Generalization in portfolio-based algorithm selection

Maria-Florina Balcan
Carnegie Mellon University
ninamf@cs.cmu.edu

Tuomas Sandholm
Carnegie Mellon University
Optimized Markets, Inc.
Strategic Machine, Inc.
Strategy Robot, Inc.
sandholm@cs.cmu.edu

Ellen Vitercik
Carnegie Mellon University
vitercik@cs.cmu.edu

December 25, 2020

Abstract

Portfolio-based algorithm selection has seen tremendous practical success over the past two decades. This algorithm configuration procedure works by first selecting a portfolio of diverse algorithm parameter settings, and then, on a given problem instance, using an *algorithm selector* to choose a parameter setting from the portfolio with strong predicted performance. Oftentimes, both the portfolio and the algorithm selector are chosen using a *training set* of typical problem instances from the application domain at hand. In this paper, we provide the first provable guarantees for portfolio-based algorithm selection. We analyze how large the training set should be to ensure that the resulting algorithm selector's average performance over the training set is close to its future (expected) performance. This involves analyzing three key reasons why these two quantities may diverge: 1) the learning-theoretic complexity of the algorithm selector, 2) the size of the portfolio, and 3) the learning-theoretic complexity of the algorithm's performance as a function of its parameters. We introduce an end-to-end learning-theoretic analysis of the portfolio construction and algorithm selection together. We prove that if the portfolio is large, overfitting is inevitable, even with an extremely simple algorithm selector. With experiments, we illustrate a tradeoff exposed by our theoretical analysis: as we increase the portfolio size, we can hope to include a well-suited parameter setting for every possible problem instance, but it becomes impossible to avoid overfitting.

1 Introduction

Algorithms for many problems have tunable parameters. With a deft parameter tuning, these algorithms can often efficiently solve computationally challenging problems. However, the best parameter setting for one problem is rarely optimal for another. *Algorithm portfolios*—which are finite sets of parameter settings—are used in practice to deal with this variability. A portfolio is often used in conjunction with an *algorithm selector*, which is a function that determines which parameter setting in the portfolio to employ on any input problem instance. Portfolio-based algorithm selection has seen tremendous empirical success, fueling breakthroughs in combinatorial auction winner determination [23, 32], SAT [38], integer programming [22, 39], planning [15, 29], and many other domains.

Both the portfolio and the algorithm selector are often chosen using a *training set* of problem instances from the application domain at hand. This training set is typically assumed to be drawn from an unknown, application-specific distribution. The portfolio and algorithm selector are chosen to have strong average performance (quantified by low average runtime, for example) over

the training set. We investigate whether the learned algorithm selector also has strong expected performance on problems from the same application domain. The difference between average performance and expected performance is known as *generalization error*. If the generalization error is small, every parameter setting's average performance over the training set is close to its expected performance, so the learned algorithm selector will not *overfit*. When overfitting occurs, the learned selector has strong average performance over the training set but poor expected performance on the true distribution. In other words, the algorithm selector is overfitting to the problem instances in the training set.

There are multiple reasons the generalization error might be large in this setting: 1) the learning-theoretic complexity of the algorithm selector, 2) the size of the portfolio, and 3) the learning-theoretic complexity of the algorithm's performance as a function of its parameters. We provide end-to-end bounds on generalization error in terms of all three elements simultaneously. The variety of factors impacting generalization error differentiates this paper from prior research on generalization guarantees in algorithm configuration [4–11, 16, 19, 26]. That research focuses on bounding the generalization error of learning a *single* good parameter setting for the entire problem instance distribution, rather than a portfolio together with an algorithm selector that selects an algorithm (e.g., its parameter values) from the portfolio *for the specific instance at hand*. In the former case, generalization error only grows with (3)—just one of the sources of error we must contend with.

Our bounds apply to the widely-applicable setting where on any fixed input, algorithmic performance is a piecewise-constant function of its parameters with at most t pieces, for some $t \in \mathbb{Z}$. This structure has been observed in algorithm configuration for integer programming, greedy algorithms, clustering, and computational biology [3–5, 8, 19]. Given a training set of size N, we prove that the generalization error is bounded[1] by $\tilde{O}\left(\sqrt{\left(\bar{d} + \kappa \log t\right)/N}\right)$, where κ is the size of the portfolio and $\bar{d}$ measures the *intrinsic complexity* of the algorithm selector, as we define in Section 3. We also prove that this bound is tight up to logarithmic factors: the generalization error can be as large as $\tilde{\Omega}\left(\sqrt{\left(\bar{d} + \kappa\right)/N}\right)$. This implies that even if the algorithm selector is extremely simple ($\bar{d}$ is small), overfitting cannot be avoided in the worst case when the portfolio size κ is large. Moreover, we instantiate our guarantees for several commonly-used families of algorithm selectors [21, 22, 38].

Finally, via experiments in the context of integer programming configuration, we illustrate the inherent tradeoff our theory exposes: as we increase the portfolio size, we can hope to include a high-performing parameter setting for any given instance, but it become increasingly difficult to avoid overfitting. We incrementally increase the size of the portfolio and with each addition we train an algorithm selector using regression forest performance models. As the portfolio size increases, the algorithm selector's training performance continues to improve, but there comes a point where the test performance begins to worsen, meaning that the algorithm selector is overfitting to the training set.

Additional related research. Gupta and Roughgarden [19] also provide generalization guarantees for algorithm configuration. They primarily analyze the problem of learning a single parameter setting with high expected performance on the underlying distribution. They do provide guarantees for the more general problem of learning a mapping from instances to parameter settings in a few special cases, but do not study the problem of learning a portfolio in conjunction with learning a selector, which we do. They study settings where for each problem instance, a domain expert has defined a number of relevant features, as do we in Section 4. Their first result applies to learning an

[1]Here we assume that algorithmic performance is a quantity in $[0, 1]$, an assumption we relax in Section 2.

algorithm selector when the set of features is finite. In contrast, our results apply to infinite feature spaces. Their second set of results is tailored to the problem of learning empirical performance models and applies when the feature space is infinite. An empirical performance model is meant to predict how long a particular algorithm will take to run on a given input. An algorithm selector can use an empirical performance model by selecting the parameter setting with best predicted performance. Gupta and Roughgarden [19] provide guarantees that bound the difference between the empirical performance model's expected error and average error over the training set. Their guarantees can be applied once the portfolio is already chosen. They do not study the problem of learning the portfolio itself, whereas we study the composite problem of learning the portfolio and the algorithm selector.

In a related theoretical direction, several papers have studied a model where there are multiple algorithms capable of computing a correct solution to a given problem, but with different costs. The user can run multiple algorithms until one terminates with the correct solution. Given a training set of problem instances, the authors provide guarantees for learning a schedule with high expected performance [33, 35, 36]. That is a distinct problem from ours, since our goal is to learn an algorithm selector rather than a schedule. Moreover, we additionally handle the problem of learning the portfolio itself.

2 Problem formulation and road map

Notation. Our theoretical guarantees apply to algorithms parameterized by a real value $\rho \in \mathbb{R}$. We use the notation $\mathcal{Z}$ to denote the set of problem instances the algorithm may take as input. For example, $\mathcal{Z}$ might consist of integer programs (IPs) if we are configuring an IP solver. There is an unknown distribution $\mathcal{D}$ over problem instances in $\mathcal{Z}$.

To describe the performance of a parameterized algorithm, we adopt the notation of prior research [8]. For every parameter setting $\rho \in \mathbb{R}$, there is a function $u_\rho : \mathcal{Z} \to [0, H]$ that measures, abstractly, the performance of the algorithm parameterized by ρ given an input $z \in \mathcal{Z}$. For example, u_ρ might measure runtime or the quality of the algorithm's output. We use the notation $\mathcal{U} = \{u_\rho : \rho \in \mathbb{R}\}$ to denote the set of all performance functions.

Problem formulation. A portfolio-based algorithm selection procedure relies on two key components: a *portfolio* and an *algorithm selector*. A portfolio is a set $\mathcal{P} = \{\rho_1, \ldots, \rho_\kappa\} \subseteq \mathbb{R}$ of κ parameter settings. An algorithm selector is a mapping $f : \mathcal{Z} \to \mathcal{P}$ from problem instances $z \in \mathcal{Z}$ to parameter settings $f(z) \in \mathcal{P}$. In practice [22, 32, 39], the portfolio and algorithm selector are typically learned using the following high-level procedure:

1. Choose a class $\mathcal{F}$ of algorithm selectors, each of which maps $\mathcal{Z}$ to $\mathbb{R}$. (In Section 4, we provide several examples of classes $\mathcal{F}$ used in practice.)
2. Draw a training set $\mathcal{S} = \{z_1, \ldots, z_N\} \sim \mathcal{D}^N$ of problem instances from the unknown distribution $\mathcal{D}$.
3. Use $\mathcal{S}$ to learn a portfolio $\hat{\mathcal{P}} = \{\rho_1, \ldots, \rho_\kappa\} \subseteq \mathbb{R}$.
4. Use $\mathcal{S}$ to learn an algorithm selector $\hat{f} \in \mathcal{F}$ that maps to parameter settings in the portfolio $\hat{\mathcal{P}}$.

Given an instance $z \in \mathcal{Z}$, the performance of the parameter setting selected by $\hat{f}$ is $u_{\hat{f}(z)}(z)$. We bound the expected quality $\mathbb{E}_{z\sim\mathcal{D}}\left[u_{\hat{f}(z)}(z)\right]$ of the learned algorithm selector.

Road map. We first analyze to what extent the average performance of the selector $\hat{f}$ over the training set generalizes to its expected performance on the distribution. We then use this analysis to relate the performance of the learned selector $\hat{f}$ and the optimal selector under the optimal choice of a portfolio. In particular, we bound the difference between $\mathbb{E}_{z\sim\mathcal{D}}\left[u_{\hat{f}(z)}(z)\right]$ and $\max_{\mathcal{P}:|\mathcal{P}|\leq\kappa}\mathbb{E}_{z\sim\mathcal{D}}\left[\max_{\rho\in\mathcal{P}}u_\rho(z)\right]$. (Equivalently, if our goal is to minimize $u_\rho(z)$, we may replace each max with a min.)

3 Sample complexity bounds

In this section, we bound the difference between the average performance of any selector $f \in \mathcal{F}$ over the training set $\mathcal{S} \sim \mathcal{D}^N$ and its expected performance. Formally, we bound

$$\left|\frac{1}{N}\sum_{z\in\mathcal{S}}u_{f(z)}(z) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[u_{f(z)}(z)\right]\right| \tag{1}$$

for any choice of an algorithm selector $f \in \mathcal{F}$. This will serve as a building block for our general analysis of portfolio-based algorithm selection.

Our bounds apply in the widely-applicable setting where on any fixed input, algorithmic performance is a piecewise-constant function of the algorithm's parameters. This structure has been observed in algorithm configuration for integer programming, greedy algorithms, clustering, and computational biology [3–5, 8, 19]. To describe this structure more formally, for a fixed input $z \in \mathcal{Z}$, we use the notation $u_z^* : \mathbb{R} \to \mathbb{R}$ to denote algorithmic performance as a function of the parameters (whereas the functions u_ρ defined in Section 2 measure performance as a function of the input z). Naturally, $u_z^*(\rho) = u_\rho(z)$. We refer to u_z^* as a *dual function* (as opposed to u_ρ, which is a *primal function*). We assume algorithmic performance is a piecewise-constant function of the parameters, or more formally, that each function u_z^* is piecewise constant with at most t pieces, for some $t \in \mathbb{Z}$.

