

Mixed-Integer Convex Nonlinear Optimization with Gradient-Boosted Trees Embedded

Miten Mistry

Dimitrios Letsios

Imperial College London, South Kensington, SW7 2AZ, UK.

Gerhard Krennrich

Robert M. Lee

BASF SE, Ludwigshafen am Rhein, Germany.

Ruth Misener

Imperial College London, South Kensington, SW7 2AZ, UK.

Decision trees usefully represent sparse, high dimensional and noisy data. Having learned a function from this data, we may want to thereafter integrate the function into a larger decision-making problem, e.g., for picking the best chemical process catalyst. We study a large-scale, industrially-relevant mixed-integer nonlinear nonconvex optimization problem involving both gradient-boosted trees and penalty functions mitigating risk. This mixed-integer optimization problem with convex penalty terms broadly applies to optimizing pre-trained regression tree models. Decision makers may wish to optimize discrete models to repurpose legacy predictive models, or they may wish to optimize a discrete model that accurately represents a data set. We develop several heuristic methods to find feasible solutions, and an exact, branch-and-bound algorithm leveraging structural properties of the gradient-boosted trees and penalty functions. We computationally test our methods on concrete mixture design instance and a chemical catalysis industrial instance.

Key words: Gradient-boosted trees, branch-and-bound, mixed-integer convex programming, decomposition, catalysis

1. Introduction

Consider integrating an unknown function into an optimization problem, i.e., without a closed-form formula, but with a data set representing evaluations over a box-constrained feasibility domain. Optimization in the machine learning literature usually refers to the training procedure, e.g., model accuracy maximization (Sra et al. 2012, Snoek et al. 2012). This paper investigates optimization problems after the training procedure, where the trained predictive model is embedded in the optimization problem. We consider optimization methods for problems with gradient-boosted tree (GBT) models embedded (Friedman 2001, Hastie et al. 2009). Advantages of GBTs are myriad (Chen and Guestrin 2016, Ke

et al. 2017), e.g., they are robust to scale differences in the training data features, handle both categorical and numerical variables, and can minimize arbitrary, differentiable loss functions.

Lombardi and Milano (2018) survey approaches for embedding machine learning models as parts of decision-making problems. We *encode the machine learning model using the native language* (Lombardi and Milano 2018), i.e., in an optimization modeling framework. Resulting optimization models may be addressed using local (Nocedal and Wright 2006) or deterministic global (Schweidtmann and Mitsos 2019) methods. The value of global optimization is known in engineering (Boukouvala et al. 2016), e.g., local minima can lead to infeasible parameter estimation (Singer et al. 2006) or misinterpreted data (Bollas et al. 2009). For applications where global optimization is less relevant, we still wish to develop optimization methods for discrete and non-smooth machine learning models, e.g., regression trees. Discrete optimization methods allow repurposing a legacy model, originally built for prediction, into an optimization framework. In closely related work, Donti et al. (2017) train machine learning models to capture the task for which they will be used. This work focusses on generating optimal decisions after the machine learning model is trained.

Our optimization problem incorporates an additional, convex penalty term in the objective. This penalty accounts for risky predicted values where the machine learning model is not well trained due to missing data. But penalizing distance from the candidate solution to the existing data is not the only reason to add a convex penalty function, e.g., our numerical tests consider an instance with an additional soft constraint. Duran and Grossmann (1986) document convex terms common in process engineering: any of those convex nonlinear equations could be incorporated into this framework. Another possible application area is in portfolio optimization, e.g., extending the (convex) Markowitz model with cardinality constraint and buy-in threshold constraints (Bienstock 1996). Several authors have considered more elaborate extensions, e.g., by integrating uncertainty in the expected return estimate (Bonami and Lejeune 2009) or considering concave transaction costs (Konno and Wiyayanayake 2001). But the framework presented in this paper could use GBT models to develop data-driven uncertainty or cost models.

This paper considers a mixed-integer nonlinear optimization problem with convex nonlinearities (convex MINLP). The objective sums a discrete GBT-trained function and a continuous convex penalty function. We design exact methods computing either globally

optimal solutions, or solutions within a quantified distance from the global optimum. The convex MINLP formulation enables us to solve industrial instances with commercial solvers. We develop a new branch-and-bound method exploiting both the GBTs combinatorial structure and the penalty function convexity. Numerical results substantiate our approach. The manuscript primarily discusses GBTs, but both the mixed-integer linear programming (MILP) formulation and most of the branch-and-bound methods leverage tree-ensemble structure and can be applied to other tree-ensemble models, e.g., random forests and extremely randomized trees (Breiman 2001, Geurts et al. 2006).

This paper studies a problem that is closely related to Mišić (2017). Our work differs in that (i) Mišić (2017) studies a more basic version of our problem formulation (optimizing an objective function derived from tree ensembles whereas our objective also includes a convex penalty) and (ii) our contribution is a specialized branch-and-bound algorithm designed to solve our optimization problem at a large-scale.

Paper organization Section 3 introduces the optimization problem and Section 4 formulates it as a convex MINLP. Section 5 describes our branch-and-bound method. Section 6 defines the convex penalty term. Section 7 presents numerical results. Finally, Section 8 discusses further connections to the literature and Section 9 concludes.

2. Background

This section describes gradient-boosted trees (GBTs) (Friedman 2001, 2002). In this paper, GBTs are embedded into the Section 3 optimization problem. GBTs are a subclass of boosting methods (Freund 1995). Boosting methods train many weak learners iteratively that collectively produce a strong learner, where a weak learner is at least better than random guessing. Each boosting iteration trains a new weak learner against the residual of the previously trained learners by minimizing a loss function. For GBTs, the weak learners are classification and regression trees (Breiman et al. 1984).

This paper restricts its analysis to regression GBTs without categorical input variables. A trained GBT function is a collection of binary trees and each of these trees provides its own independent contribution when evaluating at \mathbf{x} .

DEFINITION 1. A trained GBT function is defined by sets $(\mathcal{T}, \mathcal{V}_t, \mathcal{L}_t)$ and values $(i(t, s), v(t, s), F_{t,l})$. The set \mathcal{T} indexes the trees. For a given tree $t \in \mathcal{T}$, \mathcal{V}_t and \mathcal{L}_t index the split and leaf nodes, respectively. At split node $t \in \mathcal{T}$, $s \in \mathcal{V}_t$, $i(t, s)$ and $v(t, s)$ return the split variable and value, respectively. At leaf node $t \in \mathcal{T}$, $l \in \mathcal{L}_t$, $F_{t,l}$ is its contribution.

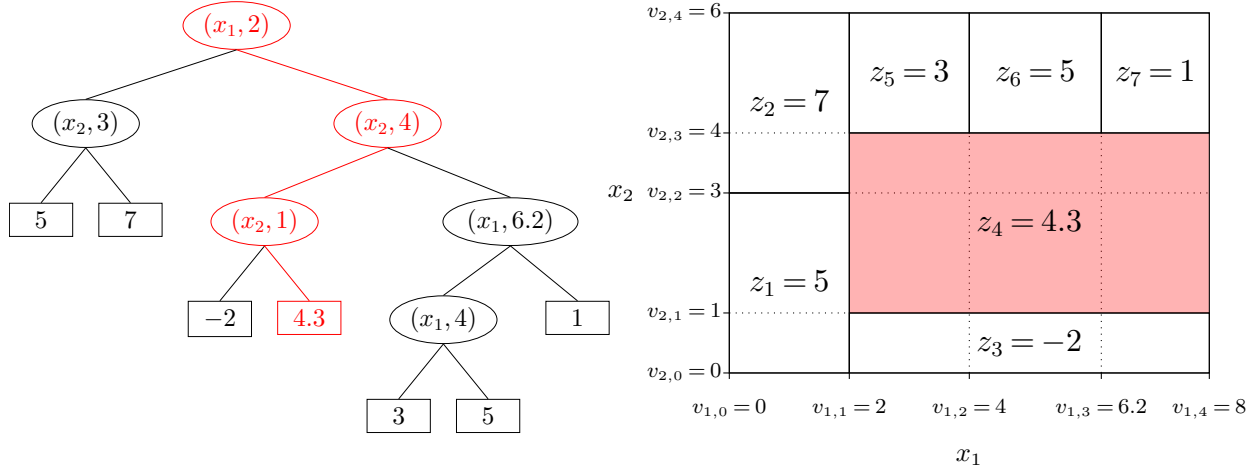


Figure 1 Gradient boosted tree, see Definition 1, trained in two dimensions. **Left:** gradient boosted tree. **Right:** recursive domain partition defined by tree on left. The highlighted path and region corresponds to the result of evaluating at $\mathbf{x} = (4.2, 2.8)^\top$ as in Example (1).

Tree $t \in \mathcal{T}$ evaluates at \mathbf{x} by following a root-to-leaf path. Beginning at the root node of t , each encountered split node $s \in \mathcal{V}_t$ assesses whether $x_{i(t,s)} < v(t, s)$ or $x_{i(t,s)} \geq v(t, s)$ and follows the left or right child, respectively. The leaf $l \in \mathcal{L}_t$ corresponding to \mathbf{x} returns t 's contribution $F_{t,l}$. Figure 1 shows how a single gradient-boosted tree recursively partitions the domain. The overall output, illustrated in Figure 2, sums all individual tree evaluations:

$$\text{GBT}(\mathbf{x}) = \sum_{t \in \mathcal{T}} \text{GBT}_t(\mathbf{x}).$$

EXAMPLE 1. Consider a trained GBT that approximates a two-dimensional function with $\mathcal{T} = \{t_1, \dots, t_{|\mathcal{T}|}\}$. To evaluate $\text{GBT}(\mathbf{x})$ where $\mathbf{x} = (4.2, 2.8)^\top$, let t_1 be the tree given by Figure 1, the highlighted path corresponds to evaluating at \mathbf{x} . The root split node query of $x_1 < 2$ is false, since $x_1 = 4.2$, so we follow the right branch. Following this branch encounters another split node. The next query of $x_2 < 4$ is true, since $x_2 = 2.8$, so we follow the left branch. The final branch reaches a leaf with value 4.3, hence $\text{GBT}_{t_1}(\mathbf{x}) = 4.3$. The remaining trees also return a value after making similar queries on \mathbf{x} . This results in $\text{GBT}(\mathbf{x}) = \sum_{i=1}^{|\mathcal{T}|} \text{GBT}_{t_i}(\mathbf{x}) = 4.3 + \sum_{i=2}^{|\mathcal{T}|} \text{GBT}_{t_i}(\mathbf{x})$.

3. Optimization Problem

This paper considers box-constrained optimization Problem (1), an MINLP where the objective sums a convex nonlinear function and a GBT-trained function:

$$\min_{\mathbf{v}^L \leq \mathbf{x} \leq \mathbf{v}^U} \underbrace{\text{cvx}(\mathbf{x})}_{\text{Convex Part}} + \underbrace{\text{GBT}(\mathbf{x})}_{\text{GBT Part}}, \quad (1)$$

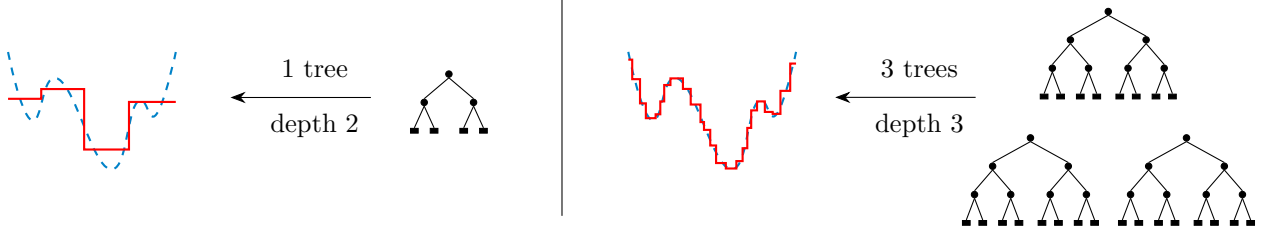


Figure 2 GBT approximations to the dashed function: 1 tree of depth 2 (left) and 3 trees of depth 3 (right).
k

Table 1 Mixed-integer convex programming model sets, parameters and variables.

Symbol	Description
v_i^L, v_i^U	Lower and upper bound of variable x_i
x_i	Continuous variable, $i \in \{1, \dots, n\}$
$t \in \mathcal{T}$	Indices of GBTs
$l \in \mathcal{L}_t$	Indices of leaves for tree t
$s \in \mathcal{V}_t$	Indices of split nodes for tree t
m_i	Number of variable x_i splitting values
$v_{i,j}$	Variable i 's j -th breakpoint, $j \in \{1, \dots, m_i\}$
$F_{t,l}$	Value of leaf (t, l)
$y_{i,j}$	Binary variable indicating whether variable $x_i < v_{i,j}$
$z_{t,l}$	Nonnegative variable that activates leaf (t, l)

where $\mathbf{x} = (x_1, \dots, x_n)^\top$ is the variable vector. $\text{GBT}(\mathbf{x})$ is the GBT-trained function value at \mathbf{x} . Table 1 defines the model sets, parameters and variables. Problem (1) is relevant, e.g., when a GBT function has been trained to data but we may trust an optimal solution close to regions with many training points. A convex penalty term may penalize solutions further from training data. For instance, consider quality maximization using historical data from a manufacturing process. The data may exhibit correlation between two process parameters, e.g., the temperature and the concentration of a chemical additive. A machine learned model of the system assigns weights to these parameters for future predictions. Lacking additional information, numerical optimization may produce candidate solutions with temperature and concentration combinations that (possibly incorrectly) suggest temperature is responsible for an observed effect. The convex penalty term helps control the optimizer's adventurousness by penalizing deviation from the training data subspace and is parameterized using principal component analysis (Vaswani et al. 2018). Large values of this risk control term generate conservative solutions. Smaller penalty values explore regions with greater possible rewards but also additional risk. Beyond modeling distance to training data, the convex penalty may represent additional soft constraints.

A given problem instance may sum independently-trained GBT functions. Without loss of generality, we equivalently optimize a single GBT function which is the union of all original GBTs.

4. Mixed-Integer Convex Formulation

Problem (1) consists of a continuous convex function and a discrete GBT function. The discrete nature of the GBT function arises from the left/right decisions at the split nodes. So we consider a mixed-integer nonlinear program with convex nonlinearities (convex MINLP) formulation. The main ingredient of the convex MINLP model is a mixed-integer linear programming (MILP) formulation of the GBT part which merges with the convex part via a linking constraint. The high level convex MINLP is:

$$\min_{\mathbf{v}^L \leq \mathbf{x} \leq \mathbf{v}^U} \text{cvx}(\mathbf{x}) + [\text{GBT MILP objective}] \quad (2a)$$

$$\text{s.t. } [\text{GBT MILP constraints}], \quad (2b)$$

$$[\text{Variable linking constraints}]. \quad (2c)$$

4.1. GBT MILP Formulation

We form the GBT MILP using the Mišić (2017) approach, which recalls the state-of-the-art in modeling piecewise linear functions (Misener et al. 2009, Misener and Floudas 2010, Vielma et al. 2010). Verwer et al. (2017) present an alternative MILP formulation. Alternative modeling frameworks include constraint programming (Rossi et al. 2006, Bonfietti et al. 2015) and satisfiability modulo theories (Lombardi et al. 2017, Mistry et al. 2018).

