sets. Our results are summarized in Figure 4, which shows across all instances the distributions of the difference in worst-case and average accuracies between the learned robust and non-robust trees. From the figure, we see that the robust model in general has higher worst-case and average-case accuracies than the non-robust model when there are distribution shifts in the data. We also see that there is a range of values of λ (namely between 0.75 and 0.9) that perform well over other values of λ in terms of both worst-case and average case accuracy. This shows us that if the budget of uncertainty is too small, then we do not allow enough room to hedge against distribution shifts in our uncertainty set. But if the budget of uncertainty is too large, then we become over-conservative and lose some accuracy in both expected and unexpected perturbations
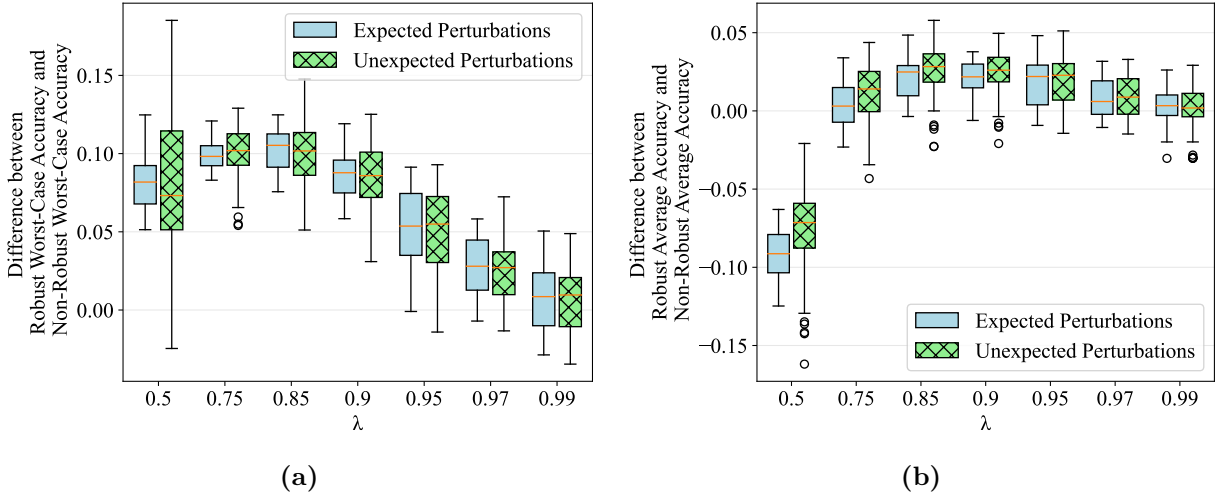
Figure 4   Companion figure to section 6.2.1. Distributions across problem instances of the gain in (a) worst-case accuracy and (b) average accuracy from using a robust tree versus a non-robust, regularized tree in dependence of λ under both expected and unexpected perturbations in the test data. The parameter λ varies between 0.5 (more robust) and 0.99 (less robust).

of our test data. We also see that there is little difference between the gains in accuracy in instances where the perturbation of our data is as we expect versus when the perturbation is not as we expect. This indicates that even if we misspecify our model, we still obtain a classification tree robust to distribution shifts within a reasonable range of the expected shift. Overall, we see that an important factor in determining the performance of our model is the budget of uncertainty, which can be conveniently tuned to create an effective robust tree.

6.2.2. Price of Robustness   We also measure the decrease in accuracy from using a robust tree versus a non-robust tree under the case of no distribution shift in the test set, i.e., the price of robustness (Bertsimas and Sim 2004), and summarize this metric in Figure 5a. From the figure, we observe that in the range of λ values that perform well in terms of accuracy (i.e., between 0.75 and 0.9, see section 6.2.1), we have an average price of robustness between 0.05 and 0.1. As the level of robustness increases (with a decreasing λ), the higher the price of robustness. This is expected with our model because a larger budget of uncertainty leads to larger deviations away from the nominal test set considered, and thus the worst-case distribution shift may look very different from the nominal test set if more drastic shifts are considered.

6.2.3. The Effect of Robustness on Sparsity   To evaluate the sparsity of our model in relation to the robustness of the tree, we measure the number of branching nodes in each

(a)   **(b)**
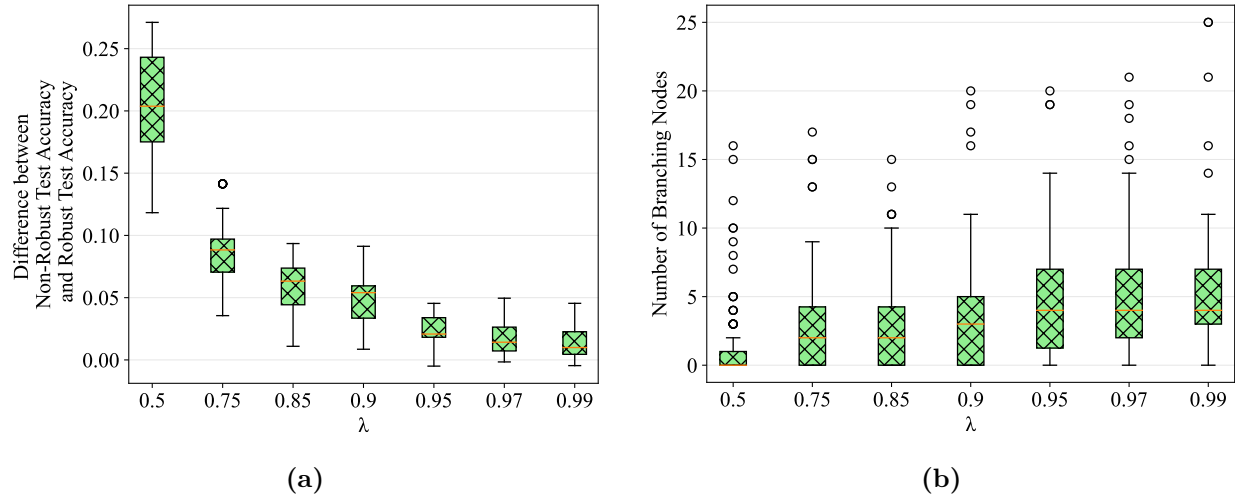
Figure 5   (a) Companion figure to section 6.2.2. Distribution across problem instances of the difference in nominal test accuracy between a non-robust, regularized tree and a robust tree in dependence of λ. The parameter λ varies between 0.5 (more robust) and 0.99 (less robust). (b) Companion figure to section 6.2.3. Distributions of the number of branching nodes across all instances with the same λ. The parameter λ varies between 0.5 (more robust) and 0.99 (less robust).

instance. Our results are summarized in Figure 5b. Overall, as the size of the uncertainty set increases with a smaller λ, the number of branching nodes decreases. Namely, the median number of branching nodes is 4 for instances with $\lambda \in \{0.95, 0.97, 0.99\}$, 3 for instances with $\lambda = 0.9$, 2 for instances with $\lambda \in \{0.75, 0.85\}$, and 0 for instances with $\lambda = 0.5$. This regularizing behavior of robustness is expected with our model: with more branching nodes in the learned model, the more opportunities there are for a low-cost perturbation of each sample in a given tree, yielding a lower worst-case accuracy. Thus, as the number of perturbations allowed expand, the number of branching nodes in the learned model decreases to yield a more favorable worst-case accuracy.