Our bounds depend on both the number of pieces t and on the *intrinsic complexity* of the class of algorithm selectors $\mathcal{F}$. We use the following notion of the *multi-class projection* of $\mathcal{F}$ to define the class's intrinsic complexity.

Definition 3.1. Given a selector $f \in \mathcal{F}$, let $\rho_1 < \rho_2 < \cdots < \rho_{\bar{\kappa}}$ be the parameter settings f maps to, with $\bar{\kappa} \leq \kappa$. The function f defines a partition $Z_1, \ldots, Z_{\bar{\kappa}}$ of the problem instances $\mathcal{Z}$ where for any $z \in \mathcal{Z}$, if $f(z) = \rho_i$, then $z \in Z_i$. For each function $f \in \mathcal{F}$ there is therefore a corresponding multi-class function $\bar{f} : \mathcal{Z} \to [\kappa]$ that indicates which set of the partition the instance z belongs to: $\bar{f}(z) = i$ when $z \in Z_i$. We use the notation $\bar{\mathcal{F}} = \left\{\bar{f} : f \in \mathcal{F}\right\}$ to denote the set of all such multi-class functions.

Defining this set of multi-class functions allows us to use classic tools from multi-class learning to reason about the algorithm selectors $\mathcal{F}$. In particular, our bounds depend on the *Natarajan [27] dimension* of the class $\bar{\mathcal{F}}$, which is a natural extension of the classic VC dimension [37] to multi-class functions.

Definition 3.2 (Natarajan dimension)**.** The set $\bar{\mathcal{F}}$ *multi-class shatters* a set of problem instances $z_1, \ldots, z_N$ if there exist labels $y_1, \ldots, y_N \in [\kappa]$ and $y_1', \ldots, y_N' \in [\kappa]$ such that:

1. For every $i \in [N]$, $y_i \neq y_i'$, and
2. For any subset $C \subseteq [N]$, there exists a function $\bar{f} \in \bar{\mathcal{F}}$ such that $\bar{f}(z_i) = y_i$ if $i \in C$ and $\bar{f}(z_i) = y_i'$ otherwise.

The *Natarajan dimension* of $\bar{\mathcal{F}}$ is the cardinality of the largest set that can be multi-class shattered by $\bar{\mathcal{F}}$.

In Section 4, we bound the Natarajan dimension of $\bar{\mathcal{F}}$ for several commonly-used classes of algorithm selectors $\mathcal{F}$. We use Natarajan dimension to quantify the intrinsic complexity of the class of selectors, which in turn allows us to bound Equation (1) for every function $f \in \mathcal{F}$. To do so, we relate the Natarajan dimension of $\bar{\mathcal{F}}$ to the *pseudo-dimension* of the function class $\mathcal{U}_{\mathcal{F}} = \left\{z \mapsto u_{f(z)}(z) : f \in \mathcal{F}\right\}$. Every function in $\mathcal{U}_{\mathcal{F}}$ is defined by an algorithm selector $f \in \mathcal{F}$. On input $z \in \mathcal{Z}$, $u_{f(z)}(z)$ equals the utility of the algorithm parameterized by $f(z)$ on input z. Pseudo-dimension [20] is a classic learning-theoretic tool for measuring the intrinsic complexity of a class of real-valued functions (whereas Natarajan dimension applies to multi-class functions). Both Natarjan dimension and pseudo-dimension are extensions of the classic VC dimension, so they bear some resemblance. Below, we define the pseudo-dimension of the class $\mathcal{U}_{\mathcal{F}}$.

Definition 3.3 (Pseudo-dimension)**.** The set $\mathcal{U}_{\mathcal{F}}$ *shatters* a set of instances $z_1, \ldots, z_N \in \mathcal{Z}$ if there exist *witnesses* $w_1, \ldots, w_N \in \mathbb{R}$ such that for any subset $C \subseteq [N]$, there exists an algorithm selector $f \in \mathcal{F}$ such that $u_{f(z_i)}(z_i) \leq w_i$ if $i \in C$ and $u_{f(z_i)}(z_i) > w_i$ otherwise. The *pseudo-dimension* of $\mathcal{U}_{\mathcal{F}}$, denoted $\operatorname{Pdim}(\mathcal{U}_{\mathcal{F}})$, is the size of the largest set of instances that can be shattered by $\mathcal{U}_{\mathcal{F}}$.

Classic learning-theoretic results allow us to provide generalization bounds once we calculate the pseudo-dimension. For example [20], with probability $1-\delta$ over the draw of the set $\{z_1, \ldots, z_N\} \sim \mathcal{D}^N$, for any selector $f \in \mathcal{F}$,

$$\left|\frac{1}{N}\sum_{i=1}^{N} u_{f(z_i)}(z_i) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[u_{f(z)}(z)\right]\right| = O\left(H\sqrt{\frac{1}{N}\left(\operatorname{Pdim}\left(\mathcal{U}_{\mathcal{F}}\right) + \log\frac{1}{\delta}\right)}\right). \tag{2}$$

We now prove a general bound on $\operatorname{Pdim}(\mathcal{U}_{\mathcal{F}})$, which allows us to bound Equation (1). The proof is in Appendix A.

Theorem 3.4. *Suppose each dual function u_z^* is piecewise-constant with at most t pieces. Let $\bar{d}$ be the Natarajan dimension of $\bar{\mathcal{F}}$. Then* $\operatorname{Pdim}(\mathcal{U}_{\mathcal{F}}) = \tilde{O}\left(\bar{d} + \kappa \log t\right)$.

At a high level, the $\tilde{O}\left(\bar{d}\right)$ term accounts for the intrinsic complexity of the algorithm selectors $\mathcal{F}$. The $O\left(\kappa \log t\right)$ term accounts for the complexity of composing selectors f with the performance functions u_ρ. In Theorem 3.5, we prove this bound is tight up to logarithmic factors.

Proof sketch of Theorem 3.4. Let $z_1, \ldots, z_N \in \mathcal{Z}$ be an arbitrary set of problem instances. Since each dual function $u_{z_i}^*$ is piecewise-constant with at most t pieces, there are $M \leq Nt$ intervals $I_1, \ldots, I_M$ partitioning $\mathbb{R}$ where for any interval I_j and any instance z_i, $u_{z_i}^*(\rho)$ is constant across all $\rho \in I_j$. Given these intervals, we partition the algorithm selectors in $\mathcal{F}$ into at most M^κ sets so that within any one set, all selectors map to the same κ (or fewer) intervals. Focusing on the selectors within one set $\mathcal{F}_0$ of the partition, we prove that the number of ways the utility functions u_f across $f \in \mathcal{F}_0$ can labels the instances $z_1, \ldots, z_N$ is upper bounded by the number of ways the multi-class projection functions $\bar{f}$ across $f \in \mathcal{F}_0$ can label the instances. We can then use the Natarajan dimension of $\bar{\mathcal{F}}$ to bound the number of ways the functions in $\mathcal{U}_{\mathcal{F}}$ label the instances $z_1, \ldots, z_N$. □

Theorem 3.4 and Equation (2) imply that with probability $1-\delta$ over the draw $\mathcal{S} \sim \mathcal{D}^N$, for any selector $f \in \mathcal{F}$,

$$\left|\frac{1}{N}\sum_{z\in\mathcal{S}} u_{f(z)}(z) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[u_{f(z)}(z)\right]\right| = O\left(H\sqrt{\frac{1}{N}\left(\bar{d} + \kappa\log t + \log\frac{1}{\delta}\right)}\right). \tag{3}$$

This theorem quantifies a fundamental tradeoff: as the portfolio size increases, we can hope to obtain better and better empirical performance $\sum_{z\in\mathcal{S}} u_{f(z)}(z)$ but the generalization error $\tilde{O}\left(H\sqrt{\left(\bar{d}+\kappa\right)/N}\right)$ will worsen.

We now prove that Theorem 3.4 is tight up to logarithmic factors. The following theorem illustrates that even if the class of algorithm selectors is extremely simple (in that the Natarajan dimension of $\bar{\mathcal{F}}$ is 0), if the portfolio size (that is, the number κ of parameters mapped to) is large, we cannot hope to avoid overfitting. The full proof is in Appendix A.

Theorem 3.5. *For any $\kappa, \bar{d} \geq 2$, there is a class of functions $\mathcal{U} = \{u_\rho : \rho \in \mathbb{R}\}$ and a class of selectors $\mathcal{F}$ such that:*

1. *Each selector $f \in \mathcal{F}$ maps to $\leq \kappa$ parameter settings.*
2. *Each dual function u_z^* is piecewise-constant with 1 discontinuity,*
3. *The Natarajan dimension of $\bar{\mathcal{F}}$ is at most $\bar{d}$, and*
4. *The pseudo-dimension of $\mathcal{U}_\mathcal{F}$ is $\Omega\left(\kappa + \bar{d}\right)$.*

Proof sketch. Let $\mathcal{Z} = (0, 1]$. For each parameter setting $\rho \in \mathbb{R}$, define $u_\rho(z) = \mathbf{1}_{\{z\leq\rho\}}$. Let $\kappa, \bar{d} \geq 2$ be two arbitrary integers. We split this proof into two cases: $\bar{d} \geq \kappa$ and $\kappa > \bar{d}$. In both cases, we construct a class of selectors $\mathcal{F}$ that satisfies the properties in the theorem statement and we prove that $\text{Pdim}\left(\mathcal{U}_\mathcal{F}\right) \geq \max\left\{\kappa, \bar{d}\right\} = \Omega\left(\kappa + \bar{d}\right)$. We sketch the proof of the case where $\kappa > \bar{d}$.

We begin by partitioning $\mathcal{Z} = (0, 1]$ into κ intervals $Z_1, \ldots, Z_\kappa$, where $Z_i = \left(\frac{i-1}{\kappa}, \frac{i}{\kappa}\right]$. For each set $C \subseteq [\kappa]$, we define an selector $f_C : \mathcal{Z} \to \mathbb{R}$ as follows. For any $z \in \mathcal{Z}$, let i be the index of the interval z lies in, i.e., $z \in Z_i$. If $i \in C$, we map $f_C(z) = \frac{i}{\kappa}$ and if $i \notin C$, we map $f_C(z) = \frac{i}{\kappa} - \frac{1}{2\kappa}$. Let $\mathcal{F} = \{f_C : C \subseteq [\kappa]\}$. The multi-class projection of $\bar{\mathcal{F}}$ is extremely simple: its Natarjan dimension is 0. Moreover, the set $\mathcal{S} = \left\{\frac{1}{\kappa}, \frac{2}{\kappa}, \ldots, \frac{\kappa-1}{\kappa}, 1\right\}$ is shattered by $\mathcal{U}_\mathcal{F}$ because—at a high level—each selector f_C maps each element $z \in \mathcal{S}$ to a parameter just above z or just below z, which allows the function class $\mathcal{U}_\mathcal{F}$ to shatter $\mathcal{S}$. □

In the proof of Theorem 3.5, each performance function u_ρ maps to $\{0, 1\}$, so we effectively prove a lower bound on the *VC dimension* of $\mathcal{U}_\mathcal{F}$. Classic results from learning theory imply the generalization error of learning a selector $f \in \mathcal{F}$ can therefore be as large as $\tilde{\Omega}\left(H\sqrt{\left(\bar{d}+\kappa\right)/N}\right)$, which matches Equation (3) up to logarithmic factors.

