The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Optimal Sparse Recovery with Decision Stumps

Kiarash Banihashem, MohammadTaghi Hajiaghayi, Max Springer

University of Maryland
kiarash@umd.edu, hajiagha@cs.umd.edu, mss423@umd.edu

Abstract

Decision trees are widely used for their low computational cost, good predictive performance, and ability to assess the importance of features. Though often used in practice for feature selection, the theoretical guarantees of these methods are not well understood. We here obtain a tight finite sample bound for the feature selection problem in linear regression using single-depth decision trees. We examine the statistical properties of these "decision stumps" for the recovery of the s active features from p total features, where $s \ll p$. Our analysis provides tight sample performance guarantees on high-dimensional sparse systems which align with the finite sample bound of $O(s \log p)$ as obtained by Lasso, improving upon previous bounds for both the median and optimal splitting criteria. Our results extend to the non-linear regime as well as arbitrary sub-Gaussian distributions, demonstrating that tree based methods attain strong feature selection properties under a wide variety of settings and further shedding light on the success of these methods in practice. As a byproduct of our analysis, we show that we can provably guarantee recovery even when the number of active features s is unknown. We further validate our theoretical results and proof methodology using computational experiments.

Introduction

Decision trees are one of the most popular tools used in machine learning. Due to their simplicity and interpretability, trees are widely implemented by data scientist, both individually, and in aggregation with ensemble methods such as random forests and gradient boosting (Friedman 2001).

In addition to their predictive accuracy, tree based methods are an important tool used for the *variable selection* problem: identifying a relevant small subset of a high-dimensional feature space of the input variables that can accurately predict the output. When the relationship between the variables is linear, it has long been known that LASSO achieves the optimal sample complexity rate for this problem (Wainwright 2009a). In practice, however, tree-based methods have been shown to be preferable as they scale linearly with the size of the data and can capture non-linear relationships between the variables (Xu et al. 2014).

Notably, various tree structured systems are implemented for this variable selection task across wide-spanning domains. For example, tree based importance measures have been used for prediction of financial distress (Qian et al. 2022), wireless signal recognition (Li and Wang 2018), credit scoring (Xia et al. 2017), and the discovery of genes in the field of bioinformatics (Breiman 2001; Bureau et al. 2005; Huynh-Thu et al. 2012; Lunetta et al. 2004) to name a small fraction.

Despite this empirical success however, the theoretical properties of these tree based methods for feature selection are not well understood. While several papers have considered variants of this problem (see the Related Work section for an overview), even in the simple linear case, the sample complexity of the decision tree is not well-characterized.

In this paper, we attempt to bridge this gap and analyze the variable selection properties of single-level decision trees, commonly referred to as "decision stumps" (DSTUMP). Considering both linear and non-linear settings, we show that DSTUMP achieves nearly-tight sample complexity rates for a variety of practical sample distributions. Compared to prior work, our analysis is simpler and applies to different variants of the decision tree, as well as more general function classes.

The remainder of the paper is organized as follows: in the next section we introduce our main results on the finite sample guarantees of DSTUMP , in the Related Work section we discuss important prior results in the field of sparse recovery and where they fall flat, in the Algorithm Description section we present the recovery algorithm, and in the subsequent section we provide theoretical guarantees for the procedure under progressively more general problem assumptions. The proofs of these results are provided in the so-labeled section. We supplement the theoretical results with computational simulations in the Experimental Results and provide concluding remarks in the final section.

Our Results

We assume that we are given a dataset $\mathcal{D} = \{(\mathbf{X}_{i,:}, Y_i)\}_{i=1}^n$ consisting of n samples from the *non-parametric regression* model $Y_i = \sum_k f_k(\mathbf{X}_{i,k}) + W_i$ where i denotes the sample number, $\mathbf{X}_{i,:} \in \mathbb{R}^p$ is the input vector with corresponding output $Y_i \in \mathbb{R}$, $W_i \in \mathbb{R}$ are i.i.d noise values and $f_k : \mathbb{R} \to \mathbb{R}$ are univariate functions that are *s-sparse*: A set of univariate functions $\{f_k\}_{k\in[p]}$ is *s-sparse* on feature set $[p]$ if there exists a set $S \subseteq [p]$ with size $s = |S| \ll p$ such that $f_k = 0 \iff k \notin S$. Given access to $(\mathbf{X}_{i,:}, Y_i)_{i=1}^n$, we consider the *sparse recovery* problem, where we attempt to reveal the