6.2.4. Computational Times   We summarize the computation times across all instances in Figure 6. For any fixed λ, the variations in computation times and optimality gaps across instances are due to differences in the maximum depth of tree, the number of data samples, the number of features, and the range of values for each feature within each instance. In general, a larger uncertainty set (smaller λ) leads to a larger optimality gap for a fixed time limit, as there are more constraints to add to the master problem to reach optimality. In particular, the number of instances that have an optimality gap larger than 50% are 72, 32, 7, and 0 for $\lambda = 0.5$, 0.85, 0.95, and 1, respectively. We also observe that there is a large
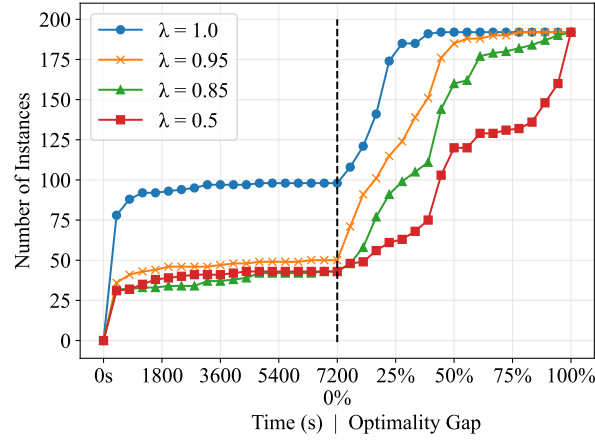
Figure 6   Companion figure to section 6.2.4. Cumulative number of problems solved to optimality for several values of λ on the 'Time' axis, and cumulative number of problems with MIO optimality gap no larger than the given value on the 'Optimality Gap' axis. The case of $\lambda = 1.0$ corresponds to the non-robust, regularized tree.

gap in number of instances solved fully to optimality within the time limit between the non-robust instances and the robust instances, with 48 more instances solved to optimality in the non-robust case in comparison to the robust case with $\lambda = 0.95$. In addition, the number of robust instances able to be solved to optimality within the time limit are fairly close, with only 7 instances between the $\lambda = 0.95$ case and the $\lambda = 0.85$ and $\lambda = 0.5$ cases.

7. Acknowledgments

References

Aghaei S, Azizi MJ, Vayanos P (2019) Learning optimal and fair decision trees for non-discriminative decision-making. *Proceedings of the AAAI Conference on Artificial Intelligence* 33:1418–1426, ISSN 2374-3468, URL http://dx.doi.org/10.1609/aaai.v33i01.33011418.

Aghaei S, Gómez A, Vayanos P (2024) Strong optimal classification trees. *Operations Research* URL http://dx.doi.org/10.1287/opre.2021.0034.

Atamtürk A, Gómez A (2017) Maximizing a class of utility functions over the vertices of a polytope. *Operations Research* 65(2):433–445.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton University Press), ISBN 9781400831050, URL http://dx.doi.org/10.1515/9781400831050.

Bertsimas D, Dunn J (2017) Optimal classification trees. *Machine Learning* 106(7):1039–1082, ISSN 15730565, URL http://dx.doi.org/10.1007/s10994-017-5633-9.

Bertsimas D, Dunn J, Pawlowski C, Zhuo YD (2019) Robust classification. *INFORMS Journal on Optimization* 1(1):2–34, ISSN 2575-1484, URL http://dx.doi.org/10.1287/ijoo.2018.0001.

Bertsimas D, Gupta V, Kallus N (2018) Data-driven robust optimization. *Mathematical Programming* 167:235–292.

Bertsimas D, Sim M (2003) Robust discrete optimization and network flows. *Mathematical Programming* 98(1):49–71.

Bertsimas D, Sim M (2004) The price of robustness. *Operations Research* 52(1):35–53.

Bickel S, Brückner M, Scheffer T (2007) Discriminative learning for differing training and test distributions. *Proceedings of the 24th International Conference on Machine Learning - ICML '07* (New York, New York, USA: ACM Press), ISBN 9781595937933, URL http://dx.doi.org/10.1145/1273496.1273507.

Borrero JS, Lozano L (2021) Modeling defender-attacker problems as robust linear programs with mixed-integer uncertainty sets. *INFORMS Journal on Computing* 33(4):1570–1589.

Boutilier JJ, Michini C, Zhou Z (2022) Shattering inequalities for learning optimal decision trees. Schaus P, ed., *Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, 74–90 (Cham: Springer International Publishing), ISBN 978-3-031-08011-1.

Breiman L, Friedman JH, Olshen RA, Stone CJ (2017) *Classification And Regression Trees* (Routledge), ISBN 9781315139470, URL http://dx.doi.org/10.1201/9781315139470.

Chen H, Zhang H, Boning D, Hsieh CJ (2019) Robust decision trees against adversarial examples. *International Conference on Machine Learning*, 1122–1131 (PMLR).

Chen R, Paschalidis IC (2020) Distributionally robust learning. *Foundations and Trends®️ in Optimization* 4(1-2):1–243, ISSN 2167-3888, URL http://dx.doi.org/10.1561/2400000026.

Dua D, Graff C (2017) UCI Machine Learning Repository. URL http://archive.ics.uci.edu/ml.

Elmachtoub A, Liang JCN, Mcnellis R (2020) Decision trees for decision-making under the predict-then-optimize framework. III HD, Singh A, eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 2858–2867 (PMLR), URL https://proceedings.mlr.press/v119/elmachtoub20a.html.

Gao R, Kleywegt A (2023) Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research* 48(2):603–655.

Gurobi Optimization, LLC (2024) Gurobi optimizer reference manual. URL https://www.gurobi.com.

Hu X, Rudin C, Seltzer M (2019) Optimal sparse decision trees. *Advances in Neural Information Processing Systems* 32.

Jaynes ET (1957) Information theory and statistical mechanics. *Physical Review* 106(4):620.

Jo N, Aghaei S, Benson J, Gómez A, Vayanos P (2023) Learning optimal fair decision trees: Trade-offs between interpretability, fairness, and accuracy. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 181–192.

Jo N, Aghaei S, Gómez A, Vayanos P (2021) Learning optimal prescriptive trees from observational data. *arXiv preprint arXiv:2108.13628* URL http://dx.doi.org/10.48550/arXiv.2108.13628.

Justin N, Aghaei S, Gomez A, Vayanos P (2022) Optimal robust classification trees. *The AAAI-22 Workshop on Adversarial Machine Learning and Beyond*, URL https://openreview.net/forum?id=HbasA9ysA3.

Kallus N (2017) Recursive partitioning for personalization using observational data. *International Conference on Machine Learning*, 1789–1798 (PMLR).

Kass GV (1980) An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 29(2):119–127.

Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. *Operations Research & Management Science in the Age of Analytics*, 130–166 (INFORMS), URL http://dx.doi.org/10.1287/educ.2019.0198.

Laporte G, Louveaux FV (1993) The integer L-shaped method for stochastic integer programs with complete recourse. *Operations Research Letters* 13(3):133–142.

Mehmanchi E, Gillen CP, Gómez A, Prokopyev OA (2020) On robust fractional 0-1 programming. *INFORMS Journal on Optimization* 2(2):96–133.

Mišić VV (2020) Optimization of tree ensembles. *Operations Research* 68(5):1605–1624, ISSN 0030-364X, URL http://dx.doi.org/10.1287/opre.2019.1928.

Mutapcic A, Boyd S (2009) Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods & Software* 24(3):381–406.

Pant R, Trafalis TB, Barker K (2011) Support vector machine classification of uncertain and imbalanced data using robust optimization. *Proceedings of the 15th WSEAS International Conference on Computers*, 369–374 (World Scientific and Engineering Academy and Society, Stevens Point, Wisconsin, USA).

Quinlan JR (1986) Induction of decision trees. *Machine Learning* 1(1):81–106.

Quinlan JR (2014) *C4. 5: Programs for Machine Learning* (Elsevier).

Quiñonero-Candela J, Sugiyama M, Lawrence ND, Schwaighofer A (2009) *Dataset Shift in Machine Learning* (MIT Press).

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5), ISSN 2522-5839, URL http://dx.doi.org/10.1038/s42256-019-0048-x.

Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C (2022) Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys* 16:1–85.

Shafieezadeh Abadeh S, Mohajerin Esfahani PM, Kuhn D (2015) Distributionally robust logistic regression. Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R, eds., *Advances in Neural Information Processing Systems*, volume 28 (Curran Associates, Inc.), URL https://proceedings.neurips.cc/paper/2015/file/cc1aa436277138f61cda703991069eaf-Paper.pdf.

Shaham U, Yamada Y, Negahban S (2018) Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing* 307:195–204, ISSN 0925-2312, URL `http://dx.doi.org/10.1016/j.neucom.2018.04.027`.