4 Application of theory to algorithm selectors

We now instantiate Theorem 3.4 for several commonly-used classes of algorithm selectors. In each of the case studies, there is a feature mapping $\phi : \mathcal{Z} \to \mathbb{R}^m$ that assigns feature vectors $\phi(z) \in \mathbb{R}^m$ to problem instances $z \in \mathcal{Z}$.

4.1 Linear performance models

We begin by providing guarantees for algorithm selectors that use a linear performance model. These have been used extensively in computational research [38, 39]. To define this type of selector, let $\boldsymbol{\rho} = (\rho_1, \ldots \rho_\kappa)$ be a set of κ distinct parameter settings. For each $i \in [\kappa]$, define a vector $\boldsymbol{w}_i \in \mathbb{R}^m$

and let

$$W = \begin{pmatrix} | & \cdots & | \\ \boldsymbol{w}_1 & \ddots & \boldsymbol{w}_\kappa \\ | & \cdots & | \end{pmatrix}$$

be a matrix containing all κ weight vectors. The dot product $\boldsymbol{w}_i \cdot \phi(z)$ is meant to estimate the performance of the algorithm parameterized by ρ_i on instance z. We define the algorithm selector $f_{\boldsymbol{\rho},W}(z) = \rho_i$ where $i = \text{argmax}_{j\in[\kappa]} \{\boldsymbol{w}_j \cdot \phi(z)\}$, which selects the parameter setting with best predicted performance. We define the class of algorithm selectors $\mathcal{F}_L = \{f_{\boldsymbol{\rho},W} : W \in \mathbb{R}^{m\times\kappa}, \boldsymbol{\rho} \in \mathbb{R}^\kappa\}$. To define the class $\bar{\mathcal{F}}_L$, for each matrix $W \in \mathbb{R}^{m\times\kappa}$, let $g_W : \mathcal{Z} \to [\kappa]$ be a function where $g_W(z) = \text{argmax}_{i\in[\kappa]} \{\boldsymbol{w}_i \cdot \phi(z)\}$. By definition, $\bar{\mathcal{F}}_L = \{g_W : W \in \mathbb{R}^{m\times\kappa}\}$, so $\bar{\mathcal{F}}_L$ is the well-studied *m-dimensional linear class* which has a Natarajan dimension of $O(m\kappa)$ [34]. This fact implies the following corollary.

Corollary 4.1. *Suppose the dual functions are piecewise-constant with at most t pieces. The pseudo-dimension of $\mathcal{U}_{\mathcal{F}_L} = \left\{z \mapsto u_{f(z)} : f \in \mathcal{F}_L\right\}$ is $O(\kappa m \log(\kappa m) + \kappa \log t)$.*

4.2 Regression tree performance models

We now analyze algorithm selectors that use a regression tree as the performance model. These have proven powerful in computational research [21]. A regression tree T's leaf nodes partition the feature space $\mathbb{R}^m$ into disjoint regions $R_1, \ldots, R_\ell$. In each region R_i, a constant value c_i is used to predict the algorithm's performance on instances in the region. The internal nodes of the tree define this partition: each performs an inequality test on some feature of the input. We use the notation $h_T(z)$ to denote tree T's prediction of the algorithm's performance on instance z. Formally, $h_T(z)$ equals the constant value corresponding to the region of the tree's partition to which $\phi(z)$ belongs.

An algorithm selector can be defined using a regression tree performance model as follows. Let $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_\kappa)$ be a set of κ distinct parameter settings. For each parameter setting ρ_i, let T_i be a tree that is meant to predict the performance of the algorithm parameterized by ρ_i, and let $\boldsymbol{T} = (T_1, \ldots, T_\kappa)$ be the set of all κ trees. We define the algorithm selector $f_{\boldsymbol{\rho},\boldsymbol{T}}(z) = \rho_i$ where $i = \text{argmax}_{j\in[\kappa]} \left\{h_{T_j}(z)\right\}$. The class of algorithm selectors $\mathcal{F}_R$ consists of all functions $f_{\boldsymbol{\rho},\boldsymbol{T}}$ across all parameter vectors $\boldsymbol{\rho} \in \mathbb{R}^\kappa$ and all κ-tuples of regression trees $\boldsymbol{T} = (T_1, \ldots, T_\kappa)$. The full proof of the following lemma is in Appendix A.1.

Lemma 4.2. *Suppose we limit ourselves to building regression trees with at most ℓ leaves. Then the Natarajan dimension of $\bar{\mathcal{F}}_R$ is $O(\ell\kappa \log(\ell\kappa m))$.*

Proof sketch. For each κ-tuple of regression trees $\boldsymbol{T} = (T_1, \ldots, T_\kappa)$, let $g_{\boldsymbol{T}} : \mathcal{Z} \to [\kappa]$ be a function where $g_{\boldsymbol{T}}(z) = \text{argmax}_{i\in[\kappa]} \{h_{T_i}(z)\}$. By definition, the set $\bar{\mathcal{F}}_R$ consists of the functions $g_{\boldsymbol{T}}$ across all κ-tuples of regression trees $\boldsymbol{T}$ with at most ℓ leaves. Let $z_1, \ldots, z_N \in \mathcal{Z}$ be a set of problem instances. Our goal is to bound the number of ways the functions $g_{\boldsymbol{T}}$ can label these instances. A single regression tree induces a partition of these N problem instances defined by which leaf each instance is mapped to as we apply the tree's inequality tests. The key step in this proof is bounding the total number of partitions we can induce by varying the tree's inequality tests. We then generalize this intuition to bound the number of partitions κ regression trees can induce as we vary all their parameters. Once the partition of each regression tree is fixed, the tree with the largest prediction for each problem instance depends on the relative ordering of the constants at the trees' leaves. There is a bounded number of possible relative orderings, and we aggregate all of these bounds to prove the lemma statement. □

Corollary 4.3. *Suppose the dual functions are piecewise-constant with at most t pieces and we limit ourselves to building regression trees with at most ℓ leaves. Then* $\operatorname{Pdim}\left(\mathcal{U}_{\mathcal{F}_R}\right) = O(\ell\kappa\log(\ell\kappa m) + \kappa\log t)$.

This pseudo-dimension bound reflects the end-to-end nature of our analysis, since the guarantee bounds the generalization error of both selecting the portfolio and training the regression tree performance model. This is why the bound grows with both the size of the portfolio (κ) and the complexity of the regression trees (ℓ and m).

4.3 Clustering-based algorithm selectors

We now provide guarantees for clustering-based algorithm selectors, which have also been used in computational research [22]. This type of selector clusters the feature vectors $\phi(z_1), \ldots, \phi(z_N) \in \mathbb{R}^m$ and chooses a good parameter setting for each cluster. On a new instance z, the selector determines which cluster center is closest to $\phi(z)$ and runs the algorithm using the parameter setting assigned to that cluster. More formally, let $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_\kappa)$ be a set of parameter settings and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\kappa \in \mathbb{R}^m$ be a set of vectors. We define the matrix

$$X = \begin{pmatrix} | & \cdots & | \\ \boldsymbol{x}_1 & \ddots & \boldsymbol{x}_\kappa \\ | & \cdots & | \end{pmatrix},$$

where each column $\boldsymbol{x}_i$ is meant to represent a cluster center. We define the algorithm selector $f_{\boldsymbol{\rho},X}(z) = \rho_i$ where $i = \operatorname{argmin}_{j\in[\kappa]}\left\{\left\|\boldsymbol{x}_j - \phi(z)\right\|_p\right\}$, for some ℓ_p-norm with $p \geq 1$. The class of algorithm selectors is $\mathcal{F}_C = \left\{f_{\boldsymbol{\rho},X} : \boldsymbol{\rho} \in \mathbb{R}^\kappa, X \in \mathbb{R}^{m\times\kappa}\right\}$. The full proof of the following lemma is in Appendix A.2.

Lemma 4.4. *For any $p \in [1,\infty)$, the Natarajan dimension of $\bar{\mathcal{F}}_C$ is $O(m\kappa\log(m\kappa p))$.*

Proof sketch. For each matrix X, let $g_X : \mathcal{Z} \to [\kappa]$ be defined such that

$$g_X(z) = \operatorname{argmin}_{i\in[\kappa]}\left\{\left\|\boldsymbol{x}_i - \phi(z)\right\|_p^p\right\}.$$

By definition, $\bar{\mathcal{F}}_C = \{g_X : X \in \mathbb{R}^{m\times\kappa}\}$. Let $z_1, \ldots, z_N \in \mathcal{Z}$ be a set of problem instances. Our goal is to bound the number of ways the functions g_X can label these instances as we vary $X \in \mathbb{R}^{m\times\kappa}$. We do so by analyzing, for each instance z_i, the boundaries in $\mathbb{R}^{m\times\kappa}$ where if we shift X from one side of the boundary to the other, the column in X closest to $\phi(z_i)$ changes. We show that these boundaries are defined by multi-dimensional polynomials. We bound the total number of regions these boundaries induce in $\mathbb{R}^{m\times\kappa}$, which implies a bound on the Natarajan dimension of $\bar{\mathcal{F}}_C$. □

Lemma 4.4 and Theorem 3.4 imply the following bound.

Corollary 4.5. *Suppose the dual functions are piecewise-constant with at most t pieces. Then* $\operatorname{Pdim}\left(\mathcal{U}_{\mathcal{F}_C}\right) = \tilde{O}\left(m\kappa + \kappa\log t\right)$.

5 Learning procedure with guarantees

In this section, we use the results from the previous section to provide guarantees for the high-level learning procedure outlined in Section 2:

1. Draw a training set of problem instances $\mathcal{S} \sim \mathcal{D}^N$.
2. Use the training set $\mathcal{S}$ to select a set of κ or fewer parameter settings $\hat{\mathcal{P}} \subseteq \mathbb{R}$.
3. Use $\mathcal{S}$ to learn an algorithm selector $\hat{f} \in \mathcal{F}$ that maps problem instances $z \in \mathcal{Z}$ to parameter settings $\hat{f}(z) \in \hat{\mathcal{P}}$.

Our guarantees depend on the quality of the portfolio $\hat{\mathcal{P}}$ and selector $\hat{f}$, as formalized by the following definition.

Definition 5.1. Given a training set $\mathcal{S} \subseteq \mathcal{Z}^N$ and parameters $\alpha \in (0, 1]$, $\beta \in [0, 1]$, and $\epsilon \in [0, 1]$, we say the portfolio $\hat{\mathcal{P}}$ and the algorithm selector $\hat{f}$ are $(\alpha, \beta, \epsilon)$-optimal if:

1. The portfolio $\hat{\mathcal{P}}$ is nearly optimal over the training set in the sense that
$$\frac{1}{N}\sum_{z\in\mathcal{S}} \max_{\rho\in\hat{\mathcal{P}}} u_\rho(z) \geq \alpha \max_{\mathcal{P}\subset\mathbb{R}:|\mathcal{P}|\leq\kappa} \frac{1}{N}\sum_{z\in\mathcal{S}} \max_{\rho\in\mathcal{P}} u_\rho(z) - \beta.$$
(The maximization means that performance is measured with respect to an oracle that selects an optimal algorithm parameter ρ from the portfolio for each instance.)
2. The algorithm selector $\hat{f}$ returns high-performing parameter settings from the set $\hat{\mathcal{P}}$ in the sense that
$$\frac{1}{N}\sum_{z\in\mathcal{S}} u_{\hat{f}(z)}(z) \geq \frac{1}{N}\sum_{z\in\mathcal{S}} \max_{\rho\in\hat{\mathcal{P}}} u_\rho(z) - \epsilon. \tag{4}$$

For example, when algorithmic performance as a function of the parameters is piecewise constant, there are only a finite number of meaningfully different parameter values to choose among, one per piece. Then, since $\sum_{z\in\mathcal{S}} \max_{\rho\in\hat{\mathcal{P}}} u_\rho(z)$ is a submodular function of the portfolio $\hat{\mathcal{P}}$, we can use a greedy algorithm to select the portfolio $\hat{\mathcal{P}}$, and we obtain $\alpha = 1 - \frac{1}{e}$ and $\beta = 0$, as we prove in Appendix C. Alternatively, an integer programming technique could be used to select the optimal portfolio from the finite set of candidate parameter values, in which case we would obtain $\alpha = 1$ and $\beta = 0$. Moreover, the value ϵ can be calculated directly from the training set.