Figure 1 shows how a GBT partitions the domain $[\mathbf{v}^L, \mathbf{v}^U]$ of \mathbf{x} . Optimizing a GBT function reduces to optimizing the leaf selection, i.e., finding an optimal interval, opposed to a specific \mathbf{x} value. Aggregating over all GBT split nodes produces a vector of ordered breakpoints $v_{i,j}$ for each x_i variable: $v_i^L = v_{i,0} < v_{i,1} < \dots < v_{i,m_i} < v_{i,m_i+1} = v_i^U$. Selecting a consecutive pair of breakpoints for each x_i defines an interval where the GBT function is constant. Each point $x_i \in [v_i^L, v_i^U]$ is either on a breakpoint $v_{i,j}$ or in the interior of an interval. Binary variable $y_{i,j}$ models whether $x_i < v_{i,j}$ for $i \in [n] = \{1, \dots, n\}$ and $j \in [m_i] = \{1, \dots, m_i\}$. Binary variable $z_{t,l}$ is 1 if tree $t \in \mathcal{T}$ evaluates at node $l \in \mathcal{L}_t$ and 0 otherwise. Denote by \mathcal{V}_t the set of split nodes for tree t . Moreover, let $\text{Left}_{t,s}$ and $\text{Right}_{t,s}$ be the sets of subtree leaf nodes rooted in the left and right children of split node s , respectively.

MILP Problem (3) formulates the GBT (Mišić 2017). Equation (3a) minimizes the total value of the active leaves. Equation (3b) selects exactly one leaf per tree. Equations (3c) and (3d) activates a leaf only if all corresponding splits occur. Equation (3e) ensures that if $x_i \leq v_{i,j-1}$, then $x_i \leq v_{i,j}$. Without loss of generality, we drop the $z_{t,l}$ integrality constraint because any feasible assignment of \mathbf{y} specifies one leaf, i.e., a single region in Figure 1.

$$\min \sum_{t \in \mathcal{T}} \sum_{l \in \mathcal{L}_t} F_{t,l} z_{t,l} \quad (3a)$$

$$\text{s.t. } \sum_{l \in \mathcal{L}_t} z_{t,l} = 1, \quad \forall t \in \mathcal{T}, \quad (3b)$$

$$\sum_{l \in \text{Left}_{t,s}} z_{t,l} \leq y_{i(s),j(s)}, \quad \forall t \in \mathcal{T}, s \in \mathcal{V}_t, \quad (3c)$$

$$\sum_{l \in \text{Right}_{t,s}} z_{t,l} \leq 1 - y_{i(s),j(s)}, \quad \forall t \in \mathcal{T}, s \in \mathcal{V}_t, \quad (3d)$$

$$y_{i,j} \leq y_{i,j+1}, \quad \forall i \in [n], j \in [m_i - 1], \quad (3e)$$

$$y_{i,j} \in \{0, 1\}, \quad \forall i \in [n], j \in [m_i], \quad (3f)$$

$$z_{t,l} \geq 0, \quad \forall t \in \mathcal{T}, l \in \mathcal{L}_t. \quad (3g)$$

4.2. Linking Constraints

Equations (4a) and (4b) relate the continuous x_i variables, from the original Problem (1) definition, to the binary $y_{i,j}$ variables:

$$x_i \geq v_{i,0} + \sum_{j=1}^{m_i} (v_{i,j} - v_{i,j-1})(1 - y_{i,j}), \quad (4a)$$

$$x_i \leq v_{i,m_i+1} + \sum_{j=1}^{m_i} (v_{i,j} - v_{i,j+1})y_{i,j}, \quad (4b)$$

for all $i \in [n]$. We express the linking constraints using non-strict inequalities to avoid computational issues when optimizing with strict inequalities. Combining Equations (2) to (4) defines the mixed-integer nonlinear program with convex nonlinearities (convex MINLP) formulation to Problem (1). Appendix A lists the complete formulation.

4.3. Worst Case Analysis

The difficulty of Problem (1) is primarily justified by the fact that optimizing a GBT-trained function, i.e., Problem (3), is an NP-hard problem (Mišić 2017). This section shows

that the number of continuous variable splits and tree depth affects complete enumeration. These parameters motivate the branching scheme in our branch-and-bound algorithm.

In a GBT ensemble, each continuous variable x_i is associated with $m_i + 1$ intervals (splits). Picking one interval $j \in \{1, \dots, m_i + 1\}$ for each x_i sums to a total of $\prod_{i=1}^n (m_i + 1)$ distinct combinations. A GBT-trained function evaluation selects a leaf from each tree. But not all leaf combinations are valid evaluations. In a feasible leaf combination where one leaf enforces $x_i < v_1$ and another enforces $x_i \geq v_2$, it must be that $v_2 < v_1$. Let d be the maximum tree depth in \mathcal{T} . Then the number of leaf combinations is upper bounded by $2^{d|\mathcal{T}|}$. Since the number of feasibility checks for a single combination is $\frac{1}{2}|\mathcal{T}|(|\mathcal{T}|-1)$, an upper bound on the total number of feasibility checks is $2^{d|\mathcal{T}|-1}|\mathcal{T}|(|\mathcal{T}|-1)$. So the worst case performance of an exact method improves as the number of trees decreases.

5. Branch-and-Bound Algorithm

This section designs an exact branch-and-bound (B&B) approach. Using a divide-and-conquer principle, B&B forms a tree of subproblems and searches the domain of feasible solutions. Key aspects of B&B are: (i) rigorous lower (upper) bounding methods for minimization (maximization) subproblems, (ii) branch variable and value selection, and (iii) feasible solution generation. In the worst case, B&B enumerates all solutions, but generally it avoids complete enumeration by pruning subproblems, i.e., removing infeasible subproblems or nodes with lower bound exceeding the best found feasible solution (Morrison et al. 2016). This section exploits *spatial branching* that splits on continuous variables (Belotti et al. 2013). Table 5 in Appendix B defines the symbols in this section.

5.1. Overview

B&B Algorithm 1 spatially branches over the $[\mathbf{v}^L, \mathbf{v}^U]$ domain. It selects a variable x_i , a point v and splits interval $[v_i^L, v_i^U]$ into intervals $[v_i^L, v]$ and $[v, v_i^U]$. Each interval corresponds to an independent subproblem and a new B&B node. To avoid redundant branches, all GBT splits define the B&B branching points. At a given node, denote the reduced node domain by $S = [\mathbf{L}, \mathbf{U}]$. Algorithm 1 solves Problem (1) by relaxing the Equation (4) linking constraints and thereby separating the convex and GBT parts. Using this separation, Algorithm 1 computes corresponding bounds $b^{\text{cvx}, S}$ and $b^{\text{GBT}, S, P}$ independently, where the latter bound requires a tree ensemble partition P initialized at the root node and dynamically refined at each non-root node.

Algorithm 1 Branch-and-Bound (B&B) Algorithm Overview

```

1:  $S = [\mathbf{L}, \mathbf{U}] \leftarrow [\mathbf{v}^L, \mathbf{v}^U]$ 
2:  $b^{\text{cvx}, S} \leftarrow \text{CONVEXBOUND}(S)$  ▷ Lemma 1, Section 5.2.1
3:  $P_{\text{root}} \leftarrow \text{ROOTNODEPARTITION}(N)$  ▷ Section 5.2.2
4:  $b^{\text{GBT}, S, P_{\text{root}}} \leftarrow \text{GBTBOUND}(S, P_{\text{root}})$  ▷ Lemma 2, Section 5.2.2
5:  $B \leftarrow \text{BRANCHORDERING}()$  ▷ Section 5.3.1
6:  $Q = \{S\}$ 
7: while  $Q \neq \emptyset$  do
8:   Select  $S \in Q$ 
9:   if  $S$  is not leaf then
10:      $S' \leftarrow S$ 
11:     repeat
12:        $S', (x_i, v) \leftarrow \text{STRONGBRANCH}(S', B)$  ▷ Algorithm 3, Section 5.3.2
13:     until strong branch not found
14:     if  $S'$  is not leaf then
15:        $(S_{\text{left}}, S_{\text{right}}) \leftarrow \text{BRANCH}(S', (x_i, v))$ 
16:        $P$ : tree ensemble partition of node  $S$ 
17:        $P' \leftarrow \text{PARTITIONREFINEMENT}(P)$  ▷ Algorithm 2, Section 5.2.2
18:        $b^{\text{GBT}, S', P'} \leftarrow \text{GBTBOUND}(S', P')$  ▷ Lemma 2, Section 5.2.2
19:       for  $S_{\text{child}} \in \{S_{\text{left}}, S_{\text{right}}\}$  do
20:         if  $S_{\text{child}}$  cannot be pruned then ▷ Section 5.2.3
21:            $Q \leftarrow Q \cup \{S_{\text{child}}\}$ 
22:         end if
23:       end for
24:     end if
25:   end if
26:    $Q \leftarrow Q \setminus \{S\}$ 
27: end while

```

Algorithm 1 begins by constructing the root node, computing a global lower bound, and determining a global ordering of all branches (lines 1–5). A given iteration: (i) extracts a node S from the unexplored node set Q , (ii) strong branches at S to cheaply identify

branches that tighten the domain resulting in node S' , (iii) updates the GBT lower bound at S' , (iv) branches to obtain the child nodes S_{left} and S_{right} , (v) assesses if each child node $S_{\text{child}} \in \{S_{\text{left}}, S_{\text{right}}\}$ may now be pruned and, if not, (vi) adds S_{child} to the unexplored node set Q (lines 8–25).

The remainder of this section is structured as follows. Section 5.2 lower bounds Problem (1). Section 5.3 introduces a GBT branch ordering and leverages strong branching for cheap node pruning. Section 5.4 discusses heuristics for computing efficient upper bounds.

5.2. Lower Bounding

5.2.1. Global lower bound The convex MINLP Problem (2) objective function consists of a *convex* (penalty) part and a *mixed-integer linear* (GBT) part. Lemma 1 computes a lower bound on the problem by handling the convex and GBT parts independently.

LEMMA 1. *Let $S = [\mathbf{L}, \mathbf{U}] \subseteq [\mathbf{v}^L, \mathbf{v}^U]$ be a sub-domain of optimization Problem (2). Denote by R^S the optimal objective value, i.e., the tightest relaxation, over the sub-domain S . Then, it holds that $R^S \geq \hat{R}^S$, where:*

$$\hat{R}^S = \underbrace{\left[\min_{\mathbf{x} \in S} \text{cvx}(\mathbf{x}) \right]}_{b^{\text{cvx}, S}} + \underbrace{\left[\min_{\mathbf{x} \in S} \sum_{t \in \mathcal{T}} \text{GBT}_t(\mathbf{x}) \right]}_{b^{\text{GBT}, S, *}}.$$

Proof Let $\mathbf{x}^* = \arg \min_{\mathbf{x} \in S} \{\text{cvx}(\mathbf{x}) + \text{GBT}(\mathbf{x})\}$ and observe that $\text{cvx}(\mathbf{x}^*) \geq b^{\text{cvx}, S}$ and $\text{GBT}(\mathbf{x}^*) \geq b^{\text{GBT}, S, *}$. \square

We may compute \hat{R}^S by removing the Equation (4) linking constraints and solving the mixed-integer model consisting of Equations (2) and (3). Computationally, the Lemma 1 separation leverages efficient algorithms for the convex part and commercial codes for the MILP GBT part. Lemma 1 treats the two Problem (1) objective terms independently, i.e., \hat{R}^S separates the convex and GBT parts. The Lemma 1 separation, while loose at the root node, may be leveraged to discard regions that are dominated by an objective term. Our approach resembles exact algorithms for multiobjective optimization (Fernández and Tóth 2009, Niebling and Eichfelder 2016, 2019). An alternative approach, e.g., in line with augmented Lagrangian methods for stochastic optimization (Bertsekas 2014), would not separate the convex penalty term as in Lemma 1, but rather tighten the lower bound by integrating the convex penalty and GBTs. This would be an interesting alternative, but would eliminate the possibility of the strong branching method used in Section 5.3.2.

5.2.2. GBT Lower Bound While we may efficiently compute $b^{\text{cvx},S}$ (Boyd and Vandenberghe 2004), deriving $b^{\text{GBT},S,*}$ is \mathcal{NP} -hard (Mišić 2017). With the aim of tractability, we calculate a relaxation of $b^{\text{GBT},S,*}$. Lemma 2 lower bounds Problem (3), i.e., the GBT part of Problem (2), by partitioning the GBT ensemble into a collection of smaller ensembles.

LEMMA 2. *Consider a sub-domain $S = [\mathbf{L}, \mathbf{U}] \subseteq [\mathbf{v}^L, \mathbf{v}^U]$ of the optimization problem. Let $P = \{\mathcal{T}_1, \dots, \mathcal{T}_k\}$ be any partition of \mathcal{T} , i.e., $\cup_{i=1}^k \mathcal{T}_i = \mathcal{T}$ and $\mathcal{T}_i \cap \mathcal{T}_j = \emptyset \forall 1 \leq i < j \leq k$. Then, it holds that $b^{\text{GBT},S,*} \geq b^{\text{GBT},S,P}$, where:*

$$b^{\text{GBT},S,P} = \sum_{\mathcal{T}' \in P} \left[\min_{\mathbf{x} \in S} \left\{ \sum_{t \in \mathcal{T}'} \text{GBT}_t(\mathbf{x}) \right\} \right].$$

Proof When evaluating $\text{GBT}(\mathbf{x})$ at a given \mathbf{x} , each tree $t \in \mathcal{T}$ provides its own independent contribution $\text{GBT}_t(\mathbf{x})$, i.e., a single leaf. A feasible selection of leaves has to be consistent with respect to the GBT node splits, i.e., if one leaf splits on $x_i < v_1$ and another splits on $x_i \geq v_2$ then $v_1 > v_2$. Relaxing this consistency requirement by considering a partition P of \mathcal{T} derives the lower bounds $b^{\text{GBT},S,P}$ for any partition P . \square

Root Node Partition B&B Algorithm 1 chooses an initial root node partition P_{root} with subsets of size N and calculates the associated Lemma 2 lower bound. Section 7 numerically decides the partition size N for the considered instances. The important factors for a subset size N are the tree depth, the number of continuous variable splits and their relation with the number of binary variables.