Shimodaira H (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* 90(2), ISSN 03783758, URL `http://dx.doi.org/10.1016/S0378-3758(00)00115-4`.

Shivaswamy PK, Bhattacharyya C, Smola AJ (2006) Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research* 7(47):1283–1314, URL `http://jmlr.org/papers/v7/shivaswamy06a.html`.

Sinha A, Namkoong H, Volpi R, Duchi J (2017) Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571* URL `http://dx.doi.org/10.48550/arXiv.1710.10571`.

Tawarmalani M, Sahinidis NV (2005) A polyhedral branch-and-cut approach to global optimization. *Mathematical Programming* 103(2):225–249.

Verwer S, Zhang Y (2019) Learning optimal classification trees using a binary linear program formulation. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, ISSN 2374-3468, URL `http://dx.doi.org/10.1609/aaai.v33i01.33011624`.

Vielma JP, Dunning I, Huchette J, Lubin M (2017) Extended formulations in mixed integer conic quadratic programming. *Mathematical Programming Computation* 9(3):369–418.

Vos D, Verwer S (2021) Efficient training of robust decision trees against adversarial examples. *International Conference on Machine Learning*, 10586–10595 (PMLR).

Vossler P, Aghaei S, Justin N, Jo N, Gómez A, Vayanos P (2023) Odtlearn: A package for learning optimal decision trees for prediction and prescription. *arXiv preprint arXiv:2307.15691* URL `http://dx.doi.org/10.48550/arXiv.2307.15691`.

Zadrozny B (2004) Learning and evaluating classifiers under sample selection bias. *Proceedings of the Twenty-First International Conference on Machine Learning*, 114.

Zeng B, Zhao L (2013) Solving two-stage robust optimization problems using a column-and-constraint generation method. *Operations Research Letters* 41(5):457–461.

E-Companion

EC.1. Connection to Wasserstein Distributionally Robust Optimization

In this section, we present a Wasserstein distributionally robust formulation for learning optimal classification trees robust to distribution shifts, and discuss the relationship between this model and the robust optimization model that we study in this paper. For general details on the mathematical relationship between distributionally robust optimization and robust optimization, we refer to corollary 2(iii) in Gao and Kleywegt (2023).

We first define the Wasserstein distance for data with discrete features. Let X be a random vector from the space of covariates $\mathcal{X} \subseteq \mathbb{Z}^{|\mathcal{F}|}$, and Y be a random variable denoting the label from the set $\mathcal{K}$. Let the set of all probability measures for (X, Y) be in set $\mathcal{P}$, and we will say that the training data is drawn from a distribution $\mathbb{P}^0 \in \mathcal{P}$. The empirical distribution of our covariates is defined as

$$\hat{\mathbb{P}}_X(\mathbf{x}) := \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{I}\left[\mathbf{x} = \mathbf{x}^i\right].$$

Let $\mathbb{Q} \in \mathcal{P}$ be the realization of the distribution of data after distribution shift, and $\mathbb{Q}_X$ denote the marginal distribution of $\mathbb{Q}$ on X.

To measure the distance between $\mathbb{Q}_X$ and the empirical distribution $\hat{\mathbb{P}}_X$, we use the Wasserstein distance, which over a discrete space is defined as

$$W(\hat{\mathbb{P}}_X, \mathbb{Q}_X) := \inf_{\mathbb{M} \in \mathcal{P}_X^2} \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{x}' \in \mathcal{X}} d(\mathbf{x}, \mathbf{x}') \mathbb{M}(\mathbf{x}, \mathbf{x}')$$

$$\text{s.t. } \sum_{\mathbf{x}' \in \mathcal{X}} \mathbb{M}(\mathbf{x}', \mathbf{x}) = \mathbb{Q}_X(\mathbf{x})$$

$$\sum_{\mathbf{x}' \in \mathcal{X}} \mathbb{M}(\mathbf{x}, \mathbf{x}') = \hat{\mathbb{P}}_X(\mathbf{x})$$

for a distance metric d on $\mathcal{X}$ and $\mathcal{P}_X^2$ the set of all joint probability measures for (X, X). To create an analogous ambiguity set as the uncertainty set defined in (3), we define the following distance metric between points indexed in the set $\mathcal{I}$ and any $\bar{\mathbf{x}} \in \mathcal{X}$:

$$d(\mathbf{x}^i, \bar{\mathbf{x}}) := \sum_{f \in \mathcal{F}} \gamma_f^i \left|x_f^i - \bar{x}_f\right|.$$

Then, letting $\mathbb{Q}_X^i(\mathbf{x})$ be the distribution of $\mathbb{M}$ conditioned on $\mathbf{x}^i$, we can represent our Wasserstein distance as

$$W(\hat{\mathbb{P}}_X, \mathbb{Q}_X) = \sum_{i \in \mathcal{I}} \sum_{\bar{\mathbf{x}} \in \mathcal{X}} \sum_{f \in \mathcal{F}} \gamma_f^i \left|x_f^i - \bar{x}_f\right| \mathbb{Q}_X^i(\bar{\mathbf{x}}). \tag{EC.1}$$

Recall that we are interested in the setting where there is only a shift in the covariates. Thus, we impose the following assumption on the distribution shifts we consider:

ASSUMPTION EC.1. *Let the observed data on individual $i \in \mathcal{I}$ be $(\hat{X}^i, \hat{Y}^i) \sim \mathbb{P}^0$, where individual i in the unknown shifted distribution has data $(X^i, Y^i) \sim \mathbb{Q}$ for unknown $\mathbb{Q} \in \mathcal{P}(\Omega)$. We assume that the shifted distribution $\mathbb{Q}$ has the property $Y^i = \hat{Y}^i$ for all $i \in \mathcal{I}$.*

Using this assumption and the definition of the Wasserstein distance defined in (EC.1), we then restrict the set of all possible distributions via a Wasserstein ball with radius $\epsilon \geq 0$:

$$\mathcal{B}(\epsilon) := \{\mathbb{Q} \in \mathcal{P} \ : \ W(\mathbb{Q}_X, \hat{\mathbb{P}}_X) \leq \epsilon, \ \mathbb{Q} \text{ satisfies Assumption EC.1}\}. \tag{EC.2}$$

Then, the analogous distributionally robust problem to $(\mathcal{P}_\Xi)$ is

$$\max_{\pi \in \Pi_d} \inf_{\mathbb{Q} \in \mathcal{B}(\epsilon)} |\mathcal{I}| \mathbb{E}_\mathbb{Q} \left[\mathbb{I} \left[\pi(X) = Y \right] \right], \tag{EC.3}$$

where we multiply the objective by $|\mathcal{I}|$ to make the objective the expected number of training samples that are correctly classified, making the objective similar to $(\mathcal{P}_\Xi)$.

Consider adding a constraint in $\mathcal{B}(\epsilon)$ that $\mathbb{Q}_X^i(\mathbf{x}) \in \{0, 1\}$ for all $\mathbf{x} \in \mathcal{X}$. For any $\mathbb{Q} \in \mathcal{B}(\epsilon)$ that satisfies this constraint, there exists exactly one $\bar{\mathbf{x}}^i \in \mathcal{X}$ such that $\mathbb{Q}_X^i(\bar{\mathbf{x}}^i) = 1$ for each $i \in \mathcal{I}$. Define a set $\bar{\mathcal{X}}(\epsilon) \subseteq \mathbb{Z}^{|\mathcal{F}|}$ that collects all of these $\bar{\mathbf{x}}^i$ values:

$$\bar{\mathcal{X}}(\epsilon) := \left\{ \bar{\mathbf{x}} \in \mathbb{Z}^{|\mathcal{I}| \times |\mathcal{F}|} \ : \ \exists \mathbb{Q} \in \mathcal{B}(\epsilon) \text{ s.t. } \mathbb{Q}_X^i(\bar{\mathbf{x}}^i) = 1 \ \forall i \in \mathcal{I} \right\}.$$

By estimating the expectation in (EC.3) via the training sample average, we can represent (EC.3) under the additional constraint that $\mathbb{Q}_X^i(\mathbf{x}) \in \{0, 1\}$ as the following problem:

$$\max_{\pi \in \Pi_d} \min_{\bar{\mathbf{x}} \in \bar{\mathcal{X}}(\epsilon)} \sum_{i \in \mathcal{I}} \mathbb{I} \left[\pi(\bar{\mathbf{x}}^i) = y^i \right], \tag{EC.4}$$

which with a change of variables to $\boldsymbol{\xi}^i := \bar{\mathbf{x}}^i - \mathbf{x}^i$, becomes equivalent to $(\mathcal{P}_\Xi)$.