The following theorem bounds the difference between the expected performance of the chosen selector $\hat{f}$ and an oracle that selects an optimal selector and an optimal portfolio. The full proof is in Appendix B.

Theorem 5.2. *Suppose that each dual function u_z^* is piecewise constant with at most t pieces. Given a training set $\mathcal{S} \subseteq \mathcal{Z}$ of size N, suppose we learn an $(\alpha, \beta, \epsilon)$-optimal portfolio $\hat{\mathcal{P}} \subset \mathbb{R}$ and algorithm selector $\hat{f} : \mathcal{Z} \to \hat{\mathcal{P}}$ in $\mathcal{F}$. With probability $1-\delta$ over the draw of the training set $\mathcal{S} \sim \mathcal{D}^N$,*

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[u_{\hat{f}(z)}(z)\right] \geq \alpha \max_{\mathcal{P}:|\mathcal{P}|\leq\kappa} \mathbb{E}\left[\max_{\rho\in\mathcal{P}} u_\rho(z)\right] - \epsilon - \beta - \tilde{O}\left(H\sqrt{\frac{\bar{d}+\kappa}{N}}\right),$$

where $\bar{d}$ is the Natarajan dimension of $\bar{\mathcal{F}}$.

Proof sketch. First, let $\mathcal{P}^*$ be the optimal portfolio in the sense that

$$\mathcal{P}^* = \mathrm{argmax}_{\mathcal{P}\subset\mathbb{R}:|\mathcal{P}|\leq\kappa} \mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[\max_{\rho\in\mathcal{P}} u_\rho(z)\right].$$

We use a Hoeffding bound to relate the expected performance of $\mathcal{P}^*$ under the oracle algorithm selector and its average performance over the training set. We then use Definition 5.1 to relate the

latter quantity to the average performance of the learned selector $\hat{f}$ over the training set. Finally, we use Theorem 3.4 to relate the average performance of $\hat{f}$ to its expected performance. Putting all of these bounds together, we prove the theorem statement. □

By a parallel argument, we can obtain symmetric guarantees when our goal is to minimize rather than maximize a performance measure.

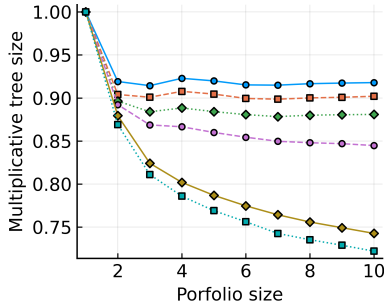
6 Experiments

We provide experiments that illustrate the tradeoff we investigated from a theoretical perspective in the previous sections: as we increase the portfolio size, we can hope to include a well-suited parameter setting for any problem instance, but it becomes increasingly difficult to avoid overfitting. We illustrate this in the context of integer programming algorithm configuration. We configure CPLEX, one of the most widely used commercial solvers. CPLEX uses the *branch-and-cut (B&C)* algorithm (branch-and-bound with cutting planes, primal heuristics, preprocessing, etc.) to solve integer programs (IPs). We tune an important parameter $\rho \in [0, 1]$ of CPLEX that controls its *variable selection policy*[2] and has been studied extensively in prior research [1, 5, 12, 13, 17, 25]. We leave CPLEX's other techniques on and unchanged in order to compare against the state of the art. We provide a more detailed overview of CPLEX and the parameter we tune in Appendix D. At a high level, B&C partitions the IP's feasible region, finding locally optimal solutions within the regions of the partition, and eventually verifies that the best solution found so far is globally optimal. It organizes this partition as a tree. As in prior research [5, 18, 40], our goal is to find parameter settings leading to small trees, so we define $u_\rho(z)$ to be the size of the tree B&C builds. We aim to learn a portfolio $\hat{\mathcal{P}}$ and selector $\hat{f}$ resulting in small expected tree size $\mathbb{E}\left[u_{\hat{f}(z)}(z)\right]$.

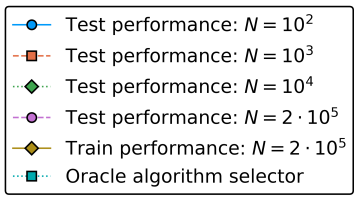
Distribution over IPs. We analyze a distribution over IPs formulating the combinatorial auction winner determination problem under the OR-bidding language [31], which we generate using the Combinatorial Auction Test Suite [24]. We use the "arbitrary" generator with 200 bids and 100 goods, resulting in IPs with around 200 variables, and the "regions" generator with 400 bids and 200 goods, resulting in IPs with around 400 variables. We define a heterogeneous distribution $\mathcal{D}$ as follows: with equal probability, we draw an instance from the "arbitrary" or "regions" distribution. To assign features to these IPs, we use all the features developed in prior research by Leyton-Brown et al. [24] and Hutter et al. [21], resulting in 140 features.

Experimental procedure. We first learn a portfolio of size 10 in the following way. We draw a training set of $M = 1000$ IPs $z_1, \ldots, z_M \sim \mathcal{D}$ and solve for the dual functions $u^*_{z_1}, \ldots, u^*_{z_M}$—which measure tree size as a function of the parameter ρ—using the algorithm described in Appendix D.1 of the paper by Balcan et al. [5]. These functions are piecewise-constant with at most t pieces, for some $t \in \mathbb{N}$. Therefore, there are at most Mt parameter settings leading to different algorithmic performance over the training set. Let $\bar{\mathcal{P}}$ be this set of parameter settings. We use a greedy algorithm to select 10 parameter settings from $\bar{\mathcal{P}}$. First, we find a parameter setting ρ_1 which minimizes average tree size over the training set: $\rho_1 \in \operatorname{argmin} \sum_{i=1}^{M} u_{z_i^*}(\rho)$. Then, we find a parameter setting ρ_2 that minimizes average tree size when the better of ρ_1 or ρ_2 is used:

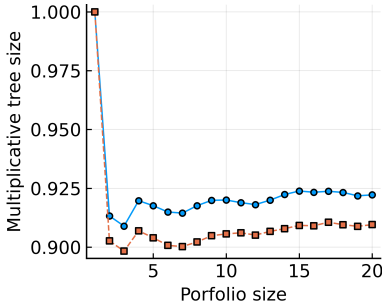
[2]We override the default variable selection of CPLEX 12.8.0.0 using the C API. All experiments were run on a 64-core machine with 512 GB of RAM, a m4.16xlarge Amazon AWS instance, and a cluster of m4.xlarge Amazon AWS instances.



(a) Plot with portfolio sizes 1 through 10.



(b) Legend for Figures 1a and 1c.



(c) Plot with portfolio sizes 1 through 20.

Figure 1: In Figures 1a and 1c, we plot the multiplicative tree size improvement we obtain as we increase both the portfolio size along the horizontal axis and the size of the training set, denoted N. Fixing a training set size and letting $\hat{v}_\kappa$ be the average tree size we obtain over the test set using a portfolio of size κ (see Equation (5)), we plot $\hat{v}_\kappa / \hat{v}_1$. In Figure 1a, the portfolio size ranges from 1 to 10 and the training set size N ranges from 100 to 200,000. In Figure 1c, the portfolio size ranges from 1 to 20 and the training set size ranges from 100 to 1000. In Figure 1a, we also plot a similar curve for the test performance of the oracle algorithm selector, as well as the training performance of the learned algorithm selector when $N = 2 \cdot 10^5$.

$\rho_2 \in \operatorname{argmin} \sum_{i=1}^{M} \min \left\{ u_{z_i^*}(\rho), u_{z_i^*}(\rho_1) \right\}$. We continue greedily until we have a portfolio $\hat{\mathcal{P}} = \{\rho_1, \ldots, \rho_{10}\}$.

We then use a regression forest to select among parameter settings in the portfolio $\hat{\mathcal{P}}$. Prior research [21] has illustrated that regression forests can be strong predictors of B&C runtime. Here, we use them to predict B&C tree size. A regression forest is a set $F = \{T_1, \ldots, T_M\}$ of regression trees (which we reviewed in Section 4.2). On an input IP z, the regression forest's prediction, denoted $h_F(z)$, is the average of the trees' predictions: $h_F(z) = \frac{1}{M} \sum_{i=1}^{M} h_{T_i}(z)$. We learn regression forests $F_1, \ldots, F_{10}$ for each of the 10 parameter settings in the portfolio $\hat{\mathcal{P}}$. We then define the algorithm selector $\hat{f}(z) = \rho_i$ where $i = \operatorname{argmin} \{h_{F_1}(z), \ldots, h_{F_{10}}(z)\}$.

To learn the regression forest, we draw a training set $z_1, \ldots, z_N \sim \mathcal{D}$ of IPs (with N specified below). For each parameter setting $\rho_i \in \hat{\mathcal{P}}$ and IP z_j, we compute $u_{\rho_i}(z_j)$, the size of the tree B&C builds using the parameter setting ρ_i. We then train the regression forest F_i corresponding to the parameter setting ρ_i using the labeled training set $\{(z_1, u_{z_1}(\rho_i)), \ldots, (z_N, u_{z_N}(\rho_i))\}$. We use Python's scikit-learn regression forest implementation [30] with the default parameter settings.

In Figure 1a, we plot the performance of the regression forests as we increase the sizes of both the training set and the portfolio. We denote the training set size as N, which ranges from 100 to 200,000. For a given choice of N, we first train the 10 regression forests $F_1, \ldots, F_{10}$ using the method described above. We then evaluate performance as a function of the portfolio size. Specifically, for each portfolio size $\kappa \in [10]$, we define an algorithm selector $\hat{f}_\kappa(z) = \rho_i$ where $i = \operatorname{argmin} \{h_{F_1}(z), \ldots, h_{F_\kappa}(z)\}$. We draw $N_t = 10^4$ test instances $\mathcal{S}_t \sim \mathcal{D}^{N_t}$ and evaluate the performance of $\hat{f}_\kappa$ on the test set. We denote the average test performance as

$$\hat{v}_\kappa = \frac{1}{N_t} \sum_{z \in \mathcal{S}_t} u_{\hat{f}_\kappa(z)}(z) \,. \tag{5}$$

In Figure 1a, we plot the multiplicative performance improvement we obtain as we increase κ. Specifically, we plot $\hat{v}_\kappa / \hat{v}_1$. These are the blue solid ($N = 10^2$), orange dashed ($N = 10^3$), green dotted ($N = 10^4$), and purple dashed ($N = 2 \cdot 10^5$) lines. By the iterative fashion we constructed the

portfolio, $\hat{v}_1$ is the performance of the best single parameter setting for the particular distribution, so $\hat{v}_1$ is already highly optimized.