Non-Root Node Partition Refinement Any non-root B&B node has reduced domain $\mathbf{x} \in S = [\mathbf{L}, \mathbf{U}] \subset [\mathbf{v}^L, \mathbf{v}^U]$. B&B Algorithm 1 only branches on GBT node splits, so modeling the reduced domain S in MILP Problem (3) is equivalent to setting $y_{i,j} = 0$ or $y_{i,j} = 1$ for any $y_{i,j}$ that corresponds to $x_i \leq L_i$ or $x_i \geq U_i$, respectively. Reducing the box-constrained domain at the node level equates to reducing the GBT instance size. In particular, we may reduce the number and height of trees by assigning fixed variable values and cancelling redundant constraints (Mišić 2017).

Assume that, at some non-root node with domain S , the algorithm is about to update $b^{\text{GBT},S',P'}$ which was calculated at the parent node with domain $S' \supset S$. Fixing binary variables $y_{i,j}$ subject to domain S reduces the worst case enumeration cost of calculating $b^{\text{GBT},S,P'}$. The GBT lower bound may further improve at S by considering an alternative partition P such that $|P| < |P'|$, i.e., reducing the number of subsets. However, reducing the

number of subsets has challenges because: (i) choosing any partition P does not necessarily guarantee $b^{\text{GBT},S,P} \geq b^{\text{GBT},S',P'}$, and (ii) a full Lemma 2 calculation of $b^{\text{GBT},S,P}$ may still be expensive when considering the cumulative time across all B&B nodes. Refinability Definition 2 addresses the choice of P such that $b^{\text{GBT},S,P} \geq b^{\text{GBT},S',P'}$.

DEFINITION 2. Given two partitions P' and P'' of set \mathcal{T} , we say that P' *refines* P'' if and only if $\forall \mathcal{T}' \in P', \exists \mathcal{T}'' \in P''$ such that $\mathcal{T}' \subseteq \mathcal{T}''$. This definition of refinement implies a partial ordering between different partitions of \mathcal{T} . We express the refinement relation by \preceq , i.e., $P' \preceq P''$ if and only if P' refines P'' .

EXAMPLE 2. Let $P = \{\{1, 2, 3\}, \{4, 5\}\}$, $P' = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$ and $P'' = \{\{1, 2\}, \{3, 4, 5\}\}$ be partitions of $\{1, \dots, 5\}$. Here P' refines P since every subset in P' is a subset of one of the P subsets. Similarly P' refines P'' . Partition P does not refine P'' nor does P'' refine P .

Lemma 3 allows bound tightening by partition refinements. Its proof is similar to Lemma 2.

LEMMA 3. Let P and P' be two partitions of \mathcal{T} . If $P' \preceq P$, then $b^{\text{GBT},P'} \leq b^{\text{GBT},P}$.

In general, for two partitions P and P' , we do not know a priori which partition results in a superior GBT lower bound. However, by Lemma 3, P' refining P suffices for $b^{\text{GBT},P} \geq b^{\text{GBT},P'}$. Therefore, given partition P' for the parent node, constructing P for the child node S by unifying subsets of P' will not result in inferior lower bounds.

Algorithm 2 improves $b^{\text{GBT},S',P'}$ at node S by computing a refined partition P . Suppose that $P' = \{\mathcal{T}_1, \dots, \mathcal{T}_k\}$. Each GBT ensemble subset $\mathcal{T}' \in P'$ corresponds to a smaller subproblem with $n^{\mathcal{T}',S}$ leaves ($z_{t,l}$ variables) over the domain S . Initially, Algorithm 2 sorts the subsets of P' in non-decreasing order of $n^{\mathcal{T}',S}$. Then, it iteratively takes the union of consecutive pairs and calculates the associated lower bound, i.e., the first calculation is for $b^{\text{GBT},S,\{\mathcal{T}_1 \cup \mathcal{T}_2\}}$, the second is for $b^{\text{GBT},S,\{\mathcal{T}_3 \cup \mathcal{T}_4\}}$ and so forth. The iterations terminate when all unions have been recalculated, or at user defined time limit q resulting in two sets of bounds: those that are combined and recalculated, and those that remain unchanged. Assuming that the final subset that is updated has index $2l$, the new partition of the trees at node S is $P = \{\mathcal{T}_1 \cup \mathcal{T}_2, \dots, \mathcal{T}_{2l-1} \cup \mathcal{T}_{2l}, \mathcal{T}_{2l+1}, \dots, \mathcal{T}_k\}$ with GBT bound $b^{\text{GBT},S,P} = \sum_{i=1}^l b^{\text{GBT},S,\{\mathcal{T}_{2i-1} \cup \mathcal{T}_{2i}\}} + \sum_{i=2l+1}^k b^{\text{GBT},S',\{\mathcal{T}_i\}}$. The second sum is a result of placing time limit q on updating the GBT lower bound. Time limit q maintains a balance between searching

Algorithm 2 Non-Root Node Partition Refinement

```

1:  $P'$ : parent node partition
2: Sort  $P' = \{\mathcal{T}_1, \dots, \mathcal{T}_k\}$  so that  $n^{\mathcal{T}_1} \leq \dots \leq n^{\mathcal{T}_k}$ 
3:  $P \leftarrow \emptyset$ 
4:  $i = 1$ 
5: while  $i < \lfloor n/2 \rfloor$  and the time limit is not exceeded do
6:    $P \leftarrow P \cup \{\mathcal{T}_{2i-1} \cup \mathcal{T}_{2i}\}$ 
7:    $i \leftarrow i + 1$ 
8: end while
9:  $P \leftarrow P \cup \{\mathcal{T}_j \in P' : j > i\}$ 
10: return  $P$ 

```

and bounding. Unifying any number of subsets satisfies Lemma 3, but Algorithm 2 unifies pairs to keep the resulting subproblems manageable. One may speed up our lower bounding procedure by reducing the height of the GBTs, thus relaxing feasibility, and converting each partition subset \mathcal{T}_k solution into a feasible one for \mathcal{T}_k using the Mišić (2017) split generating procedure for fixing violated constraints.

5.2.3. Node Pruning In the B&B algorithm, each node can access: (i) the current best found feasible objective f^* , (ii) a lower bound on the convex penalties $b^{\text{cvx},S}$, and (iii) a lower bound on the GBT part $b^{\text{GBT},S}$. The algorithm prunes node S if:

$$b^{\text{cvx},S} + b^{\text{GBT},S} > f^*, \quad (5)$$

i.e., if all feasible solutions in S have objective inferior to f^* .

5.3. Branching

5.3.1. Branch Ordering Next branch selection is a critical element of B&B Algorithm 1. Each branch is a GBT split (x_i, v) choice and eliminates a certain number of GBT leaves. Branching with respect to a GBT split that covers a larger number of leaves may lead to a smaller number of subsequent B&B iterations by reducing the GBT size.

Selecting a (x_i, v) split that most improves the GBT lower bound is challenging as it may require solving multiple expensive MILPs. So, we heuristically approximate objective improvement by quantifying splits that (i) occur often among all trees and (ii) influence a larger number of leaves in participating trees. Let $r((x_i, v), t)$ and $\text{cover}(s, t)$ return the

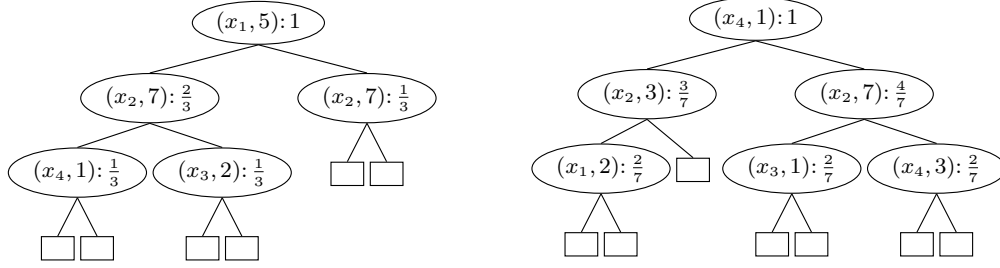


Figure 3 Example 3 node contributions to Equation (6) weight calculation. Each split node contains $'(x_i, v) : w'$ where (x_i, v) is the split pair and w is the node's contribution to (x_i, v) 's weight. We calculate w as the proportion of leaves covered relative to the total number of leaves.

set of nodes in tree t that split on (x_i, v) and the set of leaves that node $s \in \mathcal{V}_t$ covers, respectively. We initialize pseudocosts by weighting the (x_i, v) splits:

$$\text{weight}((x_i, v), t) = |\mathcal{L}_t|^{-1} \sum_{s \in r((x_i, v), t)} |\text{cover}(s, t)|, \quad (6a)$$

$$\text{weight}((x_i, v), \mathcal{T}) = \sum_{t \in \mathcal{T}} \text{weight}((x_i, v), t). \quad (6b)$$

Equation (6a) weights (x_i, v) as the fraction of leaves covered by nodes splitting on (x_i, v) in tree t . Recall that $|\mathcal{L}_t|$ is the number of leaf nodes in tree t . Equation (6b) sums all weights calculated by Equation (6a) for split (x_i, v) in each tree $t \in \mathcal{T}$. The splits are sorted in non-increasing order their pseudocosts.

EXAMPLE 3. Figure 3 shows the weight given to each node for two trees. The left tree contains 6 leaves and the right tree contains 7 leaves. Consider split $(x_2, 7)$. The left tree contains two nodes splitting on $(x_2, 7)$ one of which covers 4 out of 6 leaves and the other covers 2 out of 6 leaves therefore these nodes contribute $\frac{2}{3}$ and $\frac{1}{3}$, respectively, to the weight. Similarly, the right tree contains a single node splitting on $(x_2, 7)$ which covers 4 out of 7 leaves therefore this node contributes $\frac{4}{7}$ to the weight. We obtain the weight for $(x_2, 7)$ by summing these values, i.e., $\text{weight}((x_2, 7), \mathcal{T}) = \frac{2}{3} + \frac{1}{3} + \frac{4}{7} = 1\frac{4}{7}$.

The Equation (6) weight function initializes pseudocosts satisfying the following properties:

1. for each tree t , $\text{weight}((x_i, v), t)$ is proportional to $\sum_{s \in r((x_i, v), t)} |\text{cover}(s, t)|$,
2. if (x_i, v) and $(x_{i'}, v')$ cover the same set of leaves in tree t then $\text{weight}((x_i, v), t) = \text{weight}((x_{i'}, v'), t)$.

5.3.2. Strong Branching Branch selection is fundamental to any B&B algorithm. *Strong branching* selects a branch that enables pruning with low effort computations and achieves a non-negligible speed-up in the algorithm's performance (Morrison et al. 2016).

Algorithm 3 Strong Branching

```

1:  $S$ : B&B node with bounds  $b^{\text{GBT},S}$  and  $b^{\text{cvx},S}$ 
2:  $B^S = [(x_{i_1}, v_1), \dots, (x_{i_l}, v_l)]$ :  $l$  next branches list w.r.t. Section 5.3.1 pseudo-cost order
3: for  $(x_i, v) \in B^S$  do
4:    $S_{\text{left}}, S_{\text{right}}$ :  $S$  children by branching on  $(x_i, v)$ 
5:   Compute  $b^{\text{cvx},S_{\text{left}}}$  and  $b^{\text{cvx},S_{\text{right}}}$ 
6:   if  $\max\{b^{\text{cvx},S_{\text{left}}}, b^{\text{cvx},S_{\text{right}}}\} + b^{\text{GBT},S} < f^*$  then
7:     return  $\arg \min\{b^{\text{cvx},S_{\text{left}}}, b^{\text{cvx},S_{\text{right}}}\}, (x_i, v)$ 
8:   end if
9: end for
10: return  $S, (x_{i_1}, v_1)$ 

```

Strong branching increases the size of efficiently solvable large-scale mixed-integer problems and is a major solver component (Klabjan et al. 2001, Anstreicher et al. 2002, Anstreicher 2003, Easton et al. 2003, Belotti et al. 2009, Misener and Floudas 2013, Kılınç et al. 2014). Here, strong branching leverages the easy-to-solve convex penalty term for pruning.

At a B&B node S , branching produces two children S_{left} and S_{right} . Strong branching Algorithm 3 considers the branches in their Section 5.3.1 pseudo-cost ordering and assesses each branch by computing the associated convex bound. Under the strong branching test, one node among S_{left} and S_{right} inherits the convex bound $b^{\text{cvx},S}$ from the parent, while the other requires a new computation. Suppose that $S' \in \{S_{\text{left}}, S_{\text{right}}\}$ does not inherit $b^{\text{cvx},S}$. If $b^{\text{cvx},S'}$ satisfies the Equation (5) pruning condition without GBT bound improvement, then S' is immediately selected as the strong branch and strong branching repeats at the other child node S'' . Figure 4 illustrates strong branching. When Algorithm 1 does not find a strong branch, it performs a GBT lower bound update and branches on the first item of the branch ordering. Algorithm 1 then adds this node's children to a set of unexplored nodes and continues with the next B&B iteration.

Strong branching allows efficient pruning when the convex objective part is significant. Strong branching may reduce the computational overhead incurred by GBT bound recalculation when Algorithm 3 selects multiple strong branches between GBT bound updates. While a single strong branch assessment is negligible, the cumulative cost of calculating convex bounds for all branches may be high. Section 5.3.1 orders the branches according

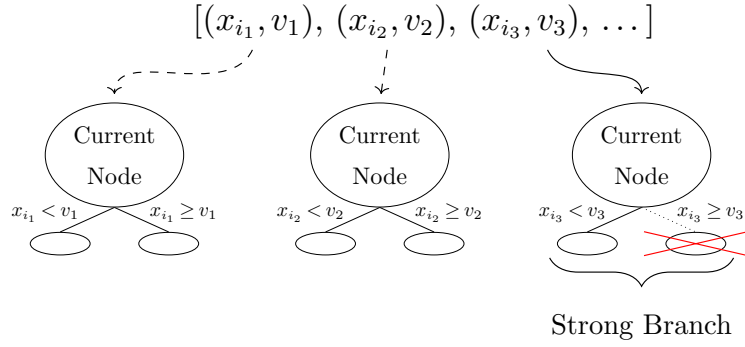


Figure 4 Strong branching for selecting the next spatial branch. A strong branch leads to a node that is immediately pruned, based on a convex bound computation.

to a measure of effectiveness aiding GBT bounding, so the time spent deriving strong branches with small weighting function may be better utilized in improving the GBT bound. Opposed to full strong branching, i.e., assessing all branches, strong branching Algorithm 3 uses a lookahead approach (Achterberg et al. 2005). Parameterized by a lookahead value $l \in \mathbb{Z}_{>0}$, Algorithm 3 investigates the first l branches. If Algorithm 3 finds a strong branch, Algorithm 1 repeats Algorithm 3, otherwise the B&B Algorithm 1 updates the GBT bound $b^{\text{GBT},S,P}$ at the current node. Algorithm 3 keeps strong branching checks relatively cheap and maintains a balance between searching and bounding.