One interpretation of the Wasserstein distributionally robust problem (EC.3) is that it allows for probabilistic realizations of our data, as represented by the value of $\mathbb{Q}$. In the approach presented in our paper $(\mathcal{P}_\Xi)$, we hedge against possible realizations of our data under the distribution seen in testing, and do not allow for these realizations to be probabilistic. A key observation is that, as shown by the problem (EC.4), the set of worst-case realizations explored in the robust approach in $(\mathcal{P}_\Xi)$ is a subset of the worst-case realizations in the distributionally robust approach (EC.3). This means that $(\mathcal{P}_\Xi)$ may generate less conservative

solutions to solving (EC.3), as $(\mathcal{P}_\Xi)$ only hedges against possible realizations of the data. Lastly, we note that since $(\mathcal{P}_\Xi)$ contains additional constraints in the inner minimization problem in comparison to (EC.3), the objective value of (EC.3) must be no more than the objective value of $(\mathcal{P}_\Xi)$. We formalize the relationship between the robust optimization problem we examine in this paper $(\mathcal{P}_\Xi)$ and the distributionally robust optimization version of this problem (EC.3) in the following proposition.

PROPOSITION EC.1. *The set of feasible solutions to the robust optimization problem* $(\mathcal{P}_\Xi)$ *are a subset of feasible solutions to the distributionally robust optimization problem* (EC.3), *and the objective value of* $(\mathcal{P}_\Xi)$ *is no less than the objective value of* (EC.3).

EC.2. Iterating over Paths in Problem (11)

We will now show why the assumption that the cut-set $\mathbf{q}^i$ incident to the path of the data sample after perturbation minimizes (11) for given $\boldsymbol{\xi}$ through the following proposition.

PROPOSITION EC.2. *For optimal solutions of* (11), *one of the following two statements holds for each sample* $i \in \mathcal{I}$

1. *The source set of a minimum cut is* $\{s\}$, *that is,* $q^i_{s,1} = 1$ *and* $q^i_{n,l(n)} = q^i_{n,r(n)} = 0$ *for all* $n \in \mathcal{N}$. *In this case* $t^i = 1$.

2. *The source set of a minimum cut is a path from* s *to a prediction node with a label other than* y^i, *that is,* $\bar{w}_{n,y^i} = 0$. *In this case* $t^i = 0$.

Proof of Proposition EC.2 First observe that setting $q^i_{s,1} = 1$ and all remaining variables to zero indeed satisfies constraints (7a)-(7d), and the associated objective value is $t^i = 1$. Suppose the first statement in the proposition does not hold and this solution is not optimal. Since, for all values of $\boldsymbol{\xi}^i$, the arc capacities are either 0 or 1, it follows that the objective values corresponding to optimal solutions is $t^i = 0$. Note that if a prediction node n with label y^i ($\bar{w}_{n,y^i} = 1$) is in the cut ($q^i_{n,t} = 1$), then $t^i \geq 1$, and no such solutions can be optimal.

If a branching node n ($\bar{b}_{nf\theta} = 1$ for some feature f and level θ) is in the cut but none of its descendants are ($q^i_{n,l(n)} = q^i_{n,r(n)} = 1$), then $\mathbb{I}\left[x^i_f + \xi^i_f \leq \theta\right] q^i_{n,l(n)} + \mathbb{I}\left[x^i_f + \xi^i_f \geq 1 + \theta\right] q^i_{n,r(n)} = 1$ for all values of ξ^i_f. Thus in this case $t^i \geq 1$ and no such solutions can be optimal. If a branching node is in the cut and both of its descendants are as well ($q^i_{n,l(n)} = q^i_{n,r(n)} = 0$), then depending on the value of $\boldsymbol{\xi}^i$, either $\mathbb{I}\left[x^i_f + \xi^i_f \leq \theta\right] = 0$ or $\mathbb{I}\left[x^i_f + \xi^i_f \geq 1 + \theta\right] = 0$. In the first case, one may set $q^i_{n,l(n)} = 1$ (and set additional cutset variables on the left subtree to zero),
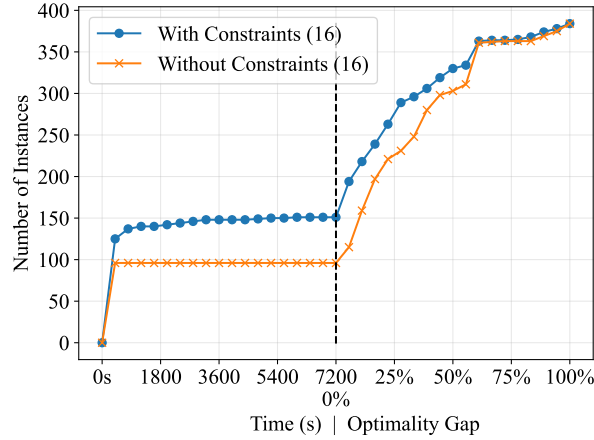
Figure EC.1   Companion figure to section EC.3. Number of `monk1`, `monk2`, and `monk3` instances solved to optimality on the 'Time' axis and the number of instances with MIO optimality gap no larger than each given value at the time limit on the 'Optimality Gap' axis. The blue line corresponds to the algorithm presented in this paper where constraints (16) are added at each subproblem, and the orange line is the same algorithm where constraints (16) are not added.

recovering a solution with the same (or less) cost; the second case is analogous. Therefore, we find that if $t^i = 0$, then there exists an optimal solution that is a path.   □

Due to proposition EC.2, we only need to solve (12) over all possible paths $\mathcal{O}(|\mathcal{L}|)$ for each sample to solve the subproblem and obtain a violated constraint (14) to add back to the main problem.

EC.3. Computation Times Improvement from Adding Constraints (16)

We add additional constraints of the form (16) to our model in order to improve on computation times. We test the magnitude of this improvement by comparing our algorithm that adds constraints (16) against the algorithm that does not add constraints (16) at each subproblem. We run such algorithms on the instances with the `monk1`, `monk2`, and `monk3` datasets, and summarize the computation times and optimality gaps in Figure EC.1. We observe that 96 out of the 384 instances are solved within the time limit on the algorithm without added constraints (16). Our algorithm with additional constraints (16) is able to solve the same number of instances within 32 seconds, corresponding to a speedup of $\frac{7200}{32} = 225$ times.

EC.4. Variants and Extensions

In this section, we describe how our model and solution approach can be adapted to handle bounded integer data, one-sided distribution shifts, distribution shifts on categorical features, and distribution shifts on mixed datasets.

EC.4.1. Calibration of Uncertainty Set Parameters for Bounded Integer Data

In many applications, there exist known bounds on the values of the integer features. There may be a bound on only one end of an integer feature (e.g., age, income) or on both ends (e.g., integer rating on a bounded scale, binary features). Here, we discuss how to tune the uncertainty set parameters for datasets involving bounded integer features with known bounds.

EC.4.1.1. Integer Features with One-Sided Bounds

We first consider the case where all covariates admit only a one-sided bound. Without loss of generality, we assume that all bounds are lower bounds, and we denote the lower bound on feature f by L_f (a symmetric argument can be made if all features are upper bounded instead).