We plot a similar curve for the test performance of the oracle algorithm selector which always selects the optimal parameter setting from the portfolio. Specifically, for each portfolio size $\kappa \in [10]$, let f^*_κ be the oracle algorithm selector $f^*_\kappa(z) = \text{argmin}_{\rho_1,\ldots,\rho_\kappa} u_{\rho_i}(z)$. Given a test set $\mathcal{S}_t \sim \mathcal{D}^{N_t}$, we define the average test performance of f^*_κ as

$$v^*_\kappa = \frac{1}{N_t} \sum_{z \in \mathcal{S}_t} u_{f^*_\kappa(z)}(z).$$

The blue dotted line equals v^*_κ / v^*_1 as a function of κ.

Finally, when the training set is of size $N = 2 \cdot 10^5$, we provide a similar curve for the training performance of the learned algorithm selectors $\hat{f}_\kappa$. Letting $z_1, \ldots, z_N$ be the training set, we denote the average training performance as

$$\tilde{v}_\kappa = \frac{1}{N} \sum_{i=1}^{N} u_{f^*_\kappa(z_i)}(z_i).$$

The yellow solid line equals $\tilde{v}_\kappa / \tilde{v}_1$ as a function of the portfolio size κ.

In Figure 1c, we plot $\hat{v}_\kappa / \hat{v}_1$ as a function of the portfolio size κ for larger portfolio sizes ranging from 1 to 20. We greedily extend the portfolio $\hat{\mathcal{P}}$ to include an additional 20 parameter settings. We then train 20 regression forests using freshly drawn training sets of size 100 and 1000. This plot illustrates the fact that as we increase the portfolio size, overfitting causes test performance to worsen.

Discussion. Focusing first on test performance using the largest training set size $N = 2 \cdot 10^5$, we see that test performance continues to improve as we increase the portfolio size, though training and test performance steadily diverge. This illustrates the tradeoff we investigated from a theoretical perspective in this paper: as we increase the portfolio size, we can hope to include a well-suited parameter setting for every instance, but the generalization error will worsen. Figure 1c shows that for a given training set size, there is a portfolio size after which test performance actually starts to get strictly worse, as our theory predicts. In other words, we observe overfitting: the learned algorithm selector has strong average performance over the training set but poor test performance.

7 Conclusions

We provided guarantees for learning a portfolio of parameter settings in conjunction with an algorithm selector for that portfolio. We provided a tight (up to log factors) bound on the number of samples sufficient and necessary to ensure that the selector's average performance on the training set generalizes to its expected performance on the real unknown problem instance distribution. Our guarantees apply in the widely-applicable setting where the algorithm's performance on any input problem instance is a piecewise-constant function of its parameters. Our theoretical bounds indicate that even with an extremely simple algorithm selector, we cannot hope to avoid overfitting in the worst-case if the portfolio is large. Thus, there is a tradeoff when increasing the portfolio size, since a large portfolio allows for the possibility of including a strong parameter setting for every instance, but this potential for performance improvement is overshadowed by a worsening propensity towards overfitting. We concluded with experiments illustrating this tradeoff in the context of integer programming. A direction for future research is to understand how the diversity of

a portfolio impacts its generalization error, since algorithm portfolios are often expressly designed to be diverse.

Acknowledgments

This material is based on work supported by the National Science Foundation under grants CCF-1535967, CCF-1733556, CCF-1910321, IIS-1617590, IIS-1618714, IIS-1718457, IIS-1901403, and SES-1919453; the ARO under awards W911NF1710082 and W911NF2010081; the Defense Advanced Research Projects Agency under cooperative agreement HR00112020003; an AWS Machine Learning Research Award; an Amazon Research Award; a Bloomberg Research Grant; a Microsoft Research Faculty Fellowship; an IBM PhD fellowship; and a fellowship from Carnegie Mellon University's Center for Machine Learning and Health.

References

[1] Tobias Achterberg. SCIP: solving constraint integer programs. *Mathematical Programming Computation*, 1(1):1–41, 2009.

[2] Martin Anthony and Peter Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 2009.

[3] Maria-Florina Balcan. Data-driven algorithm design. In Tim Roughgarden, editor, *Beyond Worst Case Analysis of Algorithms*. Cambridge University Press, 2020.

[4] Maria-Florina Balcan, Vaishnavh Nagarajan, Ellen Vitercik, and Colin White. Learning-theoretic foundations of algorithm configuration for combinatorial partitioning problems. *Conference on Learning Theory (COLT)*, 2017.

[5] Maria-Florina Balcan, Travis Dick, Tuomas Sandholm, and Ellen Vitercik. Learning to branch. In *International Conference on Machine Learning (ICML)*, 2018.

[6] Maria-Florina Balcan, Travis Dick, and Ellen Vitercik. Dispersion for data-driven algorithm design, online learning, and private optimization. In *Proceedings of the Annual Symposium on Foundations of Computer Science (FOCS)*, 2018.

[7] Maria-Florina Balcan, Travis Dick, and Colin White. Data-driven clustering via parameterized Lloyd's families. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[8] Maria-Florina Balcan, Dan DeBlasio, Travis Dick, Carl Kingsford, Tuomas Sandholm, and Ellen Vitercik. How much data is sufficient to learn high-performing algorithms? *arXiv preprint arXiv:1908.02894*, 2019.

[9] Maria-Florina Balcan, Travis Dick, and Manuel Lang. Learning to link. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[10] Maria-Florina Balcan, Tuomas Sandholm, and Ellen Vitercik. Learning to optimize computational resources: Frugal training with generalization guarantees. *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[11] Maria-Florina Balcan, Tuomas Sandholm, and Ellen Vitercik. Refined bounds for algorithm configuration: The knife-edge of dual class approximability. In *International Conference on Machine Learning (ICML)*, 2020.

[12] Evelyn Beale. Branch and bound methods for mathematical programming systems. *Annals of Discrete Mathematics*, 5:201–219, 1979.

[13] Michel Bénichou, Jean-Michel Gauthier, Paul Girodet, Gerard Hentges, Gerard Ribière, and O Vincent. Experiments in mixed-integer linear programming. *Mathematical Programming*, 1 (1):76–94, 1971.

[14] R. C. Buck. Partition of space. *Amer. Math. Monthly*, 50:541–544, 1943. ISSN 0002-9890.

[15] Isabel Cenamor, Tomás De La Rosa, and Fernando Fernández. The IBaCoP planning system: Instance-based configured portfolios. *Journal of Artificial Intelligence Research*, 56:657–691, 2016.

[16] Vikas Garg and Adam Kalai. Supervising unsupervised learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*. 2018.

[17] J-M Gauthier and Gerard Ribière. Experiments in mixed-integer linear programming using pseudo-costs. *Mathematical Programming*, 12(1):26–47, 1977.

[18] Prateek Gupta, Maxime Gasse, Elias Khalil, Pawan Mudigonda, Andrea Lodi, and Yoshua Bengio. Hybrid models for learning to branch. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[19] Rishi Gupta and Tim Roughgarden. A PAC approach to application-specific algorithm selection. *SIAM Journal on Computing*, 46(3):992–1017, 2017.

[20] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

[21] Frank Hutter, Lin Xu, Holger H Hoos, and Kevin Leyton-Brown. Algorithm runtime prediction: Methods & evaluation. *Artificial Intelligence*, 206:79–111, 2014.

[22] Serdar Kadioglu, Yuri Malitsky, Meinolf Sellmann, and Kevin Tierney. ISAC—instance-specific algorithm configuration. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, 2010.

[23] Kevin Leyton-Brown. *Resource allocation in competitive multiagent systems*. PhD thesis, Stanford University, 2003.

[24] Kevin Leyton-Brown, Mark Pearson, and Yoav Shoham. Towards a universal test suite for combinatorial auction algorithms. In *Proceedings of the ACM Conference on Electronic Commerce (ACM-EC)*, pages 66–76, Minneapolis, MN, 2000.

[25] Jeff Linderoth and Martin Savelsbergh. A computational study of search strategies for mixed integer programming. *INFORMS Journal of Computing*, 11(2):173–187, 1999.

[26] Shengcai Liu, Ke Tang, Yunwei Lei, and Xin Yao. On performance estimation in automatic algorithm configuration. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[27] Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

[28] George Nemhauser and Laurence Wolsey. *Integer and Combinatorial Optimization*. John Wiley & Sons, 1999.

[29] Sergio Núñez, Daniel Borrajo, and Carlos Linares López. Automatic construction of optimal static sequential portfolios for ai planning and beyond. *Artificial Intelligence*, 226:75–101, 2015.

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[31] Tuomas Sandholm. Algorithm for optimal winner determination in combinatorial auctions. *Artificial Intelligence*, 135:1–54, January 2002.

[32] Tuomas Sandholm. Very-large-scale generalized combinatorial multi-attribute auctions: Lessons from conducting $60 billion of sourcing. In Zvika Neeman, Alvin Roth, and Nir Vulkan, editors, *Handbook of Market Design*. Oxford University Press, 2013.

[33] Tzur Sayag, Shai Fine, and Yishay Mansour. Combining multiple heuristics. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 242–253. Springer, 2006.

[34] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[35] Matthew Streeter and Daniel Golovin. An online algorithm for maximizing submodular functions. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 1577–1584, 2009.

[36] Matthew Streeter, Daniel Golovin, and Stephen F. Smith. Combining multiple heuristics online. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2007.

[37] Vladimir Vapnik and Alexey Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[38] L. Xu, F. Hutter, H.H. Hoos, and K. Leyton-Brown. Satzilla: portfolio-based algorithm selection for SAT. *Journal of Artificial Intelligence Research*, 32(1):565–606, 2008.

[39] Lin Xu, Holger Hoos, and Kevin Leyton-Brown. Hydra: Automatically configuring algorithms for portfolio-based selection. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2010.

[40] Giulia Zarpellon, Jason Jo, Andrea Lodi, and Yoshua Bengio. Parameterizing branch-and-bound search trees to learn branching policies. *arXiv preprint arXiv:2002.05120*, 2020.

A Sample complexity proofs from Section 3

Theorem 3.4. *Suppose each dual function u_z^* is piecewise-constant with at most t pieces. Let $\bar{d}$ be the Natarajan dimension of $\bar{\mathcal{F}}$. Then* $\operatorname{Pdim}\left(\mathcal{U}_{\mathcal{F}}\right) = \tilde{O}\left(\bar{d} + \kappa \log t\right)$.