5.4. Heuristics

To prune, i.e., satisfy Equation (5), consider two heuristic methods generating good feasible solutions to Problem (1): (i) a mixed-integer convex programming (convex MINLP) approach, and (ii) particle swarm optimization (PSO) (Eberhart and Kennedy 1995, Kennedy and Eberhart 1995). The mixed-integer approach uses the decomposability of GBT ensembles, i.e., while convex MINLP solvers provide weak feasible solutions for large-scale instances of Problem (1), they may efficiently solve moderate instances to global optimality (Westerlund and Pettersson 1995, Tawarmalani and Sahinidis 2005, Vigerske 2012, Misener and Floudas 2014, Lundell et al. 2017). The PSO approach exploits trade-offs between the convex and objective GBT parts. Metaheuristics like particle swarm optimization and simulated annealing (Kirkpatrick et al. 1983) may produce heuristic solutions in preprocessing, i.e., before the branch-and-bound algorithm begins. Simpler convex MINLP heuristics may improve upper bounds at a branch-and-bound node because of their efficient running times. Appendix C in the electronic companion discusses these heuristics.

6. Case Studies: Principal Component Analysis for Penalizing Solutions far from Training Data

Our case studies consider GBT instances where training data is not evenly distributed over the $[\mathbf{v}^L, \mathbf{v}^U]$ domain. So, while $\mathbf{x} \in [\mathbf{v}^L, \mathbf{v}^U]$ is feasible, $\text{GBT}(\mathbf{x})$ may be less meaningful for \mathbf{x} far from training data. The Problem (1) $\text{cvx}(\mathbf{x})$ function, for the case studies, is a penalty function constructed with principal component analysis (PCA) (Jolliffe 2002).

PCA characterizes a large, high-dimensional input data set $D = \{\mathbf{d}^{(1)}, \dots, \mathbf{d}^{(p)}\}$ with a low-dimensional subspace capturing most of the variability (James et al. 2013). PCA defines a set of n ordered, orthogonal *loading vectors*, ϕ_i , such that ϕ_i captures more variability than $\phi_{i'}$, for $i < i'$. PCA on D defines parameters $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^n$ and $\boldsymbol{\Phi} = [\phi_1 \dots \phi_n] \in \mathbb{R}^{n \times n}$, i.e., the sample mean, sample standard deviation and loading vectors, respectively. Vectors $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ standardize D since PCA is sensitive to scaling. Often, only a few ($k < n$) leading loading vectors capture most of the variance in D and $\boldsymbol{\Phi}' = [\phi_1 \dots \phi_k]$ may effectively replace $\boldsymbol{\Phi}$. $\mathbf{P} = \boldsymbol{\Phi}' \boldsymbol{\Phi}'^\top$ defines a projection matrix to the subspace spanned by $\{\phi_1, \dots, \phi_k\}$. Penalizing solutions further from training data with PCA defined projection matrix \mathbf{P} :

$$\text{cvx}_\lambda(\mathbf{x}) = \lambda \left\| (\mathbf{I} - \mathbf{P}) \text{diag}(\boldsymbol{\sigma})^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\|_2^2 \quad (7)$$

where $\lambda > 0$ is a penalty parameter, \mathbf{I} is the identity matrix and $\text{diag}(\cdot)$ is a matrix with the argument on the diagonal. Larger λ is more conservative with respect to PCA subspace \mathbf{P} . Note in Equation (7) that our specific nonlinear convex penalty is a convex quadratic.

Equation (7) aims to characterize the region containing the training data with an affine subspace. Points in the subspace are not penalized and points close to the subspace are not heavily penalized. However, Equation (7) may be qualitatively less effective when the standardized training data is not evenly distributed within subspace \mathbf{P} .

EXAMPLE 4. Consider a data set $\{\mathbf{x}^{(i)}\}_{i=1}^{2m}$, $\mathbf{x}^{(i)} \in \mathbb{R}^3$ where $x_1^{(i)} \sim U(0, 1)$, $\forall i \in [2m]$, and $x_2^{(i)} = x_3^{(i)} = 0$, $x_2^{(m+i)} = x_3^{(m+i)} = 1$, $\forall i \in [m]$. The 2D subspace containing these points contains the origin and directions $(1, 0, 0)^T, (0, 1, 1)^T$. Equation (7) does not penalize points in this subspace. But the point $(0.5, 0.5, 0.5)$, which is contained in the subspace, is far from the training data when considering the subspace distribution. Having $x_2^{(i)}, x_3^{(i)} \sim N(0, \varepsilon)$, $x_2^{(m+i)}, x_3^{(m+i)} \sim N(1, \varepsilon)$, $\forall i \in [m]$ and small $\varepsilon > 0$, introduces an error term to the second and third variables while retaining the same clustered distribution over the subspace.

Table 2 Instance Sizes

	Concrete Mixture Design	Chemical Catalysis
<i>GBT attributes:</i>		
Number of trees	7,750	8,800
Maximum depth	16	16
Number of leaves	131,750	93,200
Number of x_i continuous variables	8	42
<i>Convex MINLP (2) attributes:</i>		
Number of $y_{i,j}$ binary variables	8,441	2,061
Number of constraints	281,073	183,791

Clustering, e.g., Example 4, may be handled by the Section 5 B&B. We could instantiate a separate instance for each cluster using a penalty that only considers training data in a given cluster and limit the solve to a reduced box domain. A single problem formulation considering more complex training data relationships may negatively affect the strong branching aspect of B&B Algorithm 1.

7. Numerical Results

This section compares the Section 5 lower bounding and branch-and-bound algorithms to black-box solvers. Appendix D of the electronic companion presents results for the Section 5.4 heuristics. Section 7.1 provides information about the system specifications and the solvers. Sections 7.2 and 7.3 investigate two GBT instances for engineering applications, namely: (i) concrete mixture design and (ii) chemical catalysis. Section 7.4 discusses observations from the Sections 7.2 and 7.3 results. The concrete mixture design instance is from the UCI machine learning repository (Dheeru and Karra Taniskidou 2017). The industrial chemical catalysis instance is provided from BASF. Table 2 presents information about these instances. For both instances, we model closeness to training data using the PCA-based function $\text{cvx}(\mathbf{x})$ defined in Equation (7).

7.1. System and Solver Specifications

Experiments are run on an Ubuntu 16.04 HP EliteDesk 800 G1 TWR with 16GB RAM and an Intel Core i7-4770@3.40GHz CPU. Implementations are in Python 3.5.3 using Pyomo 5.2 (Hart et al. 2011, 2017) for mixed-integer programming modeling and interfacing with solvers. We use CPLEX 12.7 and Gurobi 7.5.2 as: (i) black-box solvers for the entire convex MINLP (2), and (ii) branch-and-bound algorithm components for solving MILP (3) instances in the Section 5.2 GBT lower bounding procedure. Note that current versions of

CPLEX and Gurobi cannot solve general convex MINLP, so we would use a more general solver if we had non-quadratic penalty functions. All results report wall clock times.

This section evaluates the (i) objective lower bounding procedure, and (ii) branch-and-bound algorithm, both of which use CPLEX or Gurobi as a black-box MILP solver. We also apply CPLEX and Gurobi to the entire MINLP for evaluating branch-and-bound Algorithm 1. Figures 5-12 append labels -C and -G to indicate CPLEX and Gurobi, respectively, and use different line types for displaying the results. At nodes immediately following a GBT bound update, the B&B algorithm assesses solutions from solving the convex part of Problem (1) as heuristics solutions. We use the default CPLEX 12.7 and Gurobi 7.5.2 tolerances, i.e., relative MIP gap, integrality and barrier convergence tolerances of 10^{-4} , 10^{-5} and 10^{-8} , respectively.

7.2. Concrete Mixture Design

In concrete mixture design, different ingredient proportions result in different properties of the concrete, e.g., compressive strength. The relationship between ingredients and properties is complex, so black-box machine learning is well suited for the function estimation task (Chou et al. 2011, Erdal 2013, DeRousseau et al. 2018).

7.2.1. Instance We maximize concrete compressive strength where GBTs are used for modeling. Since we maximize concrete compressive strength, negating all leaf weights $F_{t,l}$ forms an equivalent GBT instance that fits the Problem (1) minimization formulation. We use the Yeh (1998) concrete compressive strength dataset from the UCI machine learning repository (Dheeru and Karra Taniskidou 2017). This dataset has $n = 8$ continuous variables. R packages `gbm` (Ridgeway 2017) and `caret` (Kuhn 2008) are used for GBT training. Root-mean-square error is used for model selection. The resulting GBT instance has 7750 trees with max depth 16. The PCA based convex penalty has $\text{rank}(\mathbf{P}) = 4$, i.e., we select the first four loading vectors. Section D.1 of the electronic companion presents the results for the Section 5.4 heuristics.

7.2.2. GBT Lower Bounding Figures 5 and 6 evaluate the Section 5.2.2 GBT lower bounding approach for different partition subset sizes. Figure 5 illustrates the global GBT lower bound improvement as the partition subset size increases. Figure 6 compares run times with either CPLEX 12.7, or Gurobi 7.5.2 as subsolvers for each partition subset size. For the entire MILP instance, i.e., solving Problem (3), black-box solving with CPLEX

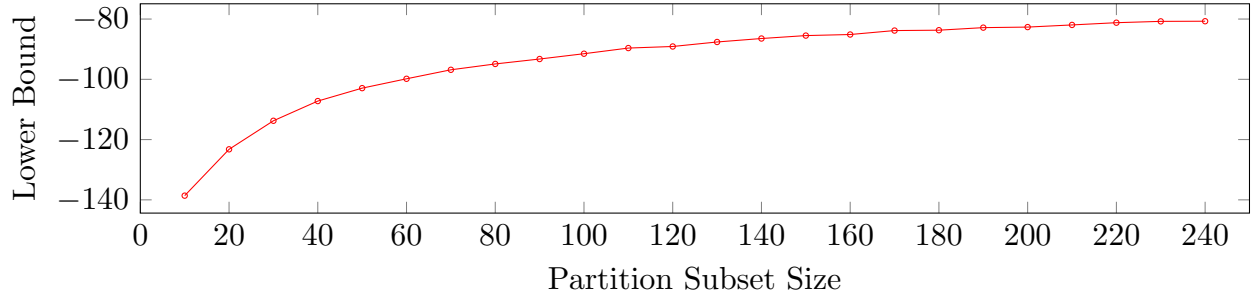


Figure 5 Concrete mixture design instance: Global GBT lower bound improvement using the Section 5.2 GBT lower bounding approach for different partition subset sizes.

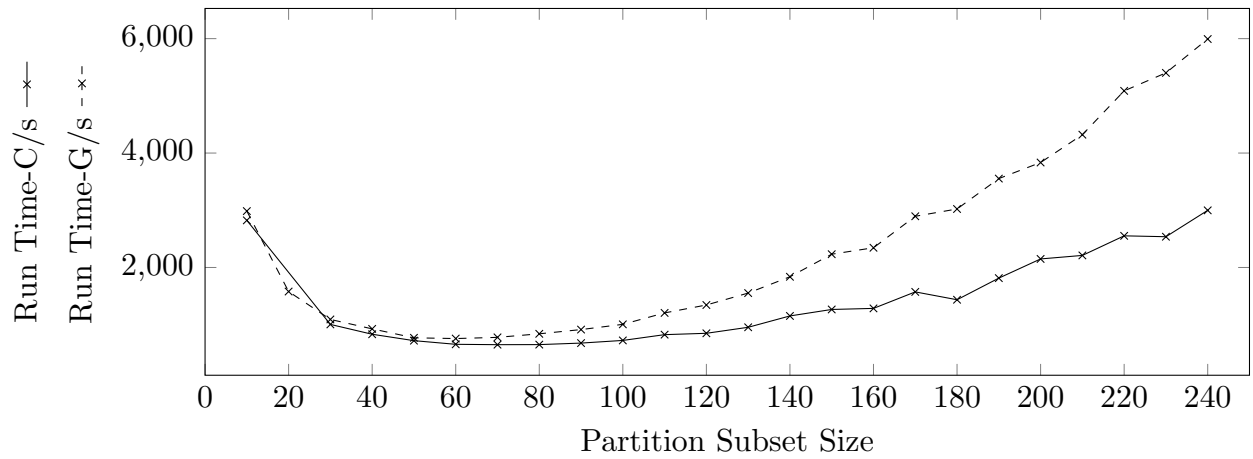


Figure 6 Concrete mixture design instance: Global GBT lower bounding wall clock time using the Section 5.2 approach for different partition subset sizes. Suffixes -C and -G denote subsolvers CPLEX 12.7 and Gurobi 7.5.2, respectively.

12.7 and Gurobi 7.5.2 achieve GBT lower bounds -97 and -547, respectively, within 1 hour. The Section 5.2.2 approach achieves a lower bound of -83 (partition size 190), in 1 hour, and improves upon black-box solver lower bounds in under 15 minutes (partition size 70).

7.2.3. Branch-and-Bound Algorithm We instantiate the branch-and-bound algorithm with a root node partition of 70 trees, and non-root lower bounding time limit of 120 seconds. All branch-and-bound tests are run with CPLEX 12.7 and Gurobi 7.5.2 as MILP subsolvers. We assess the effect of strong branching by comparing lookahead list sizes $l = 1$ vs. $l = 100$. We assess the quality of feasible solutions by comparing with the Table 6 [electronic companion] best found feasible solution. We assess the pseudocost ordering by comparing with 10 independent tests of random branch orderings for each strong branch lookahead-solver combination. We compare all branch-and-bound results, which allo-

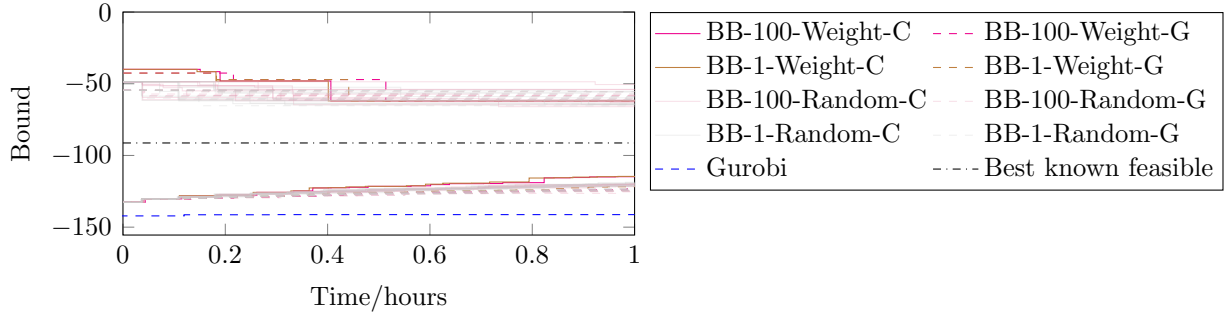


Figure 7 Concrete mixture design instance ($\lambda = 1$): B&B lower bound improvement compared to Gurobi 7.5.2, with a one hour timeout. The B&B Algorithm 1 is labeled BB- a - b - c where a , b and c denote the strong branching lookahead value, the pseudocost initialization approach, and the solver used for lower bounding and solving convex quadratics, respectively. The BB-* results sort the unexplored nodes in ascending lower bound order. The dashed-dotted line reports best found feasible solution (upper bound).