To tune the hyperparameters of (3), we assume that the probability of certainty in the nominal value of feature f for sample i, ρ_f^i, satisfies $0 < \rho_f^i < 1$. Under this assumption, the perturbation ξ_f^i is distributed as the truncated symmetric geometric distribution conditioned on $x_f^i + \xi_f^i \geq L_f$. That is, for ζ_f^i the realization of the perturbation of x_f^i, the probability that x_f^i is perturbed by ξ_f^i is

$$\mathbb{P}(\xi_f^i = \zeta_f^i) := \rho_f^i (r_f^i)^{|\zeta_f^i|} \tag{EC.5}$$

for $0 < r_f^i \leq 1$ some constant that makes $\mathbb{P}$ a distribution over the support of all ξ_f^i that satisfy $x_f^i + \xi_f^i \geq L_f$. To find the value of r_f^i for (EC.5), we utilize the following lemma.

LEMMA EC.1. *For $0 < \rho_f^i < 1$, there exists a real-valued solution r_f^i to*

$$\rho_f^i (r_f^i)^{x_f^i - L + 1} - (\rho_f^i + 1) r_f^i + 1 - \rho_f^i = 0 \tag{EC.6}$$

such that $0 < r_f^i < 1$. Furthermore, when used in (EC.5), r_f^i defines a valid distribution over the support of all $\xi_f^i \in \mathbb{Z}$ that satisfy $x_f^i + \xi_f^i \geq L_f$.

Proof of Lemma EC.1 To find an r_f^i that makes (EC.5) a probability mass function, we solve the infinite polynomial equation

$$\sum_{\xi_f^i = L_f - x_f^i}^{\infty} \rho_f^i (r_f^i)^{|\xi_f^i|} = 1, \tag{EC.7}$$

and choose the solutions such that $0 < r_f^i < 1$. We can regroup terms in the left hand side of the above sum to get

$$\rho_f^i + 2 \sum_{\xi_f^i = 1}^{x_f^i - L_f} \rho_f^i (r_f^i)^{\xi_f^i} + \sum_{\xi_f^i = x_f^i - L_f + 1}^{\infty} \rho_f^i (r_f^i)^{\xi_f^i} = 1.$$

Then, using the closed form of both a finite and infinite geometric series, an equivalent finite polynomial equation to solve problem (EC.7) is

$$\rho_f^i + 2\left(\frac{\rho_f^i r_f^i \left(1 - (r_f^i)^{x_f^i - L_f}\right)}{1 - r_f^i}\right) + \left(\frac{\rho_f^i (r_f^i)^{x_f^i - L_f}}{1 - r_f^i}\right) = 1$$

$$\Leftrightarrow \rho_f^i - \rho_f^i r_f^i + 2\rho_f^i r_f^i - 2\rho_f^i (r_f^i)^{x_f^i - L_f + 1} + \rho_f^i (r_f^i)^{x_f^i - L_f + 1} = 1 - r_f^i$$

$$\Leftrightarrow \rho_f^i (r_f^i)^{x_f^i - L + 1} - (\rho_f^i + 1)r_f^i + 1 - \rho_f^i = 0$$

which is exactly (EC.6).

Now we show that a solution $0 < r_f^i < 1$ exists in the above equation. Let

$$f(r_f^i) := \rho_f^i (r_f^i)^{x_f^i - L + 1} - (\rho_f^i + 1)r_f^i + 1 - \rho_f^i,$$

i.e. the function on the left-hand side of (EC.6) that we wish to find real roots of. Thus, $f(0) = 1 - \rho_f^i$ and $f(1) = -\rho_f^i$. Since $0 < \rho_f^i < 1$, $f(0) > 0$, and $f(1) < 0$, by Bolzano's Theorem, there exists a solution r_f^i to (EC.6) such that $0 < r_f^i < 1$. $\qquad\square$

Standard root-finding approaches (e.g., the bisection method) can be used to solve (EC.6) in order to find r_f^i. Once r_f^i is found, we set up a hypothesis test with threshold λ of the form in a similar fashion as done in section 4, which yields the following condition on ξ_f^i:

$$\frac{\prod_{i \in \mathcal{I}} \prod_{f \in \mathcal{F}} \rho_f^i (r_f^i)^{|\xi_f^i|}}{\prod_{i \in \mathcal{I}} \prod_{f \in \mathcal{F}} \rho_f^i} \geq \lambda^{|\mathcal{I}|}, \tag{EC.8}$$

reducing down to

$$\sum_{i \in \mathcal{I}} \sum_{f \in \mathcal{F}} |\xi_f^i| \log\left(\frac{1}{r_f^i}\right) \leq -|\mathcal{I}| \log \lambda. \tag{EC.9}$$

Therefore, the tuned hyperparameters for this case are $\gamma_f^i = \log \frac{1}{r_f^i}$ and $\epsilon = -|\mathcal{I}| \log \lambda$.

EC.4.1.2. Integer Features with Both Upper and Lower Bounds We now assume that there is both a lower bound L_f and an upper bound U_f on all features f. As is the case with the one-sided bounded features, we set up a hypothesis testing framework to tune the hyperparameters. We assume that $\frac{1}{U_f - L_f + 1} \leq \rho_f^i \leq 1$ for all $i \in \mathcal{I}$ and $f \in \mathcal{F}$, and utilize the same truncated symmetric geometric distribution characterized by (EC.5) with support $[L_f - x_f^i, U_f - x_f^i]$. Note that we place the lower bound $\frac{1}{U_f - L_f + 1}$ on ρ_f^i as assuming $\rho_f^i = \frac{1}{U_f - L_f + 1}$ under maximal entropy makes the perturbation of feature f at sample i uniformly distributed, meaning that there is maximal uncertainty on the value of $x_f^i + \xi_f^i$. For $\rho_f^i > \frac{1}{U_f - L_f + 1}$, we find an $0 < r_f^i < 1$ for each sample i by the following lemma:

LEMMA EC.2. *For $\frac{1}{U_f - L_f + 1} < \rho_f^i < 1$, there exists a real-valued solution r_f^i to*

$$\rho_f^i(r_f^i)^{\max\{U_f - x_f^i, x_f^i - L_f\} + 1} + \rho_f^i(r_f^i)^{\min\{U_f - x_f^i, x_f^i - L_f\} + 1} - (\rho_f^i + 1)r_f^i + 1 - \rho_f^i = 0. \quad (\text{EC}.10)$$

such that $0 < r_f^i < 1$. Furthermore, when used in (EC.5), r_f^i defines a valid distribution over the support of all $\xi_f^i \in \mathbb{Z}$ that satisfy $L_f \le x_f^i + \xi_f^i \le U_f$.

Proof of Lemma EC.2 To find an r_f^i that makes (EC.5) a probability mass function, we solve the polynomial equation

$$\sum_{\xi_f^i = L_f - x_f^i}^{U_f - x_f^i} \rho_f^i(r_f^i)^{|\xi_f^i|} = 1. \quad (\text{EC}.11)$$

Let $M = \max\{U_f - x_f^i, x_f^i - L_f\}$ and $m = \min\{U_f - x_f^i, x_f^i - L_f\}$. We can regroup terms in the left hand side of the above sum to get

$$\rho_f^i + 2\sum_{\xi_f^i = 1}^{m} \rho_f^i(r_f^i)^{\xi_f^i} + \sum_{\xi_f^i = m+1}^{M} \rho_f^i(r_f^i)^{\xi_f^i} = 1.$$

Then, using the closed form of a finite geometric series, (EC.11) takes the simplified form of

$$\rho_f^i + 2\left(\frac{\rho_f^i r_f^i \left(1 - (r_f^i)^m\right)}{1 - r_f^i}\right) + \left(\frac{\rho_f^i(r_f^i)^{m+1}\left(1 - (r_f^i)^{M-m}\right)}{1 - r_f^i}\right) = 1$$

$$\Leftrightarrow \rho_f^i - \rho_f^i r_f^i + 2\rho_f^i r_f^i - 2\rho_f^i(r_f^i)^{m+1} + \rho_f^i(r_f^i)^{m+1} - \rho_f^i(r_f^i)^{M+1} = 1 - r_f^i$$

$$\Leftrightarrow \rho_f^i(r_f^i)^{M+1} + \rho_f^i(r_f^i)^{m+1} - (\rho_f^i + 1)r_f^i + 1 - \rho_f^i = 0.$$

which is exactly (EC.10).