Proof. Let $z_1, \ldots, z_N \in \mathcal{Z}$ be a set of problem instances that is shattered by $\mathcal{U}_{\mathcal{F}}$, as witnessed by the points $t_1, \ldots, t_N \in \mathbb{R}$. By definition, this means that

$$2^N = \left| \left\{ \begin{pmatrix} \mathbf{1}_{\{u_{f(z_1)}(z_1) \leq t_1\}} \\ \vdots \\ \mathbf{1}_{\{u_{f(z_N)}(z_N) \leq t_N\}} \end{pmatrix} : f \in \mathcal{F} \right\} \right| \leq \left| \left\{ \begin{pmatrix} u_{f(z_1)}(z_1) \\ \vdots \\ u_{f(z_N)}(z_N) \end{pmatrix} : f \in \mathcal{F} \right\} \right|. \tag{6}$$

Since each dual function $u^*_{z_i}$ is piecewise-constant with at most t pieces, we know there are $M \leq Nt$ intervals $I_1, \ldots, I_M$ partitioning $\mathbb{R}$ where for any interval I_j and any problem instance z_i, $u^*_{z_i}(\rho)$ is constant across all $\rho \in I_j$. We assume that the intervals are ordered so that if $j < j'$, then the points in I_j are smaller than the points in $I_{j'}$

Let $J = (j_1, \ldots, j_{\bar{\kappa}}) \in [M]^{\bar{\kappa}}$ be a vector of $\bar{\kappa} \leq \kappa$ interval indices with $j_1 \leq j_2 \leq \cdots \leq j_{\bar{\kappa}}$. Let $\mathcal{F}_J \subseteq \mathcal{F}$ be the set of functions $f \in \mathcal{F}$ with the following property: letting $\rho_1 < \rho_2 < \cdots < \rho_{\bar{\kappa}}$ be the parameter settings f maps to (i.e., $\{f(z) : z \in \mathcal{Z}\} = \{\rho_1, \ldots \rho_{\bar{\kappa}}\}$), we have that the i^{th} parameter setting is in the i^{th} interval: $\rho_1 \in I_{j_1}, \ldots, \rho_{\bar{\kappa}} \in I_{j_{\bar{\kappa}}}$. Since $I_1, \ldots, I_M$ partition $\mathbb{R}$ and since each function $f \in \mathcal{F}$ maps to at most κ parameter settings, $\mathcal{F} = \cup_J \mathcal{F}_J$. Together with Equation (6), this means that

$$2^N \leq \sum_{\bar{\kappa}=1}^{\kappa} \sum_{J \in [M]^{\bar{\kappa}}} \left| \left\{ \begin{pmatrix} u_{f(z_1)}(z_1) \\ \vdots \\ u_{f(z_N)}(z_N) \end{pmatrix} : f \in \mathcal{F}_J \right\} \right|. \tag{7}$$

Fix a particular set $J = (j_1, \ldots, j_{\bar{\kappa}}) \in [M]^{\bar{\kappa}}$ as defined above. For each algorithm selector $f \in \mathcal{F}_J$, let $f_0 : \mathcal{Z} \to J$ be a function that indicates which of the $\bar{\kappa}$ intervals $I_{j_1}, \ldots, I_{j_{\bar{\kappa}}}$ the parameter setting $f(z)$ falls in. In other words, $f_0(z) = j$ if and only if $f(z) \in I_j$. Recall that for any $i \in [N]$ and $j \in J$, $u_{f(z_i)}(z_i)$ is constant across all $f \in \mathcal{F}_J$ with $f(z_i) \in I_j$. Therefore, even if we only know which of the $\bar{\kappa}$ intervals $f(z_i)$ falls in and not the function f itself, we can correctly infer the value $u_{f(z_i)}(z_i)$. Said another way, if we only know the value $f_0(z_i) \in [\bar{\kappa}]$, we can infer the value $u_{f(z_i)}(z_i)$. Aggregating this logic across all N problem instances, given a vector $(f_0(z_1), \ldots, f_0(z_N))$ we can directly infer the vector $\big(u_{f(z_1)}(z_1), \ldots, u_{f(z_N)}(z_N)\big)$. This implies that

$$\left| \left\{ \begin{pmatrix} u_{f(z_1)}(z_1) \\ \vdots \\ u_{f(z_N)}(z_N) \end{pmatrix} : f \in \mathcal{F}_J \right\} \right| \leq \left| \left\{ \begin{pmatrix} f_0(z_1) \\ \vdots \\ f_0(z_N) \end{pmatrix} : f \in \mathcal{F}_J \right\} \right|. \tag{8}$$

Next, we use a similar logic to show that

$$\left| \left\{ \begin{pmatrix} f_0(z_1) \\ \vdots \\ f_0(z_N) \end{pmatrix} : f \in \mathcal{F}_J \right\} \right| \leq \left| \left\{ \begin{pmatrix} \bar{f}(z_1) \\ \vdots \\ \bar{f}(z_N) \end{pmatrix} : f \in \mathcal{F}_J \right\} \right|. \tag{9}$$

To see why, suppose we only know the value $\bar{f}(z_i)$ and not the function f itself. For ease of notation, say $\ell = \bar{f}(z_i)$. By definition of $\bar{f}$, we know that $f(z_i)$ is the ℓ^{th}-smallest parameter setting that the function f maps to. By definition of the function f_0, this implies that $f_0(z_i) = j_\ell$. Therefore, if we only know the value $\bar{f}(z_i)$ and not the function f itself, we can correctly infer the value $f_0(z_i)$. Again, aggregating this logic across all N problem instances, given a vector $\big(\bar{f}(z_1), \ldots, \bar{f}(z_N)\big)$ we can directly infer the vector $(f_0(z_1), \ldots, f_0(z_N))$. This implies that Equation (9) holds.

Combining Equations 8 and (9) with Natarajan's lemma [27], we have that

$$\left|\left\{\begin{pmatrix} u_{f(z_1)}(z_1) \\ \vdots \\ u_{f(z_N)}(z_N) \end{pmatrix} : f \in \mathcal{F}_J\right\}\right| \leq N^{\bar{d}}\bar{\kappa}^{2\bar{d}}.$$

Combining this fact, the fact that $M \leq Nt$, and Equation (7), we have that $2^N \leq \kappa(Nt)^\kappa N^{\bar{d}}\kappa^{2\bar{d}}$, which implies that $N = O\left(\left(\kappa+\bar{d}\right)\log\left(\kappa+\bar{d}\right)+\kappa\log t\right)$. □

Theorem 3.5. *For any $\kappa, \bar{d} \geq 2$, there is a class of functions $\mathcal{U} = \{u_\rho : \rho \in \mathbb{R}\}$ and a class of selectors $\mathcal{F}$ such that:*

1. *Each selector $f \in \mathcal{F}$ maps to $\leq \kappa$ parameter settings.*
2. *Each dual function u_z^* is piecewise-constant with 1 discontinuity,*
3. *The Natarajan dimension of $\bar{\mathcal{F}}$ is at most $\bar{d}$, and*
4. *The pseudo-dimension of $\mathcal{U}_\mathcal{F}$ is $\Omega\left(\kappa+\bar{d}\right)$.*

Proof. Let $\mathcal{Z} = (0,1]$. For each parameter setting $\rho \in \mathbb{R}$, define $u_\rho(z) = \mathbf{1}_{\{z \leq \rho\}}$. As claimed, each dual function $u_z^* : \mathbb{R} \to \mathbb{R}$ is piecewise-constant with 1 discontinuity. In this case, the function in $\mathcal{U}_\mathcal{F}$ map $\mathcal{Z}$ to $\{0,1\}$. In the special case where the range of the function class is $\{0,1\}$, pseudo-dimension is typically referred to as *VC dimension*, which we denote as $\text{VCdim}\left(\mathcal{U}_\mathcal{F}\right)$.

Let $\kappa, \bar{d} \geq 2$ be two arbitrary integers. We split this proof into two cases: $\bar{d} \geq \kappa$ and $\kappa > \bar{d}$. In both cases, we exhibit a class of selectors $\mathcal{F}$ that satisfies the properties in the theorem statement and we prove that $\text{VCdim}\left(\mathcal{U}_\mathcal{F}\right) \geq \max\left\{\kappa, \bar{d}\right\} = \Omega\left(\kappa+\bar{d}\right)$.

Claim A.1. *Suppose $\bar{d} \geq \kappa$. There exists a class of selectors $\mathcal{F}$ that satisfies the properties in the theorem statement and $\text{VCdim}\left(\mathcal{U}_\mathcal{F}\right) = \bar{d}$.*

Proof of Claim A.1. Let $\mathcal{F} \subseteq \{0,1\}^\mathcal{Z}$ be any set of binary functions with VC dimension $\bar{d}$. As required, each selector $f \in \mathcal{F}$ maps to at most κ parameter settings ($|\{f(z) : z \in \mathcal{Z}\}| \leq 2 \leq \kappa$). Moreover, $\bar{\mathcal{F}} = \mathcal{F}$, so the Natarajan dimension of $\bar{\mathcal{F}}$ equals the VC dimension of $\mathcal{F}$, which is $\bar{d}$.

For any instance $z \in \mathcal{Z}$ and function $f \in \mathcal{F}$,

$$u_{f(z)}(z) = \begin{cases} 1 & \text{if } z \leq f(z) \\ 0 & \text{if } z > f(z). \end{cases}$$

Since $z \in (0,1]$ and $f(z) \in \{0,1\}$, this implies that $u_{f(z)}(z) = f(z)$. Therefore, $\text{VCdim}\left(\mathcal{U}_\mathcal{F}\right) = \text{VCdim}(\mathcal{F}) = \bar{d}$. □

Claim A.2. *Suppose $\kappa > \bar{d}$. There exists a class of selectors $\mathcal{F}$ that satisfies the properties in the theorem statement and $\text{VCdim}\left(\mathcal{U}_\mathcal{F}\right) \geq \kappa$.*

Proof of Claim A.2. We begin by partitioning $\mathcal{Z} = (0,1]$ into κ intervals $Z_1, \ldots, Z_\kappa$, where $Z_i = \left(\frac{i-1}{\kappa}, \frac{i}{\kappa}\right]$. For each set $T \subseteq [\kappa]$, define an selector $f_T : \mathcal{Z} \to \mathbb{R}$ as follows. For any $z \in \mathcal{Z} = (0,1]$, let $i \in [\kappa]$ be the index of the interval z lies in, i.e., $z \in Z_i$. We define

$$f_T(z) = \begin{cases} \frac{i}{\kappa} & \text{if } i \in T \\ \frac{i}{\kappa} - \frac{1}{2\kappa} & \text{if } i \notin T. \end{cases}$$

Let $\mathcal{F} = \{f_T : T \subseteq [\kappa]\}$. For every function $f \in \mathcal{F}$, $\bar{f}(z)$ equals the index $i \in [\kappa]$ such that $z \in Z_i$. Therefore, $\left|\bar{\mathcal{F}}\right| = 1$, so the Natarajan dimension of $\bar{\mathcal{F}}$ is $0 < \bar{d}$.

Define $\mathcal{S} = \left\{\frac{1}{\kappa}, \frac{2}{\kappa}, \ldots, \frac{\kappa-1}{\kappa}, 1\right\} \subset \mathcal{Z}$. We prove that $\mathcal{S}$ is shattered by $\mathcal{U}_{\mathcal{F}}$. Let $T \subseteq [\kappa]$ be an arbitrary subset. If $i \in T$, then $f_T\left(\frac{i}{\kappa}\right) = \frac{i}{\kappa}$, so

$$u_{f_T\left(\frac{i}{\kappa}\right)}\left(\frac{i}{\kappa}\right) = u_{\frac{i}{\kappa}}\left(\frac{i}{\kappa}\right) = \mathbf{1}_{\left\{\frac{i}{\kappa} \leq \frac{i}{\kappa}\right\}} = 1.$$

If $i \notin T$, then $f_T\left(\frac{i}{\kappa}\right) = \frac{i}{\kappa} - \frac{1}{2\kappa}$, so

$$u_{f_T\left(\frac{i}{\kappa}\right)}\left(\frac{i}{\kappa}\right) = u_{\frac{i}{\kappa} - \frac{1}{2\kappa}}\left(\frac{i}{\kappa}\right) = \mathbf{1}_{\left\{\frac{i}{\kappa} \leq \frac{i}{\kappa} - \frac{1}{2\kappa}\right\}} = 0.$$

Therefore, $\mathcal{S}$ is shattered by $\mathcal{U}_{\mathcal{F}}$, so the VC dimension of $\mathcal{U}_{\mathcal{F}}$ is at least κ. □

These two claims illustrate that $\text{VCdim}\left(\mathcal{U}_{\mathcal{F}}\right) \geq \max\left\{\kappa, \bar{d}\right\} = \Omega\left(\kappa + \bar{d}\right)$. □

A.1 Regression tree performance models

Lemma 4.2. *Suppose we limit ourselves to building regression trees with at most ℓ leaves. Then the Natarajan dimension of $\bar{\mathcal{F}}_R$ is $O(\ell\kappa \log(\ell\kappa m))$.*

Proof. In this proof, to simplify notation, we will denote the feature vector $\phi(z)$ as $\boldsymbol{z} \in \mathbb{R}^m$. For each κ-tuple of regression trees $\boldsymbol{T} = (T_1, \ldots, T_\kappa)$, let $g_{\boldsymbol{T}} : \mathcal{Z} \to [\kappa]$ be a function where $g_{\boldsymbol{T}}(z) = \text{argmax}_{i \in [\kappa]}\{h_{T_i}(z)\}$. By definition, the set $\bar{\mathcal{F}}_R$ consists of the functions $g_{\boldsymbol{T}}$ across all κ-tuples of regression trees $\boldsymbol{T}$ with at most ℓ leaves. We will assume, without loss of generality, that all trees are full.