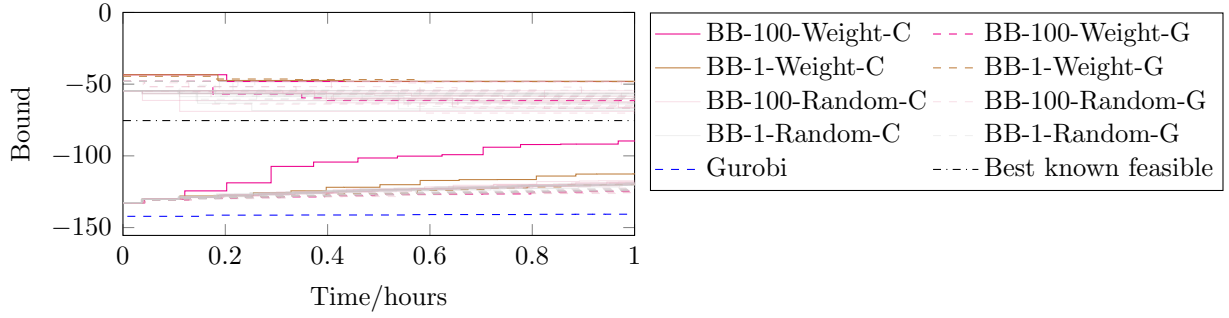


Figure 8 Concrete mixture design instance ($\lambda = 1000$): B&B lower bound improvement compared to Gurobi 7.5.2, with a one hour timeout. The B&B Algorithm 1 is labeled BB- a - b - c where a , b and c denote the strong branching lookahead value, the pseudocost initialization approach, and the solver used for lower bounding and solving convex quadratics, respectively. The BB-* results sort the unexplored nodes in ascending lower bound order. The dashed-dotted line reports best found feasible solution (upper bound).

cate 1 hour for GBT lower bounding at the root node and 1 hour for the B&B search, to 3 hour black-box runs of CPLEX 12.7 and Gurobi 7.5.2 for the entire convex MINLP.

Figures 7 and 8 plot the bound improvement for $\lambda = 1$ and $\lambda = 1000$, respectively. For the entire convex MINLP, the black-box CPLEX 12.7 bounds are outside the figure axis limits. For $\lambda = 1$, a larger strong branching lookahead value does not noticeably improve the lower bound, but a larger lookahead does significantly improve the lower bound for $\lambda = 1000$. Figure 7 depicts the lower bound improvement. The B&B algorithm lower bound improves over time, but there is still a non-negligible gap from the best-known feasible solution after 1 hour. This gap appears to be due to a cluster-like effect caused by the GBTs (Du and

Table 3 Concrete mixture design instance: Results comparing 24 hour runs of the B&B algorithm with Gurobi 7.5.2. The B&B algorithm uses a strong branching lookahead value of 100, a root node partition of 70 trees, a non-root lower bounding time limit of 120 seconds, and CPLEX 12.7 as a subsolver.

λ	BB-C			Gurobi 7.5.2		
	UB	LB	Gap	UB	LB	Gap
1	-80.56	-99.87	24%	-85.48	-140.75	64%
10	-74.96	-99.39	33%	-85.06	-121.10	42%
100	-73.74	-96.43	31%	-77.98	-121.27	55%
1000	-74.86	-90.75	22%	-72.29	-121.23	67%

Kearfott 1994, Wechsung et al. 2014, Kannan and Barton 2017), where the variable split points are quite close. In the B&B algorithm, if the current lookahead list contains these clusters, strong branching is less effective. CPLEX 12.7 results in an out-of-memory error prior to beginning the branch-and-bound search therefore its lower bounds are relatively poor. Gurobi 7.5.2 returns an incumbent of -85 and a lower bound of -141, after 2 hours, and these do not improve further in the subsequent hour. The B&B algorithm, at 2 hours, i.e., prior to tree search, has an incumbent of -91 and a lower bound not less than -133. Given an additional hour for tree search, the gap reduces further. Table 3 compares the B&B algorithm to Gurobi 7.5.2 with 24 hours time limit. The Gurobi heuristics generally outperform the B&B algorithm, but the B&B algorithm derives better lower bounds. In all cases, $\geq 22\%$ optimality gap remains. Because regions close to training data have many GBT breakpoints, optimal solutions lie in highly discretized areas of the feasibility domain.

7.3. Chemical Catalysis

BASF uses catalysts to improve yield and operating efficiency. But, modeling catalyst effectiveness is highly nonlinear and varies across different applications. BASF has found GBTs effective for modeling catalyst behavior. Capturing the high-dimensional nature of catalysis over the entire feasible domain requires many experiments, too many to run in practice. Running a fewer number of experiments necessitates penalizing solutions further from where the GBT function is trained.

7.3.1. Instance The BASF industrial instance contains $n = 42$ continuous variables. The convex part of the instance takes the following form:

$$\text{cvx}_\lambda(\mathbf{x}) = \lambda \left\| (\mathbf{I} - \mathbf{P}) \text{diag}(\boldsymbol{\sigma})^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\|_2^2 + \left(100 - \sum_{i \in \mathcal{I}^\%} x_i \right)^2 \quad (8)$$

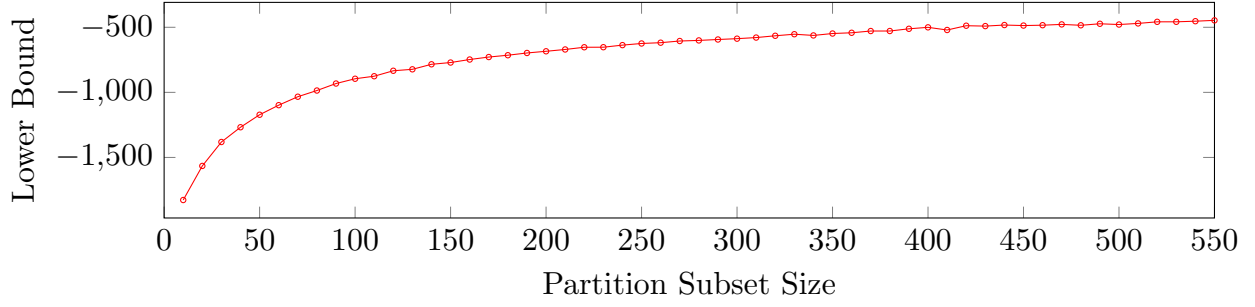


Figure 9 Chemical catalysis BASF instance: Global GBT lower bound improvement using the Section 5.2 GBT lower bounding approach for different partition subset sizes.

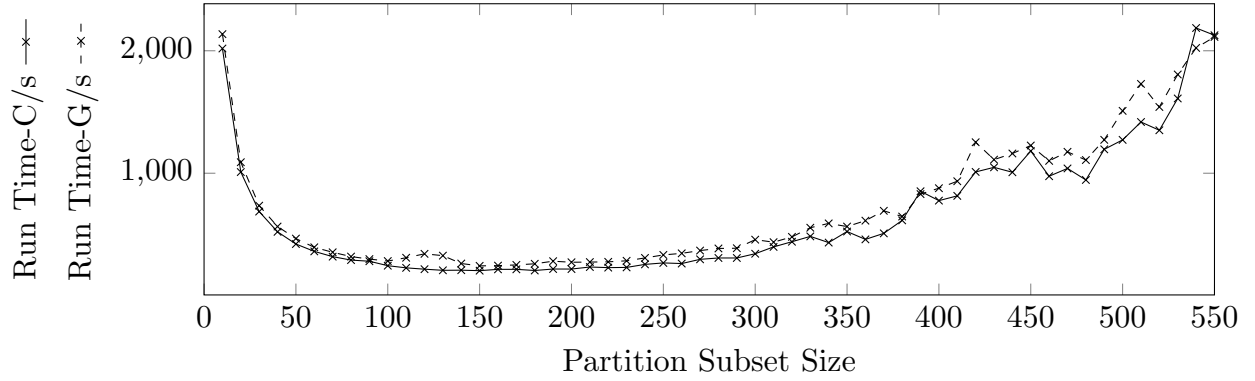


Figure 10 Chemical catalysis BASF instance: Global GBT lower bounding wall clock time using the Section 5.2 approach for different partition subset sizes. Suffixes -C and -G denote subsolvers CPLEX 12.7 and Gurobi 7.5.2, respectively.

Equation (8) differs from Equation (7) in its addend which aims to generate solutions where $x_i \in \mathcal{I}^\%$, i.e., proportions of the chemicals being mixed, sum to 100%. The test instance has $\text{rank}(\mathbf{P}) = 2$ and $|\mathcal{I}^\%| = 37$. The GBT part contains 8800 trees where 4100 trees have max depth 16, the remaining trees have max depth 4, the total number of leaves is 93,200 and the corresponding Problem (3) MILP model has 2061 binary variables. Section D.2 of the electronic companion presents the results for the Section 5.4 heuristics.

7.3.2. GBT Lower Bounding Figures 9 and 10 evaluate the Section 5.2.2 GBT lower bounding approach for different partition subset sizes. Figure 9 illustrates the global GBT lower bound improvement as the partition subset size increases. Figure 10 compares run times when using either CPLEX 12.7, or Gurobi 7.5.2 as subsolvers for each partition subset size. These results resemble Figures 5 and 6. In particular, (i) the lower bound is improved with larger subset sizes, (ii) there is a time-consuming modeling overhead for solving many small MILPs for small subset sizes, and (iii) the running time increases exponentially, though non-monotonically, for larger subset sizes. We compare the lower

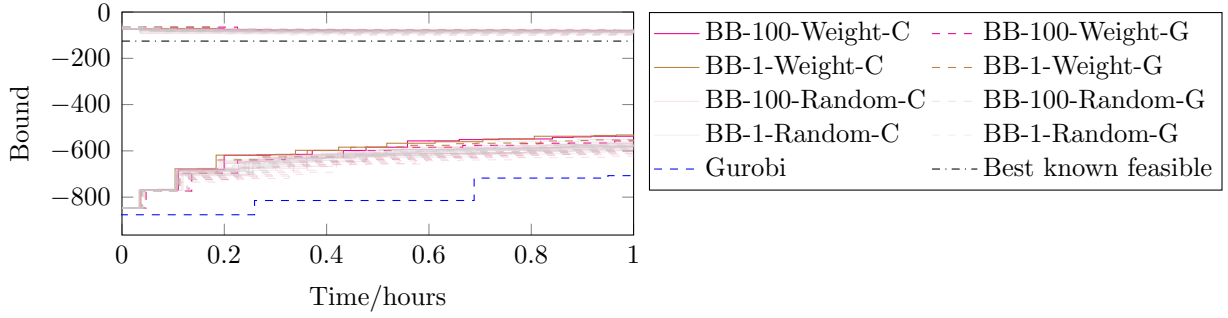


Figure 11 Chemical catalysis BASF instance ($\lambda = 1$): B&B lower bound improvement compared to Gurobi 7.5.2 with one hour timeout. The B&B Algorithm 1 is labeled BB- a - b - c where a , b and c denote the strong branching lookahead value, the pseudocost initialization approach, and the solver used for lower bounding and solving convex quadratics, respectively. The BB-* results sort the unexplored nodes in ascending lower bound order. The dashed-dotted line reports best found feasible solution (upper bound).

bounding approach with solving the entire MILP (3) using CPLEX 12.7, or Gurobi 7.5.2 as black-box solvers. Our lower bounding approach exhibits a superior time-to-lower bound performance: (i) it improves the Gurobi 7.5.2 lower bound with subset size 140 and 4 minutes of execution, and (ii) it improves the CPLEX 12.7 lower bound with subset size 360 and 8 minutes of execution.

7.3.3. Branch-and-Bound Algorithm We instantiate the branch-and-bound algorithm with a root node partition of 150 trees, and non-root lower bounding time limit of 120 seconds. All branch-and-bound tests are run with CPLEX 12.7 and Gurobi 7.5.2 as subsolvers. We assess the effect of strong branching by comparing lookahead list sizes $l = 1$ vs. $l = 100$. We assess the quality of feasible solutions by comparing with the Table 7 [electronic companion] best found feasible solution. We assess the pseudocost ordering by comparing with 10 independent tests of random branch orderings for each strong branch lookahead-subsolver combination. We compare all branch-and-bound results, which allocate 1 hour for GBT lower bounding at the root node and 1 hour for the B&B search, to 3 hour black-box runs of CPLEX 12.7 and Gurobi 7.5.2 for the entire convex MINLP.

Figures 11 and 12 plot the bound improvement for $\lambda = 1$ and $\lambda = 1000$, respectively. For the entire convex MINLP, CPLEX 12.7 reports a poor lower bound and does not find a feasible solution within 3 hours. The B&B algorithm terminates with a tighter lower bound and closes a larger gap than the black-box solvers, across all tested parameter combinations. The B&B algorithm performs better for $\lambda = 1000$ because the convex part dominates the

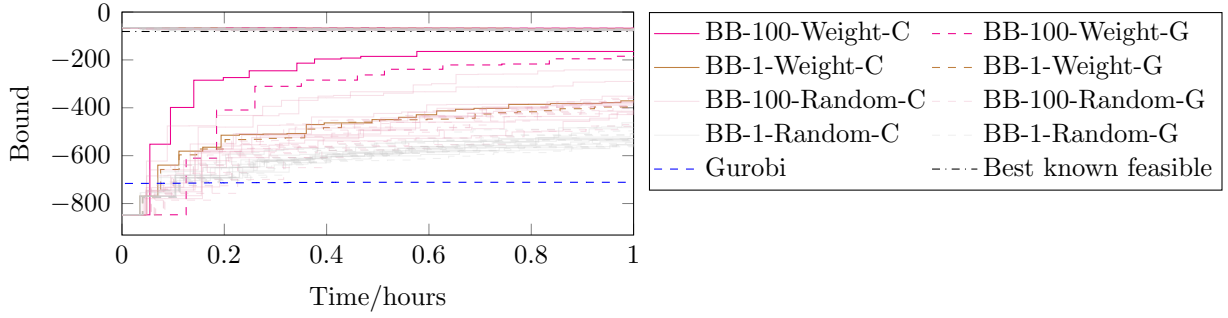


Figure 12 Chemical catalysis BASF instance ($\lambda = 1000$): B&B lower bound improvement compared to Gurobi 7.5.2 with one hour timeout. The B&B Algorithm 1 is labeled BB- a - b - c where a , b and c denote the strong branching lookahead value, the pseudocost initialization approach, and the solver used for lower bounding and solving convex quadratics, respectively. The BB-* results sort the unexplored nodes in ascending lower bound order. The dashed-dotted line reports best found feasible solution (upper bound).