We now show that the a real-valued $r_f^i \in (0,1)$ exists. Let $f(r_f^i)$ be the function on the left-hand side of (EC.10)

$$f(r_f^i) := \rho_f^i(r_f^i)^{M+1} + \rho_f^i(r_f^i)^{m+1} - (\rho_f^i + 1)r_f^i + 1 - \rho_f^i$$

that we wish to find real roots of. Thus, $f(0) = 1 - \rho_f^i$ and $f(1) = 0$. The first derivative of f is

$$f'(r_f^i) = (M+1)\rho_f^i(r_f^i)^M + (m+1)\rho_f^i(r_f^i)^m - (\rho_f^i + 1). \quad (\text{EC}.12)$$

Thus, $f'(1) = (M+m+1)\rho_f^i - 1 = (U_f - L_f + 1)\rho_f^i - 1$. For $\rho_f^i > \frac{1}{U_f - L_f + 1}$, we see that $f'(1) > 0$. Thus, there must exist an $\bar{r} \in (0,1)$ such that $f(\bar{r}) < 0$. And since $f(0) > 0$, by Bolzano's Theorem, there must exist an $r_f^i \in (0,1)$ that satisfies $f(r_f^i) = 0$. $\qquad\square$

Similar to section EC.4.1.1, the hypothesis test is set up with the found r_f^i values through Lemma EC.2. The hypothesis test is (EC.9), and the tuned parameters of uncertainty set (3) are $\gamma_f^i = \log \frac{1}{r_f^i}$ and $\epsilon = -|\mathcal{I}| \log \lambda$.

Note that binary features are a special case of integer features with lower bound 0 and upper bound 1. Using the truncated symmetric geometric distribution (EC.5) with $\frac{1}{2} \leq \rho_f^i \leq 1$, a perturbation ξ_f^i is distributed as

$$\mathbb{P}(\xi_f^i = \zeta_f^i) := \rho_f^i \left(\frac{1 - \rho_f^i}{\rho_f^i} \right)^{|\zeta_f^i|} = \begin{cases} \rho_f^i & \text{if } |\zeta_f^i| = 0 \\ 1 - \rho_f^i & \text{otherwise.} \end{cases} \tag{EC.13}$$

Thus, it follows that the tuned hyperparameters for an uncertainty set (3) with binary features ($L_f = 0$ and $U_f = 1$) is $\gamma_f^i = \log \frac{\rho_f^i}{1 - \rho_f^i}$ and $\epsilon = -|\mathcal{I}| \log \lambda$.

EC.4.2. One-Sided Distribution Shifts

In some applications, the direction of the distribution shift may be known. For instance, say that only nonnegative shifts in the values of the covariates are possible between training and deployment phases. Such a scenario can occur when, for example, there is a change in the framing of a survey question between training and deployment that reduces the sensitivity of a question, skewing the distribution of answers in the positive direction. In such settings, a model that uses set (3) hedges against infeasible shifts and the uncertainty set

$$\Xi_+ := \left\{ \boldsymbol{\xi} \in \mathbb{Z}_+^{|\mathcal{I}| \times |\mathcal{F}|} \; : \; \sum_{i \in \mathcal{I}} \sum_{f \in \mathcal{F}} \gamma_f^i \xi_f^i \leq \epsilon \right\} \tag{EC.14}$$

that only allows for nonnegative values of $\boldsymbol{\xi}$ is more appropriate. To solve $(\mathcal{P}_{\Xi_+})$, we can use the same approach as described in section 3, only changing the subproblem slightly by restricting the perturbations $\boldsymbol{\xi}^i$ considered to be in $\mathbb{Z}_+^{|\mathcal{F}|}$ for each $i \in \mathcal{I}$ (namely, in minimization problem (12)). A similar uncertainty set can likewise be defined to allow only nonpositive shifts or to allow a mixture of different one-sided shifts, where $(\mathcal{P}_\Xi)$ with such an uncertainty set can be solved with a similar adaptation to the subproblem.

For unbounded features, the uncertainty set calibration method in section 4 can be applied in the same way as the two-sided perturbation method, as the hypothesis test is on the magnitude of perturbation regardless of direction of perturbation. For bounded features, the uncertainty set calibration method in section EC.4.1 can also be used by setting L_f to the value of x_f^i.

EC.4.3. Distribution Shifts on Categorical Features

With some modifications, the modeling and solution approaches in this paper can be applied to handle general (not necessarily binary) categorical features. Assuming the entire original dataset consists of categorical features, we index categorical features in the original data in the set $\mathcal{C}$ and one-hot encode all features to obtain the new dataset features $\mathcal{F}$, wherein features indexed in the set $\mathcal{F}_c \subseteq \mathcal{F}$ are used in the one-hot encoding of categorical feature c (i.e., the sets $\{\mathcal{F}_c\}_{c \in \mathcal{C}}$ constitute a partition of $\mathcal{F}$). Then only one of two scenarios can materialize for each sample $i \in \mathcal{I}$ and categorical feature $c \in \mathcal{C}$: either the categorical feature c is not perturbed (in which case $\xi_f^i = 0$ for all $f \in \mathcal{F}_c$) or it is, in which case the category value is changed from $f' \in \mathcal{F}_c$ to $\bar{f} \in \mathcal{F}_c \backslash \{f'\}$, i.e., $\xi_{f'}^i = -1$ for $f' \in \mathcal{F}_c$ such that $x_{f'}^i = 1$, $\xi_{\bar{f}}^i = 1$ for some $\bar{f} \in \mathcal{F}_c \backslash \{f'\}$ and $\xi_f^i = 0 \ \forall f \in \mathcal{F}_c \backslash \{f', \bar{f}\}$. We characterize the values of ξ by the following uncertainty set:

$$
\begin{aligned}
\Xi_{\text{cat}} := \big\{ \ \boldsymbol{\xi} \in \{-1, 0, 1\}^{|\mathcal{I}| \times |\mathcal{F}|} \ : \ & \\
\sum_{i \in \mathcal{I}} \sum_{f \in \mathcal{F}} \gamma_f^i |\xi_f^i| \leq \epsilon, & \\
0 \leq x_f^i + \xi_f^i \qquad & \forall i \in \mathcal{I}, f \in \mathcal{F}_k, c \in \mathcal{C}, \\
\sum_{f \in \mathcal{F}_c} x_f^i + \xi_f^i = 1 \qquad & \forall i \in \mathcal{I}, c \in \mathcal{C} \ \big\}.
\end{aligned}
\tag{EC.15}
$$

where the first constraint of (EC.15) is the same as the other previously mentioned uncertainty sets, which penalize any change in feature $f \in \mathcal{F}$ with γ_f^i with a total budget of ϵ and the other constraints define values of ξ that either perturb or do not perturb each sample and categorical feature. The second and third constraints of (EC.15) ensure that $\mathbf{x} + \boldsymbol{\xi}$ maintains a one-hot encoding of each categorical feature.

When solving problem (12) in algorithm 1 for each $i \in \mathcal{I}$, the perturbations that lead down each path must satisfy the constraints on $\boldsymbol{\xi}$ in uncertainty set (EC.15) to ensure that the perturbation on each categorical feature corresponds to a change of value in the one-hot encoding. In other words, for each categorical feature $c \in \mathcal{C}$ that is perturbed in sample i, we must have $\xi_{f'}^i = -1$ and $\xi_{\bar{f}}^i = 1$ for two distinct $f', \bar{f} \in \mathcal{F}_c$. So by the left-hand side of the first constraint of (EC.15), perturbing one categorical feature incurs a cost of perturbation of $\gamma_{f'}^i + \gamma_{\bar{f}}^i$. All else in our solution method remains the same.