Let N be the the Natarajan dimension of $\bar{\mathcal{F}}_R$ and let $z_1, \ldots, z_N \in \mathcal{Z}$ be a set of N problem instances that are multi-class shattered by $\bar{\mathcal{F}}_R$. This implies that

$$2^N \leq \left| \left\{ \begin{pmatrix} g_{\boldsymbol{T}}(z_1) \\ \vdots \\ g_{\boldsymbol{T}}(z_N) \end{pmatrix} : \boldsymbol{T} \text{ is a } \kappa\text{-tuple of regression trees} \right\} \right|. \tag{10}$$

In this proof, we show that the right-hand-side of this inequality is bounded by $m^{\kappa(\ell-1)}(N\ell)^{\ell\kappa}(\kappa\ell)^{\kappa\ell}$, which implies that $N = O(\ell\kappa \log(\ell\kappa m))$.

To this end, we begin by focusing on a single regression tree T. We analyze the number of ways the tree can partition the instances $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ as we vary the parameters of T. Each internal node of T performances an inequality test on some feature of the input, so it is defined by a feature index $i \in [m]$ and a threshold $\theta \in \mathbb{R}$. Since there are ℓ leaves, there are $\ell - 1$ internal nodes. First, we fix the indices of all internal nodes, which leaves $\ell - 1$ real-valued thresholds to analyze. At a particular internal node ν, let j be the index of the feature on which the node performs an inequality test and let θ_ν be the threshold (where the index j is fixed by the threshold θ_ν is not fixed). Whether or not the instance $\boldsymbol{z}_i$ would be sorted into the left or right child of the node depends on whether or not

$$z_i[j] \leq \theta_\nu \tag{11}$$

(where $z_i[j]$ is the j^{th} coordinate of the vector $\boldsymbol{z}_i$). For each problem instance $\boldsymbol{z}_i$, there are therefore $\ell - 1$ hyperplanes splitting the set of thresholds $\mathbb{R}^{\ell-1}$ into regions where if we use thresholds from within any one region, the path that the instance $\boldsymbol{z}_i$ takes through the tree (from root to leaf) is

constant. The same holds for all N problem instances, leading to a total of $N(\ell-1)$ hyperplanes in $\mathbb{R}^{\ell-1}$. In total, these hyperplanes split $\mathbb{R}^{\ell-1}$ into at most $(N\ell)^\ell$ regions where if we use the thresholds from within any one region, the path that each of the N problem instances takes through the tree is constant [14]. Since this is true no matter how we fix the feature indices of each interval node, tuning all parameters of the tree T (both the feature indices and the thresholds) can induce at most $m^{\ell-1}(N\ell)^\ell$ different partitions of the N problem instances.

Said another way, for any tree T and instance $z \in \mathcal{Z}$, let $\lambda_T(z) \in [\ell]$ be the index of the leaf that the instance z is mapped to as we apply the inequality tests defined by the internal nodes of T. As we vary the tree T, the vector $(\lambda_T(z_1), \ldots, \lambda_T(z_N)) \in [\ell]^N$ will take on at most $m^{\ell-1}(N\ell)^\ell$ different values.

We now aggregate this reasoning across all κ regression trees. For any instance $z \in \mathcal{Z}$ and any κ-tuple of regression trees $\boldsymbol{T} = (T_1, \ldots, T_\kappa)$, let $\Lambda_{\boldsymbol{T}}(z) \in [\ell]^\kappa$ be a vector where for each $j \in [\kappa]$, the j^{th} component of $\Lambda_{\boldsymbol{T}}(z)$ is the index of the leaf that the instance z is mapped to as we apply the inequality tests defined by the tree T_j. In other words, $\Lambda_{\boldsymbol{T}}(z) = (\lambda_{T_1}(z), \ldots, \lambda_{T_\kappa}(z))$. As we vary $\boldsymbol{T}$, the matrix

$$\begin{pmatrix} \Lambda_{\boldsymbol{T}}(z_1) & \ldots & \Lambda_{\boldsymbol{T}}(z_N) \end{pmatrix} = \begin{pmatrix} \lambda_{T_1}(z_1) & \cdots & \lambda_{T_1}(z_N) \\ \vdots & \ddots & \vdots \\ \lambda_{T_\kappa}(z_1) & \cdots & \lambda_{T_\kappa}(z_N) \end{pmatrix}$$

will take on at most $m^{(\ell-1)\kappa}(N\ell)^{\ell\kappa}$ different values. After all, the first row of the matrix can take on at most $m^{\ell-1}(N\ell)^\ell$ different values as we vary the tree T_1, the second row can take on at most $m^{\ell-1}(N\ell)^\ell$ different values as we vary T_2, and so on.

Now, consider the set of all κ-tuples of regression trees $\boldsymbol{T}$ where the matrix $(\Lambda_{\boldsymbol{T}}(z_1), \ldots, \Lambda_{\boldsymbol{T}}(z_N))$ is constant. Across all such $\boldsymbol{T} = (T_1, \ldots, T_\kappa)$, we know exactly which leaf each instance z_i maps to for all κ trees. For each each instance z_i, the tree with the largest label—or in other words, the value of the multi-class function $g_{\boldsymbol{T}}(z_i)$—only depends on the relative order of the leaves' predictions. Since there is a total of $\kappa\ell$ leaves, there are at most $(\kappa\ell)^{\kappa\ell}$ such orderings. Combining this bound with the bound from the previous paragraph, we have that

$$\left| \left\{ \begin{pmatrix} g_{\boldsymbol{T}}(z_1) \\ \vdots \\ g_{\boldsymbol{T}}(z_N) \end{pmatrix} : \boldsymbol{T} \text{ is a } \kappa\text{-tuple of regression trees} \right\} \right| \leq m^{(\ell-1)\kappa}(N\ell)^{\ell\kappa}(\kappa\ell)^{\kappa\ell}.$$

From Equation (10), we have that $2^N \leq m^{(\ell-1)\kappa}(N\ell)^{\ell\kappa}(\kappa\ell)^{\kappa\ell}$, so $N = O(\ell\kappa\log(\ell\kappa m))$. □

A.2 Clustering-based algorithm selectors

Lemma 4.4. *For any $p \in [1, \infty)$, the Natarajan dimension of $\bar{\mathcal{F}}_C$ is $O(m\kappa\log(m\kappa p))$.*

Proof. In this proof, to simplify notation, we will denote the feature vector $\phi(z)$ as $\boldsymbol{z} \in \mathbb{R}^m$. For each matrix $X \in \mathbb{R}^{m\times\kappa}$, let $g_X : \mathcal{Z} \to [\kappa]$ be a function where

$$g_X(z) = \text{argmin}_{i\in[\kappa]} \left\{ \|\boldsymbol{x}_i - \boldsymbol{z}\|_p^p \right\}.$$

By definition, $\bar{\mathcal{F}}_C = \{g_X : X \in \mathbb{R}^{m\times\kappa}\}$. Let N be the Natarajan dimension of $\bar{\mathcal{F}}_C$ and let $z_1, \ldots, z_N \in \mathcal{Z}$ be a set of N problem instances that are multi-class shattered by $\bar{\mathcal{F}}_C$. This implies that

$$2^N \leq \left| \left\{ \begin{pmatrix} g_X(z_1) \\ \vdots \\ g_X(z_N) \end{pmatrix} : X \in \mathbb{R}^{m\times\kappa} \right\} \right|. \tag{12}$$

In this proof, we analyze the partition of the parameter space $\mathbb{R}^{m\times\kappa}$ into regions where in any one region $R \subseteq \mathbb{R}^{m\times\kappa}$, across all matrices $X \in R$, the vector $(g_X(z_1), \ldots, g_X(z_N))$ is constant.

We begin by subdividing $\mathbb{R}^{m\times\kappa}$ into regions $P_1, \ldots, P_T \subseteq \mathbb{R}^{m\times\kappa}$ where in any one region P, across all $X \in P$, either the ℓ^{th} component of $\boldsymbol{z}_q$ is smaller than the ℓ^{th} component of $\boldsymbol{x}_j$, i.e. $z_q[\ell] \leq x_j[\ell]$, or vice versa (but not both) for all $q \in [N]$, $j \in [\kappa]$, and $\ell \in [m]$. This is partition is defined by $N\kappa m$ hyperplanes in $\mathbb{R}^{m\kappa}$, so there are $T \leq (N\kappa m + 1)^{m\kappa}$ such regions [14].

Next, fix one of these T regions $P \subseteq \mathbb{R}^{m\times\kappa}$. Without loss of generality, assume that the ℓ^{th} component of $\boldsymbol{z}_q$ is smaller than the ℓ^{th} component of $\boldsymbol{x}_j$, i.e. $z_q[\ell] \leq x_j[\ell]$ for all $q \in [N]$, $j \in [\kappa]$, and $\ell \in [m]$. For any two labels $i, j \in [\kappa]$ and any $q \in [N]$, whether or not

$$\|\boldsymbol{x}_i - \boldsymbol{z}_q\|_p^p \geq \|\boldsymbol{x}_j - \boldsymbol{z}_q\|_p^p \tag{13}$$

directly depends on the sign of the polynomial

$$h_{q,i,j}(X) := \sum_{\ell=1}^{m} (x_i[\ell] - z_q[\ell])^p - (x_j[\ell] - z_q[\ell])^p \,.$$

We know there are at most $\left(N\kappa^2 p\right)^{m\kappa}$ regions partitioning P so that in any one region R, across all $X \in R$, either $h_{q,i,j}(X) \leq 0$ or $h_{q,i,j}(X) > 0$ (but not both) for all $q \in [N]$ and $i, j \in [\kappa]$ [2]. For any such region R, across all $X \in R$, all pairwise comparisons as in Equation (13) are fixed, so the vector $(g_X(z_1), \ldots, g_X(z_N))$ is constant. In total, there are at most $(N\kappa m + 1)^{m\kappa}\left(N\kappa^2 p\right)^{m\kappa} \leq (2N^2\kappa^3 p\ell)^{m\kappa}$ regions, which implies that

$$\left| \left\{ \begin{pmatrix} g_X(z_1) \\ \vdots \\ g_X(z_N) \end{pmatrix} : X \in \mathbb{R}^{m\times\kappa} \right\} \right| \leq (2N^2\kappa^3 p\ell)^{m\kappa}.$$

Combining this inequality with Equation (12), we have that $2^N \leq (2N^2\kappa^3 p\ell)^{m\kappa}$, so $N = O(mk\log(m\kappa p))$. □

B Proof of Theorem 5.2

Theorem 5.2. *Suppose that each dual function u_z^* is piecewise constant with at most t pieces. Given a training set $\mathcal{S} \subseteq \mathcal{Z}$ of size N, suppose we learn an $(\alpha, \beta, \epsilon)$-optimal portfolio $\hat{\mathcal{P}} \subset \mathbb{R}$ and algorithm selector $\hat{f} : \mathcal{Z} \to \hat{\mathcal{P}}$ in $\mathcal{F}$. With probability $1-\delta$ over the draw of the training set $\mathcal{S} \sim \mathcal{D}^N$,*

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[u_{\hat{f}(z)}(z)\right] \geq \alpha \max_{\mathcal{P}:|\mathcal{P}|\leq\kappa} \mathbb{E}\left[\max_{\rho\in\mathcal{P}} u_\rho(z)\right] - \epsilon - \beta - \tilde{O}\left(H\sqrt{\frac{\bar{d}+\kappa}{N}}\right),$$

where $\bar{d}$ is the Natarajan dimension of $\bar{\mathcal{F}}$.