set S with only a minimal number of samples. Note that this is different from the *prediction* problem where the goal is to learn the functions f_k. In accordance with prior work (Han et al. 2020; Kazemitabar et al. 2017; Klusowski 2020) we assume the $\mathbf{X}_{i,j}$ are i.i.d draws from the uniform distribution on $[0, 1]$ with Gaussian noise, $W_i \sim \mathcal{N}(0, 1)$. In the Main Results section, we will discuss how this assumption can be relaxed using our non-parametric results to consider more general distributions.

For the recovery problem, we consider the DSTUMP algorithm as first proposed by (Kazemitabar et al. 2017). The algorithm, shown in Algorithm 1, fits a single-level decision tree (stump) on each feature using either the "median split" or the "optimal split" strategy and ranks the features by the error of the corresponding trees. The median split provides a substantially simplified implementation as we do not need to optimize the stump construction, further providing an improved run time over the comparable optimal splitting procedure. Indeed, the median and optimal split have time complexity at most $\mathcal{O}(np)$ and $\mathcal{O}(np\log(n))$ respectively.

In spite of this simplification, we show that in the widely considered case of *linear design*, where the f_k are linear, DSTUMP can correctly recover S with a sample complexity bound of $\mathcal{O}(s\log p)$, matching the minimax optimal lower bound for the problem as achieved by LASSO (Wainwright 2009a,b). This result is noteworthy and surprising since the DSTUMP algorithm (and decision trees in general) is not designed with a linearity assumption, as is the case with LASSO . For this reason, trees are in general utilized for their predictive power in a non-linear model, yet our work proves their value in the opposite. We further extend these results for non-linear models and general sub-Gaussian distributions, improving previously known results using simpler analysis. In addition, our results do not require the sparsity level s to be known in advance. We summarize our main technical results as follows:

- We obtain a sample complexity bound of $\mathcal{O}(s\log p)$ for the DSTUMP algorithm in the linear design case, matching the optimal rate of LASSO and improving prior bounds in the literature for both the median and optimal split. This is the first *tight* bound on the sample complexity of single-depth decision trees used for sparse recovery and significantly improves on the existing results.

- We extend our results to the case of non-linear f_k, obtaining tighter results for a wider class of functions compared to the existing literature. We further use these results to generalize our analysis to sub-Gaussian distributions via the extension to nonlinear f_k.

- As a byproduct of our improved analysis, we show that our results hold for the case where the number of active features s is not known. This is the first theoretical guarantee on decision stumps that does not require s to be known.

- We validate our theoretical results using numerical simulations that show the necessity of our analytic techniques.

Related Work

While our model framework and the sparsity problem as a whole have been studied extensively (Fan and Lv 2006; Lafferty and Wasserman 2008; Wainwright 2009a,b), none have replicated the well known optimal lower bound for the classification problem under the given set of assumptions. Our work provides improved finite sample guarantees on DSTUMP for the regression problem that nearly match that of LASSO by means of weak learners.

Most closely related to our work is that of (Kazemitabar et al. 2017), which formulates the DSTUMP algorithm and theoretical approach for finite sample guarantees of these weak learners. Unlike our nearly tight result on the number of samples required for recovery, (Kazemitabar et al. 2017) provides a weaker $\mathcal{O}(s^2\log p)$ bound when using the *median splitting* criterion. We here demonstrate that the procedure can obtain the near optimal finite sample guarantees by highlighting a subtle nuance in the analysis of the stump splitting (potentially of independent interest to the reader); instead of using the variance of one sub-tree as an impurity measure, we use the variance of both sub-trees. As we will show both theoretically and experimentally, this consideration is vital for obtaining tight bounds. Our analysis is also more general than that of (Kazemitabar et al. 2017), with applications to both median and optimal splits, a wider class of functions f_k, and more general distributions.