Table 4 Chemical catalysis instance: Results comparing 24 hour runs of the B&B algorithm with Gurobi 7.5.2. The B&B algorithm uses a strong branching lookahead value of 100, a root node partition of 150 trees, a non-root lower bounding time limit of 120 seconds, and CPLEX 12.7 as a subsolver.

λ	BB-C			Gurobi 7.5.2		
	Upper Bound	Lower Bound	Gap	Upper Bound	Lower Bound	Gap
1	-81.7	-366.0	348%	-154.8	-580.6	275%
10	-80.6	-336.8	318%	-118.8	-577.1	386%
100	-87.3	-187.0	114%	-94.2	-424.5	350%
1000	-86.0	-86.0	0%	-85.9	-92.1	7%

GBT part more, making strong branching more effective. Finally, we see that the branch-and-bound algorithm finds a relatively good heuristic solution at the root node for $\lambda = 1000$. For $\lambda = 1$, there is a larger gap between the B&B upper bounds and the best known feasible solution this is expected as solving this problem is closer to optimizing only over the GBT MILP where an optimal solution may be further from the PCA subspace. Table 4 compares the B&B algorithm to Gurobi 7.5.2 with 24 hours time limit. The Gurobi heuristic solutions generally outperform the B&B algorithm. Nevertheless, the B&B algorithm derives better lower bounds. For $\lambda = 1000$, the B&B algorithm succeeds in proving global optimality whereas Gurobi terminates with a 7% gap.

7.4. Observations

The Sections 7.2.2 and 7.3.2 GBT lower bounding results show that, for large-scale GBT instances, selecting an appropriate partition subset size in the decomposition approach results has a better time-to-lower bound performance than 1 hour black-box MILP solvers.

Both problem instances show that, for larger subset sizes, the running time exponentially increases, while the lower bound improvement rate exponentially decreases. This is an expected result for GBT instances with deep trees as deeper tree induce more infeasible combinations of branches. For shallower GBT instances, individual trees to may interact less with each other, hence the decomposition strategy may derive a poorer bound than a black-box MILP solver. For small subset sizes, the partition-based lower bounding has decreasing running time because of the overhead from many sequential subproblems.

The Sections 7.2.3 and 7.3.3 B&B results also show common features. Comparing the BB- \ast - b - \ast results for $b \in \{\text{Weight}, \text{Random}\}$ in Figures 7, 8, 11 and 12 assess the pseudocost effect. The Equation (6) initialization outperforms random ordering (for matching lookahead values), showing that the pseudocosts select branches that aid GBT lower bounding. This pseudocost effect is more pronounced with a lookahead value of 100 since multiple branches are selected between branch-and-bound iterations. For $\lambda = 1000$, a lookahead list size $l = 100$ closes more gap than $l = 1$ (comparing BB-100- \ast to BB-1- \ast), as the B&B algorithm accepts more branches for strong branching. The difference between $l = 100$ and $l = 1$ implies that increased strong branching improves the GBT lower bound earlier and more often. For $\lambda = 1$, using a larger strong branching lookahead size does not have a noticable effect. However, this last finding does not depreciate strong branching. Since the GBT part dominates the convex aspect for small λ values, tighter GBT lower bounds might be essential for taking full advantage of strong branching. Testing the B&B algorithm and Gurobi 7.5.2 with a 24 hour run time shows that the branch-and-bound algorithm tends to result in superior lower bounds and closes a larger proportion of the optimality gap whereas Gurobi 7.5.2 produces better heuristic solutions. Closing any outstanding gap proves difficult as the domains of the remaining unexplored nodes are highly discretized by the GBTs.

8. Discussion

Our optimization problem consists of: (i) the GBTs, and (ii) the PCA-based penalty. Functions obtained from limited, known evaluations with machine learning are approximate by default and may deviate from the ground truth, thus, resulting in false optima. The final solution error depends on the training data distribution, noise, and machine learning model. Our PCA-based approach may deteriorate for clustered data, e.g., Example 4, when regions of the PCA subspace are far from training observations. A remedy is using data

analysis, e.g., clustering (Hastie et al. 2009), to assess uniformity in the training data distribution. An alternative direction is using proximity measures (Liaw and Wiener 2002, Mišić 2017). The proximity measures may require adjusting when using GBTs since the boosting procedure results in some trees being more relevant than others. Finally, other convex penalties are relevant in a variety of applications (Duran and Grossmann 1986).

Finally, we acknowledge other approaches for decision-making with optimization problems whose input is specified by machine learning models. Donti et al. (2017) consider end-to-end task-based learning where probabilistic models are trained to be subsequently used within stochastic programming tasks. Elmachoub and Grigas (2017) develop a framework for training predictive models with a specific loss function so that the resulting optimization problem has desirable convexity properties and is statistically consistent. Wilder et al. (2018) propose a two-stage approach for integrating machine learning predictions with combinatorial optimization problem decisions. The main difference with our work is that we are more focused on the optimization side.

9. Conclusion

As machine learning methods mature, decision makers want to move from solely making predictions on model inputs to integrating pre-trained machine learning models into larger decision-making problems. This paper addresses a large-scale, industrially-relevant gradient-boosted tree model by directly exploiting: (i) advanced mixed-integer programming technology with strong optimization formulations, (ii) GBT tree structure with priority towards searching on commonly-occurring variable splits, and (iii) convex penalty terms with enabling fewer mixed-integer optimization updates. The general form of the optimization problem appears whenever we wish to optimize a pre-trained gradient-boosted tree with convex terms in the objective, e.g., penalties. It would have been alternatively possible to train and then optimize a smooth and continuous machine learning model, but applications with legacy code may start with a GBT. Our numerical results test against concrete mixture design and chemical catalysis, two applications where the global solution to an optimization problem is often particularly useful. Our methods not only generate good feasible solutions to the optimization problem, but they also converge towards proving the exact solution.

Acknowledgments

The support of: BASF SE, the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems to M.M. (EP/L016796/1), and an EPSRC Research Fellowship to R.M. (EP/P016871/1).

References

- Achterberg T, Koch T, Martin A (2005) Branching rules revisited. *Oper. Res. Lett.* 33(1):42–54.
- Anstreicher K, Brixius N, Goux JP, Linderoth J (2002) Solving large quadratic assignment problems on computational grids. *Math. Program.* 91(3):563–588.
- Anstreicher KM (2003) Recent advances in the solution of quadratic assignment problems. *Math. Program.* 97(1):27–42.
- Belotti P, Kirches C, Leyffer S, Linderoth J, Luedtke J, Mahajan A (2013) Mixed-integer nonlinear optimization. *Acta Numer.* 22:1–131.
- Belotti P, Lee J, Liberti L, Margot F, Wächter A (2009) Branching and bounds tightening techniques for non-convex MINLP. *Optim. Method. Softw.* 24(4-5):597–634.
- Bertsekas DP (2014) *Constrained optimization and Lagrange multiplier methods* (Academic press).
- Bienstock D (1996) Computational study of a family of mixed-integer quadratic programming problems. *Math. Program.* 74(2):121–140, ISSN 1436-4646.
- Bollas GM, Barton PI, Mitsos A (2009) Bilevel optimization formulation for parameter estimation in vapor-liquid(-liquid) phase equilibrium problems. *Chem. Eng. Sci.* 64(8):1768–1783.
- Bonami P, Lejeune MA (2009) An exact solution approach for portfolio optimization problems under stochastic and integer constraints. *Oper. Res.* 57(3):650–670.
- Bonfietti A, Lombardi M, Milano M (2015) Embedding decision trees and random forests in constraint programming. *Integration of AI and OR Techniques in Constraint Programming*, 74–90.
- Boukouvala F, Misener R, Floudas CA (2016) Global optimization advances in mixed-integer nonlinear programming, MINLP, and constrained derivative-free optimization, CDFO. *Eur. J. Oper. Res.* 252(3):701–727.
- Boyd S, Vandenberghe L (2004) *Convex optimization* (Cambridge university press).
- Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32, ISSN 1573-0565.
- Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) *Classification & Regression Trees* (Wadsworth).
- Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- Chou JS, Chiu CK, Farfoura M, Al-Taharwa I (2011) Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques. *J. Comput. Civil Eng.* 25(3):242–253.
- DeRousseau M, Kasprzyk J, Srubar W (2018) Computational design optimization of concrete mixtures: A review. *Cement Concrete Res.* 109:42–53.

- Dheeru D, Karra Taniskidou E (2017) UCI machine learning repository. URL <http://archive.ics.uci.edu/ml>.
- Donti P, Amos B, Kolter JZ (2017) Task-based end-to-end model learning in stochastic optimization. *Advances in Neural Information Processing Systems* 30, 5484–5494.
- Du K, Kearfott RB (1994) The cluster problem in multivariate global optimization. *J. Global Optim.* 5(3):253–265.
- Duran MA, Grossmann IE (1986) A mixed-integer nonlinear programming algorithm for process systems synthesis. *AIChE Journal* 32(4):592–606.
- Easton K, Nemhauser G, Trick M (2003) Solving the travelling tournament problem: A combined integer programming and constraint programming approach. *Practice and Theory of Automated Timetabling IV*, 100–109, ISBN 978-3-540-45157-0.
- Eberhart R, Kennedy J (1995) A new optimizer using particle swarm theory. *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, 39–43.
- Elmachtoub AN, Grigas P (2017) Smart “Predict, then Optimize”. *arXiv e-prints* arXiv:1710.08005.
- Erdal HI (2013) Two-level and hybrid ensembles of decision trees for high performance concrete compressive strength prediction. *Eng. Appl. Artif. Intel.* 26(7):1689–1697.
- Fernández J, Tóth B (2009) Obtaining the efficient set of nonlinear biobjective optimization problems via interval branch-and-bound methods. *Comput. Optim. Appl.* 42(3):393–419.
- Freund Y (1995) Boosting a weak learning algorithm by majority. *Inform. Comput.* 121(2):256–285.
- Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29(5):1189–1232.
- Friedman JH (2002) Stochastic gradient boosting. *Comput. Stat. Data Anal.* 38(4):367–378.
- Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Machine Learning* 63(1):3–42.
- Hart WE, Laird CD, Watson JP, Woodruff DL, Hackebeil GA, Nicholson BL, Siirola JD (2017) *Pyomo—optimization modeling in Python*, volume 67 (Springer Science & Business Media), second edition.
- Hart WE, Watson JP, Woodruff DL (2011) Pyomo: modeling and solving mathematical programs in Python. *Math. Program. Comput.* 3(3):219–260.
- Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning* (Springer-Verlag New York), second edition.
- James G, Witten D, Hastie T, Tibshirani R (2013) *An Introduction to Statistical Learning* (Springer-Verlag New York).
- Jolliffe IT (2002) *Principal Component Analysis* (Springer-Verlag New York), second edition.
- Kannan R, Barton PI (2017) The cluster problem in constrained global optimization. *J. Global Optim.* 69(3):629–676.

- Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 30, 3149–3157.
- Kennedy J, Eberhart R (1995) Particle swarm optimization. *Proceedings of the IEEE International Conference on Neural Networks*, volume 4, 1942–1948.
- Kılınç M, Linderoth J, Luedtke J, Miller A (2014) Strong-branching inequalities for convex mixed integer nonlinear programs. *Comput. Optim. Appl.* 59(3):639–665.
- Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680.
- Klabjan D, Johnson EL, Nemhauser GL, Gelman E, Ramaswamy S (2001) Solving large airline crew scheduling problems: Random pairing generation and strong branching. *Comput. Optim. Appl.* 20(1):73–91.
- Konno H, Wiyayanayake A (2001) Portfolio optimization problem under concave transaction costs and minimal transaction unit constraints. *Math. Program.* 89(2):233–250, ISSN 1436-4646.
- Kuhn M (2008) Building predictive models in R using the caret package. *J. Stat. Softw.* 28(5):1–26.
- Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 18–22, ISSN 1609-3631.
- Lombardi M, Milano M (2018) Boosting combinatorial problem modeling with machine learning. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 5472–5478.
- Lombardi M, Milano M, Bartolini A (2017) Empirical decision model learning. *Artificial Intelligence* 244:343–367, combining Constraint Solving with Mining and Learning.
- Lundell A, Kronqvist J, Westerlund T (2017) SHOT – a global solver for convex MINLP in Wolfram Mathematica. Espua A, Graells M, Puigjaner L, eds., *27th European Symposium on Computer Aided Process Engineering*, volume 40 of *Computer Aided Chemical Engineering*, 2137–2142 (Elsevier).
- Miranda LJ (2018) PySwarms: A research toolkit for Particle Swarm Optimization in Python. *J. Open Source Softw.* 3.
- Misener R, Floudas CA (2010) Piecewise-linear approximations of multidimensional functions. *J. Optim. Theory Appl.* 145(1):120–147.
- Misener R, Floudas CA (2013) GloMIQO: Global Mixed-Integer Quadratic Optimizer. *J. Global Optim.* 57(1):3–50.
- Misener R, Floudas CA (2014) ANTIGONE: Algorithms for continuous / integer global optimization of nonlinear equations. *J. Global Optim.* 59(2):503–526.
- Misener R, Gounaris CE, Floudas CA (2009) Global optimization of gas lifting operations: A comparative study of piecewise linear formulations. *Ind. Eng. Chem. Res.* 48(13):6098–6104.
- Mistry M, Callia D’Iddio A, Huth M, Misener R (2018) Satisfiability modulo theories for process systems engineering. *Comput. Chem. Eng.* 113:98–114.