Proof. First, let

$$\mathcal{P}^* = \mathrm{argmax}_{\mathcal{P}\subset\mathbb{R}:|\mathcal{P}|\leq\kappa} \mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[\max_{\rho\in\mathcal{P}} u_\rho(z)\right].$$

A Hoeffding bound implies that with probability $1-\delta$,

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[\max_{\rho\in\mathcal{P}^*} u_\rho(z)\right] \leq \frac{1}{N}\sum_{z\in\mathcal{S}} \max_{\rho\in\mathcal{P}^*} u_\rho(z) + \tilde{O}\left(H\sqrt{\frac{1}{N}}\right).$$

Combining this inequality with Definition 5.1, we have that

$$\begin{aligned}\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[\max_{\rho\in\mathcal{P}^*} u_\rho(z)\right] &\le \frac{1}{\alpha}\left(\frac{1}{N}\sum_{z\in\mathcal{S}}\max_{\rho\in\hat{\mathcal{P}}} u_\rho(z)+\beta\right)+O\left(H\sqrt{\frac{1}{N}\log\frac{1}{\delta}}\right)\\ &\le \frac{1}{\alpha}\left(\frac{1}{N}\sum_{z\in\mathcal{S}} u_{\hat{f}(z)}(z)+\epsilon+\beta\right)+O\left(H\sqrt{\frac{1}{N}\log\frac{1}{\delta}}\right).\end{aligned}$$

From Theorem 3.4, we know that with probability $1-\delta$,

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[\max_{\rho\in\mathcal{P}^*} u_\rho(z)\right] \le \frac{1}{\alpha}\left(\mathop{\mathbb{E}}_{z\sim\mathcal{D}}\left[u_{\hat{f}(z)}(z)\right]+\tilde{O}\left(H\sqrt{\frac{\bar{d}+\kappa}{N}}\right)+\epsilon+\beta\right).$$

Therefore, the theorem statement holds. □

C Connection to submodularity

Since each dual function $u_z^*(\rho)$ is piecewise-constant with at most t pieces, on any training set $\mathcal{S}=\{z_1,\dots,z_N\}\subseteq\mathcal{Z}$, there are at most Nt parameter settings leading to different algorithmic performance over this training set. In other words,

$$\left|\left\{\begin{pmatrix}u_{z_1}^*(\rho)\\ \vdots\\ u_{z_N}^*(\rho)\end{pmatrix}:\rho\in\mathbb{R}\right\}\right|\le Nt.$$

Let $\bar{\mathcal{P}}\subseteq\mathbb{R}$ be a set of at most Nt parameters such that

$$\left\{\begin{pmatrix}u_{z_1}^*(\rho)\\ \vdots\\ u_{z_N}^*(\rho)\end{pmatrix}:\rho\in\mathbb{R}\right\}=\left\{\begin{pmatrix}u_{z_1}^*(\rho)\\ \vdots\\ u_{z_N}^*(\rho)\end{pmatrix}:\rho\in\bar{\mathcal{P}}\right\}.$$

For any $T\subseteq\bar{\mathcal{P}}$, let

$$U(T)=\sum_{i=1}^{N}\max_{\rho\in T} u_\rho(z_i).$$

Theorem C.1. *The function U is monotone and submodular.*

Proof. For any z_i, let $U_i:2^{\mathcal{P}^*}\to\mathbb{R}$ be the function $U_i(T)=\max_{\rho\in T}u_\rho(z_i)$. We will prove that each function U_i is submodular. The theorem then follows because the class of submodular functions is closed under non-negative linear combinations. To this end, let $T\subseteq\mathcal{P}^*$ be an arbitrary subset of $\mathcal{P}^*$ and let $\rho_1,\rho_2\in\mathcal{P}^*\setminus T$ be any two parameter settings in $\mathcal{P}^*$ but not in T. We want to prove that

$$\max_{\rho\in T\cup\{\rho_1\}} u_\rho(z_i)+\max_{\rho\in T\cup\{\rho_2\}} u_\rho(z_i)\ge\max_{\rho\in T\cup\{\rho_1,\rho_2\}} u_\rho(z_i)+\max_{\rho\in T} u_\rho(z_i). \tag{14}$$

Without loss of generality, suppose that $u_{\rho_1}(z_i)\ge u_{\rho_2}(z_i)$. Let $\bar{\rho}\in\operatorname{argmax}_{\rho\in T}u_\rho(z_i)$. There are three cases:

- In the first case, $u_{\bar{\rho}}(z)\ge u_{\rho_1}(z_i)\ge u_{\rho_2}(z_i)$, so

  $$\max_{\rho\in T\cup\{\rho_1\}} u_\rho(z_i)+\max_{\rho\in T\cup\{\rho_2\}} u_\rho(z_i)=2u_{\bar{\rho}}(z_i)=\max_{\rho\in T\cup\{\rho_1,\rho_2\}} u_\rho(z_i)+\max_{\rho\in T} u_\rho(z_i),$$

  so Equation (14) holds.

- In the second case, $u_{\rho_1}(z_i) \geq u_{\bar{\rho}}(z) \geq u_{\rho_2}(z_i)$, so

$$\max_{\rho \in T \cup \{\rho_1\}} u_\rho(z_i) + \max_{\rho \in T \cup \{\rho_2\}} u_\rho(z_i) = u_{\rho_1}(z_i) + u_{\bar{\rho}}(z) = \max_{\rho \in T \cup \{\rho_1, \rho_2\}} u_\rho(z_i) + \max_{\rho \in T} u_\rho(z_i),$$

  so Equation (14) holds.

- In the third and final case, $u_{\rho_1}(z_i) \geq u_{\rho_2}(z_i) \geq u_{\bar{\rho}}(z)$, so

$$\max_{\rho \in T \cup \{\rho_1\}} u_\rho(z_i) + \max_{\rho \in T \cup \{\rho_2\}} u_\rho(z_i) = u_{\rho_1}(z_i) + u_{\rho_2}(z_i) \geq u_{\rho_1}(z_i) + u_{\bar{\rho}}(z) = \max_{\rho \in T \cup \{\rho_1, \rho_2\}} u_\rho(z_i) + \max_{\rho \in T} u_\rho(z_i),$$

  so Equation (14) holds.

Therefore, the function U is monotone and submodular. □

For any cardinality constraint $\kappa \in \mathbb{N}$, let $\hat{\mathcal{P}} \subseteq \mathcal{P}^*$ be the set of κ parameter settings that the greedy algorithm selects to optimize the function U. Theorem C.1 implies that

$$\sum_{i=1}^{N} \max_{\rho \in \hat{\mathcal{P}}} u_\rho(z_i) \geq \left(1 - \frac{1}{e}\right) \max_{T \subseteq \mathcal{P} : |T| \leq \kappa} \sum_{i=1}^{N} \max_{\rho \in T} u_\rho(z_i).$$

D Additional details about experiments

Branch-and-bound. We begin with a high-level overview of branch-and-cut (B&C) and refer the reader to the textbook by Nemhauser and Wolsey [28], for example, for more details. B&C is an algorithm for solving integer programs (IPs). An IP is defined by an objective vector $\boldsymbol{c} \in \mathbb{R}^n$, a constraint matrix $A \in \mathbb{R}^{m \times n}$, a constraint vector $\boldsymbol{b} \in \mathbb{R}^m$, and a set of indices $I \subseteq [m]$. The goal is to solve the following optimization problem:

$$\begin{array}{ll} \text{maximize} & \boldsymbol{c} \cdot \boldsymbol{x} \\ \text{subject to} & A\boldsymbol{x} \leq \boldsymbol{b} \\ & x[i] \in \mathbb{Z} \quad \forall i \in I. \end{array} \tag{15}$$

In keeping with Section 2, we use the notation $z = (\boldsymbol{c}, A, \boldsymbol{b}, I)$ to denote the IP. B&C builds a search tree to solve an input IP z, with z stored at the root. It begins by solving the LP relaxation of the input IP z. We use the notation $\breve{\boldsymbol{x}}_z$ to denote the solution to this LP relaxation. B&C then uses a *variable selection policy* to choose a variable $i \in I$ and it *branches* on this variable. This means that it defines a new IP z_i^- which is identical to the original IP z but with the additional constraint that $x[i] \leq \lfloor \breve{x}_z[i] \rfloor$. It stores the IP z_i^- in the left child of the root node. Similarly, it defines another IP z_i^+ which is identical to the original IP z but with the additional constraint that $x[i] \geq \lfloor \breve{x}_z[i] \rfloor$. It stores the IP z_i^+ in the right child of the root node. It then uses a *node selection policy* to choose one of the two leaves and repeats this process—solving the LP relaxation of the node's IP, choosing a variable to branch on, and so on. Eventually, one of the solutions to an LP relaxation B&C solves will in fact be the optimal solution to the original IP (Equation (15)), and B&C will be able to verify its optimality (this verification procedure is straightforward, but we do not go into the details here).

Parameterized variable selection policy. We analyze a parameterized variable selection policy that has been studied extensively in prior research [1, 5, 11–13, 17, 25]. To define this variable selection policy, we use the notation $\breve{c}_{\bar{z}} = \boldsymbol{c} \cdot \breve{\boldsymbol{x}}_{\bar{z}}$ for any IP $\bar{z}$. Given a parameter setting $\rho \in [0, 1]$ and the IP $\bar{z}$ contained at the leaf of the search tree, this variable selection policy chooses to branch on the variable $i \in I$ that maximizes

$$(1 - \rho) \min \left\{ \breve{c}_{\bar{z}} - \breve{c}_{\bar{z}_i^+}, \breve{c}_{\bar{z}} - \breve{c}_{\bar{z}_i^-} \right\} + \rho \max \left\{ \breve{c}_{\bar{z}} - \breve{c}_{\bar{z}_i^+}, \breve{c}_{\bar{z}} - \breve{c}_{\bar{z}_i^-} \right\}.$$

The parameter ρ thus balances a pessimistic approach to branching—which always chooses the variable leading to the minimal change in the LP objective value—with an optimistic approach—which chooses the variable leading to the maximal change in the LP objective value.