In a recent work, (Klusowski and Tian 2021) provide an indirect analysis of the DSTUMP formulation with the *optimal splitting* criterion, by studying its relation to the SIS algorithm of (Fan and Lv 2006). Designed for linear models specifically, SIS sorts the features based on their Pearson correlation with the output, and has optimal sample complexity for the linear setting. (Klusowski and Tian 2021) show that when using the optimal splitting criterion, DSTUMP is *equivalent* to SIS up to logarithmic factors, which leads to a sample complexity of

$$n \gtrsim s\log(p)\cdot(\log(s)+\log(\log(p)))\,.$$

This improves the results of (Kazemitabar et al. 2017), though the analysis is more involved. A similar technique is also used to study the non-linear case, but the conditions assumed on f_k are hard to interpret and the bounds are weakened. Indeed, instantiating the non-parametric results for the linear case implies a *sub-optimal* sample complexity rate of $\mathcal{O}(s^2\log p)$. In addition, (Klusowski and Tian 2021) describe a heuristic algorithm for the case of *unknown* $|S|$, though they fail to prove any guarantees on its performance.

In contrast, we provide a direct analysis of DSTUMP . This allows us to obtain optimal bounds for *both* the *median* and *optimal* split in the linear case, as well as improved and generalized bounds in the non-linear case. Our novel approach further allows us to analyze the case of unknown sparsity level, as we provide the first formal proof for the heuristic algorithm suggested in (Klusowski and Tian 2021) and (Fan, Feng, and Song 2011). Compared to prior work, our analysis is considerably simpler and applies to more general settings.

Additionally, various studies have leveraged the simplicity of median splits in decision trees to produce sharp bounds

on mean-squared error for the regression problem with *random forests* (Duroux, Roxane and Scornet, Erwan 2018; Klusowski 2021). In each of these studies, analysis under the median split assumption allows for improvements in both asymptotic and finite sample bounds on prediction error for these ensemble weak learners. In the present work, we extend this intuition to the feature selection problem for high-dimensional sparse systems, and further emphasize the utility of the median split even in the singular decision stump case.

Algorithm Description

Notation and Problem Setup

For mathematical convenience, we adopt matrix notation and use $\mathbf{X} = (\mathbf{X}_{ij}) \in \mathbb{R}^{n\times p}$ and $Y = (Y_i) \in \mathbb{R}^n$ to denote the input matrix and the output vector respectively. We use $\mathbf{X}_{i,:}$ and X^k to refer the i-th row and k-th column of the matrix $\mathbf{X}$. We will also extend the definition of the univariate functions f_k to vectors by assuming that the function is applied to each coordinate separately: for $v \in \mathbb{R}^d$, we define $f_k(v) \in \mathbb{R}^d$ as $(f_k(v))_i = f_k(v_i)$.

We let $\mathbb{E}[\cdot]$ and $\mathrm{Var}(\cdot)$ denote the expectation and variance for random variables, with $\widehat{\mathbb{E}}[\cdot]$ and $\widehat{\mathrm{Var}}(\cdot)$ denoting their empirical counterparts i.e, for a generic vector $v \in \mathbb{R}^d$,

$$\widehat{\mathbb{E}}[v] = \frac{\sum_{i=1}^d v_i}{d} \quad \text{and} \quad \widehat{\mathrm{Var}}(v) = \frac{\sum_{i=1}^d (v_i - \widehat{\mathbb{E}}[v])^2}{d}.$$

We will also use $[d]$ to denote the set $\{1, \dots, d\}$. Finally, we let $\mathrm{Unif}(a, b)$ denote the uniform distribution over $[a, b]$ and use $C, c > 0$ to denote generic universal constants.

Throughout the paper, we will use the concept of *sub-Gaussian* random variables for stating and proving our results. A random variable $Z \in \mathbb{R}$ is called sub-Gaussian if there exists a t for which $\mathbb{E}\left[e^{(Z/t)^2}\right] \le 2$ and its *sub-Gaussian norm* is defined as

$$\|Z\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\left[e^{(Z/t)^2}\right] \le 2\right\}.$$

Sub-Gaussian random variables are well-studied and have many desirable properties, (see (Vershynin 2018) for a comprehensive overview), some of which we outline below as they are leveraged throughout our analysis.