To calibrate the uncertain parameters of (EC.15), we again note that due to the second and third constraints in (EC.15), any perturbation of one feature $f' \in \mathcal{F}_c$ will cause an equal

and opposite perturbation in some other feature $\bar{f} \in \mathcal{F}_c \backslash \{f'\}$. We assume that the probability of no perturbation in categorical feature $c \in \mathcal{C}$ of sample i is $\frac{1}{|\mathcal{F}_c|} \leq \rho_c^i \leq 1$, and that the realizations of all other values are all equally likely. Similar to the bounded integer case (see section EC.4.1.2), we place the lower bound of $\frac{1}{|\mathcal{F}_c|}$ on ρ_c^i – this allows for maximal uncertainty on the value of c. Then, the probability of the perturbation of categorical feature $c \in \mathcal{C}$ encoded by the collection of perturbations $\{\xi_f^i\}_{f \in \mathcal{F}_c}$ on the one-hot encoding of c, is

$$\mathbb{P}(\{\xi_f^i\}_{f \in \mathcal{F}_c} = \{\zeta_f^i\}_{f \in \mathcal{F}_c}) := \mathbb{I}\left[\sum_{f \in \mathcal{F}_c} |\zeta_f^i| = 0\right] \rho_c^i + \mathbb{I}\left[\sum_{f \in \mathcal{F}_c} |\zeta_f^i| = 2\right]\left(\frac{1 - \rho_c^i}{|\Phi_k| - 1}\right). \tag{EC.16}$$

As done before, we set up our hypothesis test as a likelihood ratio test with threshold $\lambda^{|\mathcal{I}|}$:

$$\frac{\prod_{i \in \mathcal{I}} \prod_{c \in \mathcal{C}} \mathbb{P}(\{\xi_f^i\}_{f \in \mathcal{F}_c})}{\prod_{i \in \mathcal{I}} \prod_{c \in \mathcal{C}} \mathbb{P}(\{0\}_{f \in \mathcal{F}_c})} \geq \lambda^{|\mathcal{I}|}. \tag{EC.17}$$

Plugging in (29) into (30) yields

$$\sum_{i \in \mathcal{I}} \sum_{c \in \mathcal{C}} -\log\left(\mathbb{I}\left[\sum_{f \in \mathcal{F}_c} |\xi_f^i| = 0\right] + \mathbb{I}\left[\sum_{f \in \mathcal{F}_c} |\xi_f^i| = 2\right]\left(\frac{1 - \rho_c^i}{\rho_c^i(|\mathcal{F}_c| - 1)}\right)\right) \leq -|\mathcal{I}| \log \lambda. \tag{EC.18}$$

Thus, if $\sum_{f \in \mathcal{F}_c} |\xi_f^i| = 0$, i.e., no perturbation occurs, no cost is incurred; if $\sum_{f \in \mathcal{F}_c} |\xi_f^i| = 2$, i.e., a perturbation occurs, then a cost of $-\log\left(\frac{1 - \rho_c^i}{\rho_c^i(|\mathcal{F}_c| - 1)}\right)$ incurred. We can represent (EC.18) more compactly as

$$\sum_{i \in \mathcal{I}} \sum_{f \in \mathcal{F}} \frac{1}{2} \log\left(\frac{\rho_c^i(|\{\mathcal{F}_c\}| - 1)}{1 - \rho_c^i}\right) |\xi_f^i| \leq -|\mathcal{I}| \log \lambda. \tag{EC.19}$$

It follows from the above that the tuned hyperparameters in the uncertainty set (EC.15) are $\gamma_f^i = \frac{1}{2} \log\left(\frac{\rho_c^i(|\{\mathcal{F}_c\}| - 1)}{1 - \rho_c^i}\right)$ and $\epsilon = -|\mathcal{I}| \log \lambda$.

We note that if c is binary and one-hot encoded into two features following the process described in this section, the tuned uncertainty set parameter values are equivalent to that described in section EC.4.1.2.

EC.4.4. Mixed Datasets and Distribution Shifts

Often, datasets have a mixture of unbounded integer, bounded integer, binary, and categorical data, and known directions of distribution shifts vary across features. In such scenarios, we can adapt the models and calibration methods presented in section 2.2 and 4, as well as the Electronic Companion sections EC.4.1, EC.4.2, and EC.4.3 to create uncertainty sets that capture this information.

Consider the following uncertainty set

$$\Xi_{\mathrm{mixed}} := \left\{ \boldsymbol{\xi} \in \mathcal{M} : \sum_{i \in \mathcal{I}} \sum_{f \in \mathcal{F}} \gamma_f^i |\xi_f^i| \leq \epsilon \right\} \tag{EC.20}$$

where $\mathcal{M} \subseteq \mathbb{Z}^{|\mathcal{I}| \times |\mathcal{F}|}$ can be use to place restrictions on the kinds of shifts allowed for each datapoint (e.g., to capture one-sided perturbations, bounded shifts, or categorical features). Letting ρ_f^i denote the probability of certainty in the nominal value of the data, the tuned hyperparameters are $\epsilon = -|\mathcal{I}| \log \lambda$ and

$$\gamma_f^i := \begin{cases} \log\left(\frac{1}{1-\rho_f^i}\right) & \text{for } f \text{ an unbounded integer} \\ \log\left(\frac{1}{r_f^i}\right) & \text{for } f \text{ a bounded integer} \\ \log\left(\frac{\rho_f^i}{1-\rho_f^i}\right) & \text{for } f \text{ binary} \\ \frac{1}{2}\log\left(\frac{\rho_f^i(|\{\Phi_k \; : \; f \in \Phi_k\}|-1)}{1-\rho_f^i}\right) & \text{for } f \text{ part of a categorical feature,} \end{cases} \tag{EC.21}$$

where $0 < r_f^i < 1$ is found by Lemma EC.1 for one-sided bounds or Lemma EC.2 for two-sided bounds (see Electronic Companion EC.4.1). The solution method described in section 3 remains the same, only needing to adapt the subproblem algorithm 1 to only allow perturbations from the set $\mathcal{M}$.

EC.5. Numerical Comparison to Alternative Robust Method

We empirically analyze the performance of our method against the method of Bertsimas et al. (2019), see section 5 for comparisons between each model. For simplicity, we will refer to our method as `RCT` and the method of Bertsimas et al. (2019) as `RCT-B` in this section.

EC.5.1. Setup

For fair comparison to the method of Bertsimas et al. (2019), we only evaluate `RCT-B` on instances corresponding to datasets in Table 2 with only binary features (i.e., `house-votes-84` and `spect`). For datasets with non-binary features, `RCT-B` cannot tackle problems with the uncertainty sets we consider in this paper.

To adapt our uncertainty set to a comparable uncertainty set in the `RCT-B` formulation, we first transformed the uncertainty set used in the above experiment (i.e. the uncertainty set (3) tuned using (EC.13)) into the uncertainty set (21) for $p = \infty$. For each instance of

our original problem, let ρ be the average probability of certainty used to generate the ρ^i_f values in the experiments on our methods. Then, (EC.13) becomes for each $i \in \mathcal{I}$

$$\sum_{f \in \mathcal{F}} |\xi^i_f| \leq -\frac{\log \lambda}{\log \left(\frac{\rho}{1-\rho} \right)}.$$

This matches the form of (21) for $p = 1$, and therefore can be used to define the parameters of (21) and have a more similar comparison to our methodology. Additionally, we apply no regularization term for all experiments and set a time limit of 7200 seconds, which is the same as in our formulation.

EC.5.2. Accuracy Under Distribution Shifts

We analyze the worst-case and average-case accuracies on the same 5000 perturbed datasets for each instance of RCT and its analogous instance of RCT-B. The results for the worst-case and average case on expected perturbations are shown in Figure EC.2. We see that on binary datasets, there is on average a decrease in worst-case performance in the RCT-B method across all values of λ in comparison to the non-robust tree, whereas RCT has on average a favorable gain in worst-case performance on the same instances. Specifically, RCT returns a higher gain in worst-case performance than that of RCT-B by up to about 15% on average in the worst-case. In addition, the average-case performance of RCT-B suffers in comparison to the non-robust tree on average, whereas RCT on average remains on par with the average-case performance of a non-robust tree in the instances where λ is 0.75 or above. These results match the conclusions from proposition 1, where RCT-B is more conservative than RCT, potentially causing a bad performance in testing.