- Mišić VV (2017) Optimization of Tree Ensembles. *ArXiv e-prints* ArXiv:1705.10883.
- Morrison DR, Jacobson SH, Sauppe JJ, Sewell EC (2016) Branch-and-bound algorithms: A survey of recent advances in searching, branching, and pruning. *Discrete Optim.* 19:79–102.
- Niebling J, Eichfelder G (2016) A branch-and-bound algorithm for bi-objective problems. *Proceedings of the XIII Global Optimization Workshop*, 57–60.
- Niebling J, Eichfelder G (2019) A branch-and-bound-based algorithm for nonconvex multiobjective optimization. *SIAM J. Optim.* 29(1):794–821.
- Nocedal J, Wright SJ (2006) *Sequential Quadratic Programming*, 529–562. ISBN 978-0-387-40065-5.
- Ridgeway G (2017) Package ‘gbm’. URL <https://cran.r-project.org/web/packages/gbm/index.html>.
- Rossi F, Van Beek P, Walsh T (2006) *Handbook of constraint programming* (Elsevier).
- Schweidtmann AM, Mitsos A (2019) Deterministic global optimization with artificial neural networks embedded. *J. Optim. Theory Appl.* 180(3):925–948.
- Singer AB, Taylor JW, Barton PI, Green WH (2006) Global dynamic optimization for parameter estimation in chemical kinetics. *J. Phys. Chem. A* 110(3):971–976.
- Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems 25*, 2951–2959.
- Sra S, Nowozin S, Wright SJ (2012) *Optimization for Machine Learning* (MIT Press).
- Tawarmalani M, Sahinidis NV (2005) A polyhedral branch-and-cut approach to global optimization. *Math. Program.* 103:225–249.
- Vaswani N, Bouwmans T, Javed S, Narayanamurthy P (2018) Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery. *IEEE Signal Proc. Mag.* 35(4):32–55.
- Verwer S, Zhang Y, Ye QC (2017) Auction optimization using regression trees and linear models as integer programs. *Artificial Intelligence* 244:368–395, combining Constraint Solving with Mining and Learning.
- Vielma JP, Ahmed S, Nemhauser G (2010) Mixed-integer models for nonseparable piecewise-linear optimization: Unifying framework and extensions. *Oper. Res.* 58(2):303–315.
- Vigerske S (2012) *Decomposition in Multistage Stochastic Programming and a Constraint Integer Programming Approach to Mixed-Integer Nonlinear Programming*. PhD in Mathematics, Humboldt-University Berlin.
- Wechsung A, Schaber SD, Barton PI (2014) The cluster problem revisited. *J. Global Optim.* 58(3):429–438.
- Westerlund T, Pettersson F (1995) An extended cutting plane method for solving convex MINLP problems. *Comput. Chem. Eng.* 19:131–136.
- Wilder B, Dilkina B, Tambe M (2018) Melding the Data-Decisions Pipeline: Decision-Focused Learning for Combinatorial Optimization. *arXiv e-prints* arXiv:1809.05504.

- Xiang Y, Gubian S, Suomela B, Hoeng J (2013) Generalized simulated annealing for efficient global optimization: the GenSA package for R. *R J.* 5.
- Yeh IC (1998) Modeling of strength of high-performance concrete using artificial neural networks. *Cement Concrete Res.* 28(12):1797–1808.

Electronic supplementary material: Mixed-Integer Convex Nonlinear Optimization with Gradient-Boosted Trees Embedded

Miten Mistry • Dimitrios Letsios • Gerhard Krennrich • Robert M. Lee • Ruth Misener

Appendix A: Full convex MINLP formulation

$$\min_{\mathbf{v}^L \leq \mathbf{x} \leq \mathbf{v}^U} \text{cvx}(\mathbf{x}) + \sum_{t \in \mathcal{T}} \sum_{l \in \mathcal{L}_t} F_{t,l} z_{t,l} \quad (9a)$$

$$\text{s.t. } \sum_{l \in \mathcal{L}_t} z_{t,l} = 1, \quad \forall t \in \mathcal{T}, \quad (9b)$$

$$\sum_{l \in \text{Left}_{t,s}} z_{t,l} \leq y_{i(s),j(s)}, \quad \forall t \in \mathcal{T}, s \in \mathcal{V}_t, \quad (9c)$$

$$\sum_{l \in \text{Right}_{t,s}} z_{t,l} \leq 1 - y_{i(s),j(s)}, \quad \forall t \in \mathcal{T}, s \in \mathcal{V}_t, \quad (9d)$$

$$y_{i,j} \leq y_{i,j+1}, \quad \forall i \in [n], j \in [m_i - 1], \quad (9e)$$

$$x_i \geq v_{i,0} + \sum_{j=1}^{m_i} (v_{i,j} - v_{i,j-1})(1 - y_{i,j}), \quad \forall i \in [n], \quad (9f)$$

$$x_i \leq v_{i,m_i+1} + \sum_{j=1}^{m_i} (v_{i,j} - v_{i,j+1})y_{i,j}, \quad \forall i \in [n], \quad (9g)$$

$$y_{i,j} \in \{0, 1\}, \quad \forall i \in [n], j \in [m_i], \quad (9h)$$

$$z_{t,l} \geq 0, \quad \forall t \in \mathcal{T}, l \in \mathcal{L}_t. \quad (9i)$$

Appendix B: Table of Notation

Name	Description
GBT Ensemble Definition	
n	Number of the GBT-trained function (continuous) variables
i	Continuous variable index
x_i	Continuous variable
\mathbf{x}	Vector $(x_1, \dots, x_n)^T$
\mathcal{T}	Set of gradient boosted trees
t	Gradient boosted tree
\mathcal{V}_t	Set of split nodes (vertices) in tree t
\mathcal{L}_t	Set of leaf nodes in tree t
s	Split node associated with a tree t and mainly referred to as (t, s)
$i(t, s)$	Continuous variable index associated with split node s in tree t
$v(t, s)$	Splitting value of variable $x_{i(t,s)}$ at split node s in tree t
$\text{GBT}_t(\mathbf{x})$	Tree t evaluation at point \mathbf{x}
$\text{GBT}(\mathbf{x})$	GBT ensemble evaluation at point \mathbf{x}
Convex MINLP with GBTs Problem Definition	
$\text{cvx}(\mathbf{x})$	Convex function evaluation at point \mathbf{x}
m_i	Number of variable x_i splitting values
$v_{i,j}$	j -th greatest variable x_i splitting value
v_i^L or $v_{i,0}$	Variable x_i lower bound
v_i^U or v_{i,m_i+1}	Variable x_i upper bound
\mathbf{v}^L	Vector (v_1^L, \dots, v_n^L)
\mathbf{v}^U	Vector (v_1^U, \dots, v_n^U)
$\text{Left}_{t,s}$	Set of leaves in the subtree rooted in the left child of s in tree t
$\text{Right}_{t,s}$	Set of leaves in the subtree rooted in the right child of s in tree t

$F_{t,l}$	Contribution of leaf node l in tree t
$y_{i,j}$	Binary variable indicating whether $x_i \leq v_{i,j}$, or not
$z_{t,l}$	Binary variable specifying whether tree t evaluates at leaf l
d	Maximum tree depth
Branch-and-Bound Algorithm Overview	
$[\mathbf{v}^L, \mathbf{v}^U]$	Optimization problem global domain
$S = [\mathbf{L}, \mathbf{U}]$	Optimization problem subdomain / B&B node
(x_i, v)	GBT splitting point / B&B branch
$S_{\text{left}}, S_{\text{right}}, S_c, S'$	B&B nodes
Q	Set of unexplored B&B nodes
P_{root}	Initial GBT ensemble partition at B&B root node
P, P', P''	GBT ensemble partitions
$b^{\text{cvx}, S}$	Convex lower bound over domain S
$b^{\text{GBT}, S, P}$	GBT lower bound over domain S with respect to partition P
Lower Bounding	
R^S	Optimal objective value, i.e., tightest relaxation
\hat{R}^S	Relaxation dropping linking constraints
$b^{\text{GBT}, S, *}$	Optimal GBT lower bound over domain S
\mathbf{x}^*	Optimal solution
i, j, l	Subset indices of a GBT ensemble partition
k	GBT ensemble partition size
$\mathcal{T}_i, \mathcal{T}_j, \mathcal{T}', \mathcal{T}''$	Subsets of GBTs
N	GBT ensemble subset size
$n^{\mathcal{T}, S}$	Number of leaves in GBT subset \mathcal{T} over domain S
f^*	Best found feasible objective
q	Time limit on lower bound improvement algorithm
Branching	
B	Branch ordering
$r((x_i, v), t)$	Set of nodes in tree t that split on (x_i, v)
$d(s)$	Depth of split node s (root node has zero depth)
$w(s)$	Weight of split node s
$i(s)$	Number of inactive leaves below split s when branching with respect to (x_i, s)
$\text{weight}((x_i, v), t)$	Weight assigned to (x_i, v) in tree t
$\text{weight}((x_i, v), \mathcal{T})$	Weight assigned to (x_i, v) in GBT ensemble \mathcal{T}
$\text{inactive}((x_i, v), \mathcal{T})$	Number of inactive leaves when branching on pair (x_i, v) in \mathcal{T}
$\text{cover}(t, s)$	Set of leaves covered by split node s at tree t
$S, S_{\text{left}}, S_{\text{right}}, S_0$	B&B nodes denoted by their corresponding domain
l	Strong branching lookahead parameter

Table 5: Nomenclature

Appendix C: Heuristics

C.1. Mixed-Integer Convex Programming Heuristic

For a given a subset $\mathcal{T}' \subseteq \mathcal{T}$ of trees, let $f_{\mathcal{T}'}(\cdot)$ be the objective function obtained by ignoring the trees $\mathcal{T} \setminus \mathcal{T}'$. Then, $\min_{\mathbf{v}^L \leq \mathbf{x} \leq \mathbf{v}^U} \{f_{\mathcal{T}'}(\mathbf{x})\}$ may be significantly more tractable than the original problem instance when $|\mathcal{T}'| \ll |\mathcal{T}|$. So, the Algorithm 4 heuristic solves the original convex MINLP by sequentially solving smaller convex MINLP sub-instances of increasing size. A sub-instance is restricted to a subset $\mathcal{T}' \subseteq \mathcal{T}$ of GBTs. Let $\mathcal{T}^{(k)}$ be the subset of trees when the k -th heuristic iteration begins. Initially, $\mathcal{T}^{(0)} = \emptyset$, i.e., $f_{\mathcal{T}^{(0)}}(\cdot)$ consists only of the convex part. Denote by $\mathbf{x}^{(k)}$ the sub-instance optimal solution minimizing $f_{\mathcal{T}^{(k)}}(\cdot)$. Note that $\mathbf{x}^{(k)}$ is feasible for the full instance. Each iteration k chooses a set of N additional trees $\mathcal{T}^{\text{next}} \subseteq \mathcal{T} \setminus \mathcal{T}^{(k)}$

and constructs $\mathcal{T}^{(k+1)} = \mathcal{T}^{(k)} \cup \mathcal{T}^{\text{next}}$, i.e., $\mathcal{T}^{(k)} \subseteq \mathcal{T}^{(k+1)}$. Consider two approaches for picking the N trees between consecutive iterations: (i) training-aware selection and (ii) best improvement selection. Termination occurs when the time limit is exceeded and Algorithm 4 returns the best computed solution.

Training-aware selection Let T_1, T_2, \dots, T_m be the tree generation order during training. This approach selects the trees $\mathcal{T}^{\text{next}}$ according to this predefined order. That is, in the k -th iteration, $\mathcal{T}^{(k)} = \{T_1, \dots, T_{kN}\}$ and $\mathcal{T}^{\text{next}} = \{T_{kN+1}, \dots, T_{(k+1)N}\}$. A GBT training algorithm constructs the trees iteratively, so each new tree reduces the current GBT ensemble error with respect to the training data. Thus, we expect that the earliest-generated trees better approximate the learned function than the latest-generated trees. Specifically, for two subsets $\mathcal{T}_A, \mathcal{T}_B \subseteq \mathcal{T}$ with the property that $t_a < t_b$ for each $T_{t_a} \in \mathcal{T}_A$ and $T_{t_b} \in \mathcal{T}_B$, we expect that $|f_{\mathcal{T}_A}(\mathbf{x}) - f^*(\mathbf{x})| \leq |f_{\mathcal{T}_B}(\mathbf{x}) - f^*(\mathbf{x})|$, for each $\mathbf{v}^L \leq \mathbf{x} \leq \mathbf{v}^U$, where f^* is the original objective function, i.e., the optimal approximation. Intuitively, earlier trees place the GBT function within the correct vicinity, while later trees have a fine tuning role.

Best improvement selection In this approach, the k -th iteration picks the N trees with the maximum contribution when evaluating at $\mathbf{x}^{(k)}$. We select $\mathcal{T}^{\text{next}} \subseteq \mathcal{T} \setminus \mathcal{T}^{(k)}$ so that, for each pair of trees $T_t \in \mathcal{T}^{\text{next}}$ and $T_{t'} \in \mathcal{T} \setminus (\mathcal{T}^{(k)} \cup \mathcal{T}^{\text{next}})$, it holds that $f_t(\mathbf{x}^{(k)}) \geq f_{t'}(\mathbf{x}^{(k)})$. Assuming that approximation $\mathcal{T}^{(k)}$ is poor, then $\mathcal{T}^{\text{next}}$ contains the trees that refute optimality of $\mathbf{x}^{(k)}$ the most, from the perspective of $f_t(\mathbf{x}^{(k)})$ $t \in \mathcal{T} \setminus \mathcal{T}^{(k)}$.

Algorithm 4 Mixed-integer convex programming heuristic

```

1:  $k \leftarrow 0$ 
2:  $\mathcal{T}^{(k)} \leftarrow \emptyset$ 
3: while the time limit is not exceeded do
4:    $\mathbf{x}^{(k)} \leftarrow \arg \min_{\mathbf{v}^L \leq \mathbf{x} \leq \mathbf{v}^U} f_{\mathcal{T}^{(k)}}(\mathbf{x})$ 
5:   Choose  $\mathcal{T}^{\text{next}}$  from  $\{\mathcal{T}' \mid \mathcal{T}' \subseteq \mathcal{T} \setminus \mathcal{T}^{(k)}, |\mathcal{T}'| = \min\{N, |\mathcal{T} \setminus \mathcal{T}^{(k)}|\}\}$ 
6:    $\mathcal{T}^{(k+1)} \leftarrow \mathcal{T}^{(k)} \cup \mathcal{T}^{\text{next}}$ 
7:    $k \leftarrow k + 1$ 
8: end while
9: return  $\arg \min_{k \in \{0, \dots, k-1\}} f(\mathbf{x}^{(k)})$ 

```

C.2. Particle Swarm Optimization

Kennedy and Eberhart (1995) introduce PSO for optimizing continuous nonlinear functions. PSO computes a good heuristic solution by triggering m particles that collaboratively search the feasibility space. PSO picks the initial particle position $\mathbf{x}_i^{(0)}$ and search direction $\mathbf{v}_i^{(0)}$ of particle i randomly. The search occurs in a sequence of rounds. In round k , every particle chooses its next position $\mathbf{x}_i^{(k+1)}$ by following the direction specified by a weighted sum of: (i) the current trajectory direction $\mathbf{v}_i^{(k)}$, (ii) the particle's best found solution \mathbf{p}_i , (iii) the globally best found solution \mathbf{g} , and moving by a fixed step size. The *inertia term* $\omega \mathbf{v}_i^{(k)}$ controls

Algorithm 5 Particle Swarm Optimization

Compute initial position $\mathbf{x}_i^{(0)} \in \mathbb{R}^n$ and velocity $\mathbf{v}_i^{(0)} \in \mathbb{R}^n$ for each particle $i = 1, \dots, m$.
 $\mathbf{p}_i \leftarrow \mathbf{x}_i^{(0)}$
 $\mathbf{g} \leftarrow \arg \min \{f(\mathbf{p}_i)\}$
 $k \leftarrow 0$
while the time limit is not exceeded **do**
 for $i = 1, \dots, m$ **do**
 Choose random values $r_1, r_2 \sim U(0, 1)$
 $\mathbf{v}_i^{(k+1)} \leftarrow \omega \mathbf{v}_i^{(k)} + c_1 \cdot r_1 \cdot (\mathbf{p}_i - \mathbf{x}_i^{(k)}) + c_2 \cdot r_2 \cdot (\mathbf{g} - \mathbf{x}_i^{(k)})$
 $\mathbf{x}_i^{(k+1)} \leftarrow \mathbf{x}_i^{(k)} + \mathbf{v}_i^{(k+1)}$
 if $f(\mathbf{x}_i^{(k+1)}) < f(\mathbf{p}_i)$ **then**
 $\mathbf{p}_i \leftarrow \mathbf{x}_i^{(k+1)}$
 end if
 end for
 $\mathbf{g} \leftarrow \arg \min \{f(\mathbf{p}_i)\}$
 $k \leftarrow k + 1$
end while

how quickly a particle changes direction. The *cognitive term* $c_1 \cdot r_1 \cdot (\mathbf{p}_i - \mathbf{x}_i^{(k)})$ controls the particle tendency to move to the best observed solution by that particle. The *social term* $c_2 \cdot r_2 \cdot (\mathbf{g} - \mathbf{x}_i^{(k)})$ controls the particle tendency to move toward the best solution observed by any particle. Coefficients ω , c_1 , and c_2 are tunable parameters. Termination occurs either when all particles are close, or within a specified time limit. Algorithm 5 lists the PSO algorithm.