EC.5.3. Price of Robustness

For both RCT and RCT-B, we measure the decrease in accuracy from using a robust tree versus a non-robust tree under the case of no distribution shift in the test set, i.e., the price of robustness (Bertsimas and Sim 2004), and summarize this metric in Figure EC.3a. From the figure, we observe that for each value of λ, RCT has an average a price of robustness less than 0.25. In contrast, RCT-B has an average price of robustness of about 0.4 for all λ.
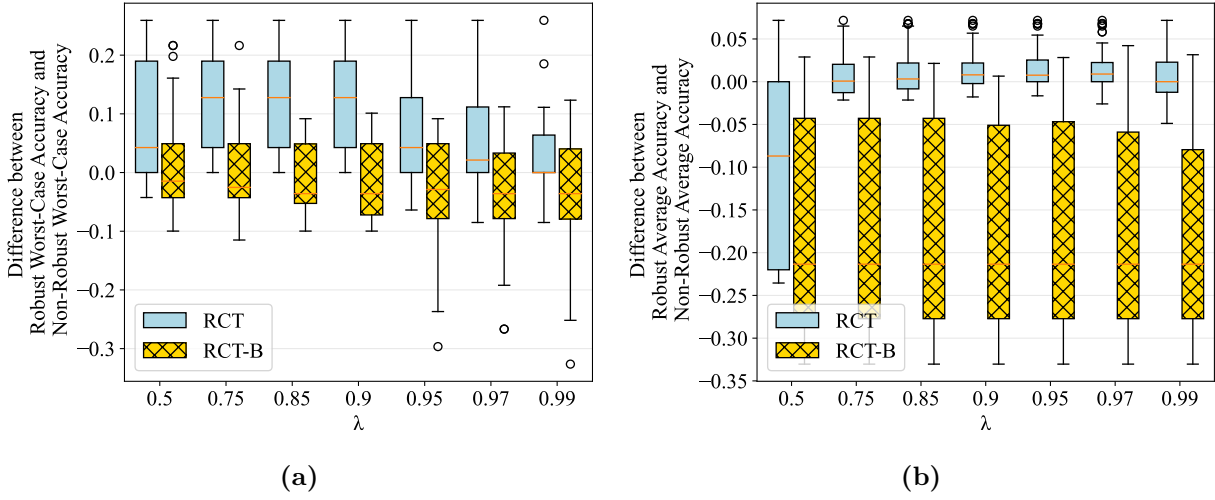
(a)

(b)

Figure EC.2   Companion figure to section EC.5.2. Distribution across problem instances of the gain in (a) worst-case accuracy and (b) average accuracy from using `RCT` versus `RCT-B` in dependence of λ on binary datasets. The parameter λ varies between 0.5 (more robust) and 0.99 (less robust).
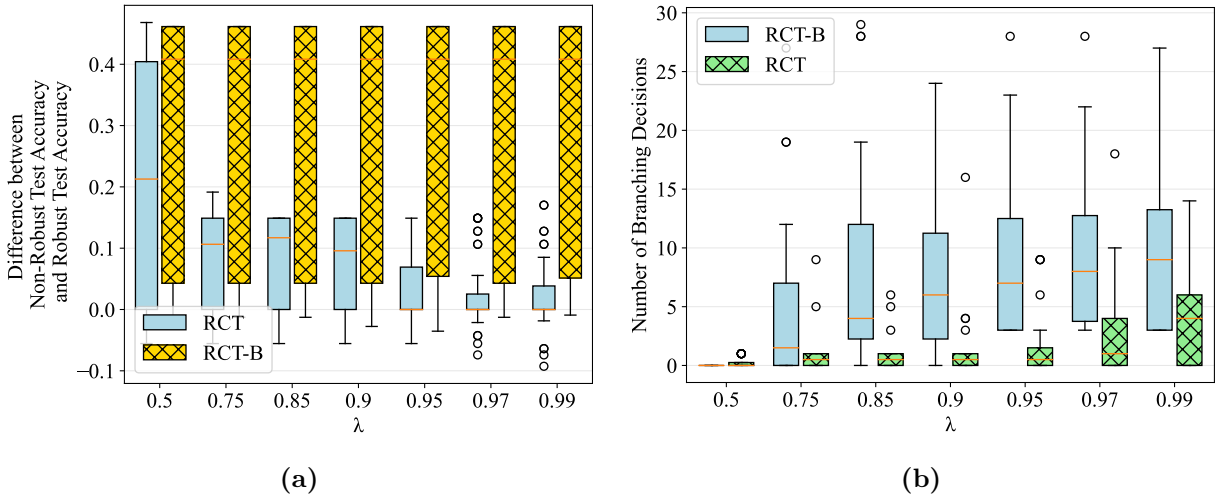

(a)

(b)

Figure EC.3   (a) Companion figure to section EC.5.3. Distribution across problem instances of the loss in nominal test accuracy from using `RCT` versus `RCT-B` in dependence of λ on binary datasets. The parameter λ varies between 0.5 (more robust) and 0.99 (less robust). (b) Companion figure to section EC.5.4. Distribution of the number of branching nodes across all instances with the same budget of uncertainty for both `RCT` and `RCT-B`.

EC.5.4. The Effect of Robustness on Sparsity

For each value of λ in Figure EC.3b, we compare the number of branching nodes of `RCT` against the method of `RCT-B`, aggregating across each comparable instance between the two methods. We first note that the number of branching nodes of `RCT-B` remains relatively

consistent for values of λ above 0.75, up until the budget of uncertainty becomes too large (i.e., $\lambda = 0.5$) where many instances yield a tree with no branching nodes. As previously mentioned, the method of Bertsimas et al. (2019) constrains every sample to have the same predictions across all realizations of the covariates. So with a large enough uncertainty set, all samples are given the majority label regardless of the covariates, which is the same as yielding a tree with no branching nodes.

In comparison, we see that `RCT` does not have the same behavior as `RCT-B`, as the number of branching nodes drops smoothly from a smaller uncertainty set to a larger uncertainty set. The tree structure changes through variations in the parameters of uncertainty more consistently in `RCT` in comparison to `RCT-B`, suggesting that the trees generated by `RCT` are closely tailored to the specific distribution shifts it hedges against.

EC.5.5. Computational Times

Figure EC.4 shows the computation times and optimality gaps of all instances of `RCT-B` in comparison to the corresponding instances of `RCT`. We see that about two times more instances of `RCT-B` can be solved within the time limit in comparison to `RCT`. However, for instances that could not be solved to optimality in the time limit, the optimality gap overall is much smaller in instances of `RCT` than instances of `RCT-B`. There are several reasons for this. In `RCT-B`, instances that resulted in no branching nodes were solved very quickly, as the only feasible trees were zero-depth ones. For other instances of `RCT-B` where the uncertainty set did allow for several feasible solutions, the convergence to optimality is noticeably less than that `RCT`, where nearly all instances of `RCT` are within fifty percent of the optimality gap. This suggests that in `RCT-B` instances where a nonzero-depth tree is the optimal tree, there is a greater computational costs of learning trees with `RCT-B` than with `RCT`.
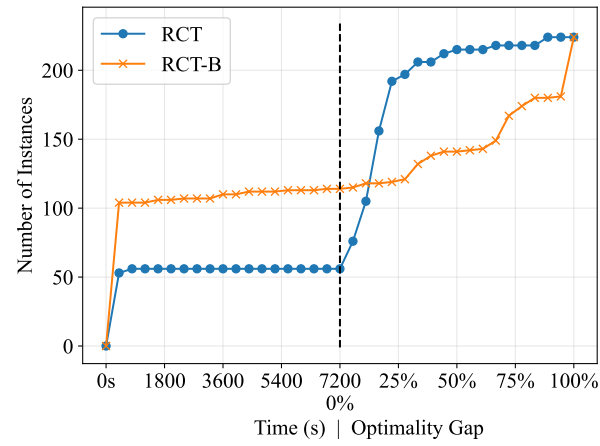
Figure EC.4   Companion figure to section EC.5.5. Number of instances solved to optimality on the 'Time' axis and number of instances with optimality gap no larger than each given value at the time limit on the 'Optimality Gap' axis for `RCT` and `RCT-B`.