For Problem (1), we improve the PSO performance by avoiding initial particle positions in feasible regions strictly dominated by the convex term. We project the initial random points close to regions where the GBT term is significant compared to the convex term.

C.3. Simulated Annealing

Algorithm 6 lists the simulated annealing algorithm (Kirkpatrick et al. 1983).

Appendix D: Numerical Results: Heuristic Solutions

This section assesses performance of the Section 5.4 heuristic algorithms compared with simulated annealing. We use CPLEX 12.7 and Gurobi 7.5.2 as: (i) black-box solvers for the entire convex MINLP (2) and (ii) heuristic components for solving convex MINLP (2) instances in the Section 5.4 convex MINLP heuristic. The R package GenSA (Xiang et al. 2013) runs the Simulated Annealing (SA) metaheuristic. We provide a SA technical description (Kirkpatrick et al. 1983) in Section C.3. The Python module PySwarms (Miranda

Algorithm 6 Simulated Annealing

```

1: Compute an initial solution  $\mathbf{x}^{(0)} \in \mathbb{R}^n$ .
2: Set initial temperature  $T^{(0)} = 1$  and probability constant  $c = 1$ .
3: Set temperature factor  $\alpha \in [0.80, 0.99]$ .
4:  $t = 0, k = 0$ 
5: while  $T^{(t)} > \epsilon$  do
6:   for  $r$  iterations do
7:     Select a neighboring solution  $\mathbf{x} \in \mathcal{N}(\mathbf{x}^{(k)})$  randomly.
8:     if  $f(\mathbf{x}) < f(\mathbf{x}^{(k)})$  then
9:        $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}$ 
10:       $k \leftarrow k + 1$ 
11:     else
12:       Choose  $p \sim U(0, 1)$ 
13:       if  $\exp(-(f(\mathbf{x}) - f(\mathbf{x}^{(k)}))/cT^{(t)}) > p$  then
14:          $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}$ 
15:          $k \leftarrow k + 1$ 
16:       end if
17:     end if
18:   end for
19:    $T^{(t+1)} \leftarrow \alpha T^{(t)}$ 
20:    $t \leftarrow t + 1$ 
21: end while

```

2018) implements the Section 5.4 Particle Swarm Optimization (PSO) metaheuristic. Each heuristic, i.e. TA, BI, and Random, uses either CPLEX, or Gurobi as a black-box convex MINLP solver. We append the labels -C or -G to indicate the underlying solver. We use the default CPLEX 12.7 and Gurobi 7.5.2 tolerances, i.e., relative MIP gap, integrality and barrier convergence tolerances of 10^{-4} , 10^{-5} and 10^{-8} , respectively. We use the default SA parameters. We parameterize PSO with inertia term $\omega = 0.5$, cognitive term $c_1 = 0.7$, social term $c_2 = 0.3$, 500 particles and an iteration limit of 100. Each particle takes a randomly generated point, $\mathbf{x}^{(0)} \in [\mathbf{v}^L, \mathbf{v}^U]$, and its projection, $\mathbf{x}^{(p)}$ on \mathbf{P} and initializes at $\mathbf{x} = h \cdot \mathbf{x}^{(0)} + (1 - h) \cdot \mathbf{x}^{(p)}$. For our tests, we use $h = 0.15$.

D.1. Concrete Mixture Design

Table 6 compares the CPLEX 12.7, Gurobi 7.5.2, SA, and PSO computed solutions for the entire convex MINLP, under 1 hour time limit. SA performs the best. PSO solution is relatively close to the SA best found

Table 6 Concrete mixture design instance: black-box solver solutions (upper bounds) by solving the entire mixed-integer convex programming (convex MINLP) model using: (i) CPLEX 12.7, (ii) Gurobi 7.5.2, (iii) Simulated Annealing (SA), and (iv) Particle Swarm Optimization (PSO), with 1 hour timeout.

λ	CPLEX 12.7	Gurobi 7.5.2	PSO	SA
1	-14.2	-17.7	-88.7	-91.3
10	422.7	112.6	-86.0	-86.6
100	4,791.6	1,413.7	-80.1	-80.3
1000	48,480.6	14,425.1	-75.9	-71.6

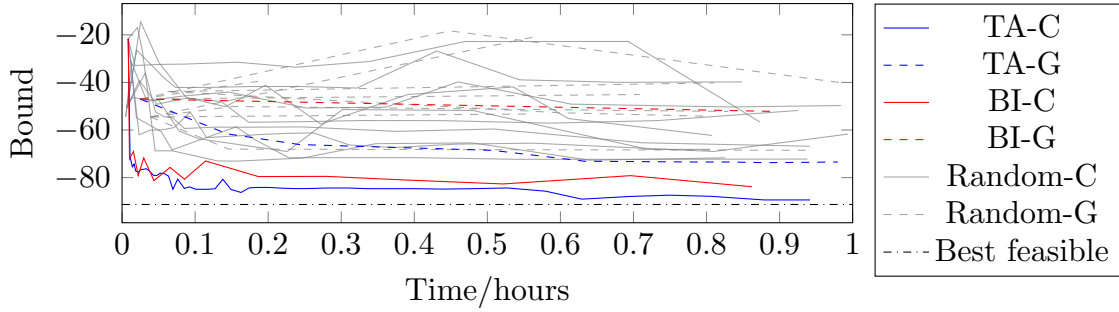


Figure 13 Concrete mixture design instance ($\lambda = 1$): Convex MINLP heuristic using training-aware (TA), best improvement (BI), or random strategies for choosing the next trees. Each iteration selects 10 new trees. The suffixes -C and -G denote using CPLEX 12.7 and Gurobi 7.5.2 as subsolvers, respectively. Best feasible is the simulated annealing solution.

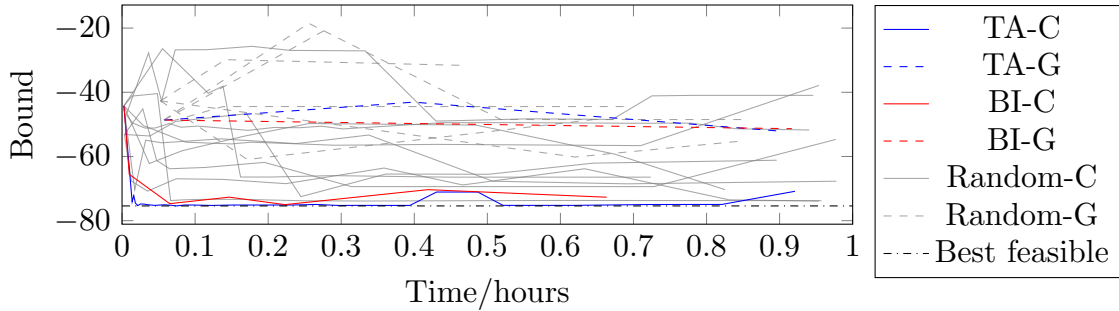


Figure 14 Concrete mixture design instance ($\lambda = 1000$): Convex MINLP heuristic using training-aware (TA), best improvement (BI), or random strategies for choosing the next trees. Each iteration selects 10 new trees. The suffixes -C and -G denote using CPLEX 12.7 and Gurobi 7.5.2 as subsolvers, respectively. Best feasible is the simulated annealing solution.

solution, compared to CPLEX 12.7 or Gurobi 7.5.2. Figures 13 and 14 evaluate the Section C.1 augmenting convex MINLP heuristic using CPLEX 12.7, Gurobi 7.5.2, and the different tree selection approaches, i.e., (i) training-aware (TA), (ii) best improvement (BI), and (iii) random selection. Figures 13 and 14 also plots the SA best-found solution. In general, both TA and BI perform better than random selection. Moreover, TA performs better than BI. Therefore, there is a benefit in choosing the earlier trees to find good heuristic solutions. Interestingly, the solution found in the first iteration of the augmenting convex MINLP heuristic,

Table 7 Chemical catalysis BASF instance (with different λ values): Black-box solver solutions (upper bounds) by solving the entire mixed-integer convex programming (convex MINLP) model using: (i) CPLEX 12.7, (ii) Gurobi 7.5.2, (iii) Simulated Annealing (SA), and (iv) Particle Swarm Optimization (PSO), with 1 hour timeout.

λ	CPLEX 12.7	Gurobi 7.5.2	PSO	SA
0	*	-158.5	-96.8	-168.2
1	*	-101.6	-89.8	-130.7
10	952	-100.1	-97.6	-102.7
100	1,040	11.5	-82.7	-84.2
1000	18,579	606.5	-76.5	-81.3

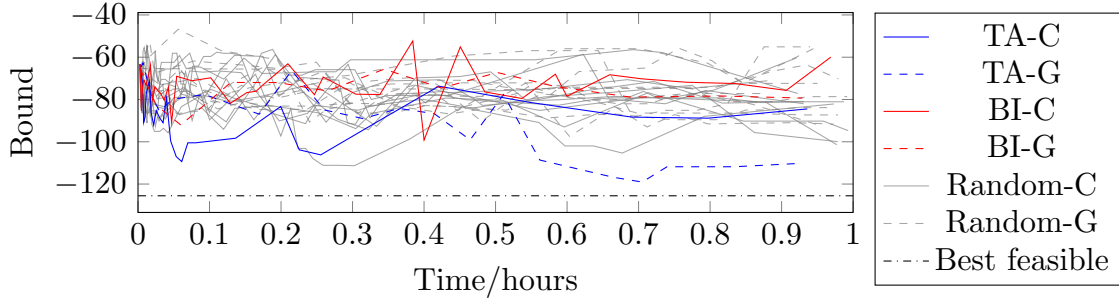


Figure 15 Chemical catalysis instance ($\lambda = 1$): Convex MINLP heuristic using training-aware (TA), best improvement (BI), or random strategies for choosing the next trees. Each iteration selects 10 new trees. The suffixes -C and -G denote using CPLEX 12.7 and Gurobi 7.5.2 as subsolvers, respectively. Best feasible is the simulated annealing solution.

i.e., by solely minimizing the convex part, is lower than -43, while the upper bounds reported by CPLEX 12.7 and Gurobi 7.5.2 after one hour of execution are greater than -18.

D.2. Chemical Catalysis

Table 7 compares the CPLEX 12.7, Gurobi 7.5.2, SA, and PSO computed solutions for the entire convex MINLP, under a 1 hour time limit. SA outperforms all others. PSO performs well for larger λ values, because it keeps the contribution of the convex part low at initialization. Gurobi 7.5.2 also performs relatively well for smaller λ values, however due to solver tolerances it may report incorrect objective values. For example, using $\lambda = 0$ the solver reports an objective of -174.1, however a manual evaluation results in -158.5. In fact, both CPLEX 12.7 or Gurobi 7.5.2, may produce incorrect outputs due to solver tolerances, hence a specialized fixing method may be necessary.

Figures 15 and 16 evaluate the Section C.1 augmenting convex MINLP heuristic for different values of the λ input parameter. We investigate the augmenting convex MINLP heuristic performance using either CPLEX 12.7, or Gurobi 7.5.2 for solving convex MINLP sub-instances and each of the: (i) training-aware (TA), (ii) best improvement (BI), and (iii) random selection strategies. The Figures 15 and 16 best feasible solution is the one produced by SA. For $\lambda = 1$, TA constructs several heuristic solutions that outperform both the BI and random selection ones. In this case, since the GBT part dominates the convex part, TA iteratively computes a better GBT approximation. For $\lambda = 1000$, TA and BI exhibit comparable performance, with BI

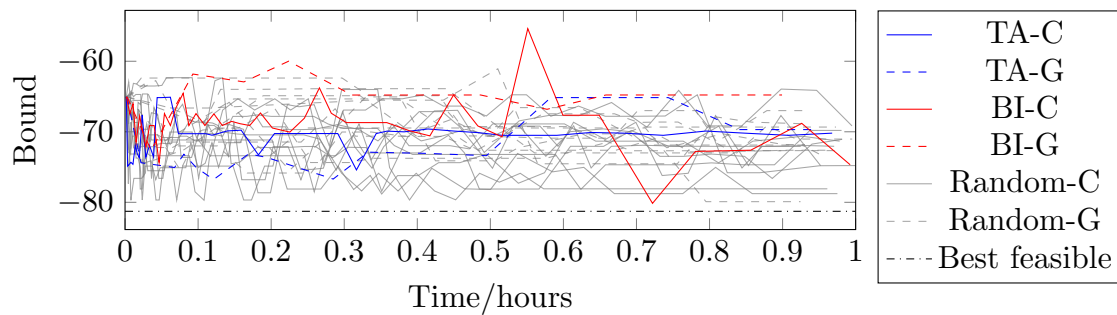


Figure 16 Chemical catalysis instance ($\lambda = 1000$): Convex MINLP heuristic using training-aware (TA), best improvement (BI), or random strategies for choosing the next trees. Each iteration selects 10 new trees. The suffixes -C and -G denote using CPLEX 12.7 and Gurobi 7.5.2 as subsolvers, respectively. Best feasible is the simulated annealing solution.

finding the best solution. Random selection also performs well because the convex part dominates the GBT part.