(**P1**) (Hoeffding's Inequality) If Z is a sub-Gaussian random variable, then for any $t > 0$,

$$\Pr\left(|Z - \mathbb{E}[Z]| \ge t\right) \le 2e^{-c\left(t/\|Z\|_{\psi_2}\right)^2}.$$

(**P2**) If $Z_1, \dots Z_n$ are independent sub-Gaussian random variables, then $\sum Z_i$ is also sub-Gaussian with norm satisfying $\|\sum Z_i\|_{\psi_2}^2 \le \sum \|Z_i\|_{\psi_2}^2$.

(**P3**) If Z is a sub-Gaussian random variable, then so is $Z - \mathbb{E}[Z]$ and $\|Z - \mathbb{E}[Z]\|_{\psi_2} \le c\|Z\|_{\psi_2}$.

DSTUMP Algorithm

We here present the recovery algorithm DSTUMP. For each feature $k \in [p]$, the algorithms fits a single-level decision tree or "stump" on the given feature and defines the impurity of the feature as the error of this decision tree. Intuitively, the active features are expected to be better predictors of Y and therefore have *lower* impurity values. Thus, the algorithm sorts the features based on these values, and outputs the $|S|$ features with smallest impurity. A formal pseudo-code of our approach is given in Algorithm 1.

Algorithm 1: Scoring using DSTUMP

Input: $\mathbf{X} \in \mathbb{R}^{n\times p}, Y \in \mathbb{R}^n, s \in \mathbb{N}$
Output: Estimate of S

for $k \in \{1, ..., p\}$ **do**
  $\tau^k = \mathrm{argsort}(X^k)$
  for $n_L \in \{1, \dots, n\}$ **do**
    $n_R = n - n_L$
    $Y_L^k = (Y_{\tau^k(1)}, ...Y_{\tau^k(n_L)})^T$
    $Y_R^k = (Y_{\tau^k(n_L+1)}, ...Y_{\tau^k(n)})^T$
    $\mathrm{imp}_{k,n_L} = \frac{n_L}{n}\widehat{\mathrm{Var}}(Y_L^k) + \frac{n_R}{n}\widehat{\mathrm{Var}}(Y_R^k)$
  if *median split* **then**
    $\mathrm{imp}_k = \mathrm{imp}_{k,\lfloor \frac{n}{2} \rfloor}$.
  else
    $\mathrm{imp}_k = \min_\ell \mathrm{imp}_{k,\ell}$.
return $\tau = \mathrm{argsort}(\mathrm{imp})$

Formally, for each feature k, the k-th column is sorted in increasing order such that $X^k_{\tau^k(1)} \le X^k_{\tau^k(2)} \cdots \le X^k_{\tau^k(n)}$ with ties broken randomly. The samples are then split into two groups: the *left half*, consisting of $X_i^k \le X^k_{\tau^k(n_L)}$ and the *right half* consisting of $X_i^k > X^k_{\tau^k(n_L)}$. Conceptually, this is the same as splitting a single-depth tree on the k-th column with a n_L to n_R ratio and collecting the samples in each group. The algorithm then calculates the variance of the output in each group, which represents the optimal prediction error for this group with a single value as $\widehat{\mathrm{Var}}\left(Y_L^k\right) = \min_t \frac{1}{n_L} \cdot \sum (Y^k_{L,i} - t)^2$. The average of these two variances is taken as the impurity. Formally,

$$\mathrm{imp}_{k,n_L} = \frac{n_L}{n}\widehat{\mathrm{Var}}(Y_L^k) + \frac{n_R}{n}\widehat{\mathrm{Var}}(Y_R^k). \tag{1}$$

For the *median split* algorithm, the value of n_L is simply chosen as $\lfloor \frac{n}{2} \rfloor$ or $\lceil \frac{n}{2} \rceil$, where as for the optimal split, the value is chosen in order to minimize the impurity of the split. Lastly, the features are sorted by their impurity values and the $|S|$ features with lowest impurity are predicted as S. In the Unknown Sparsity Level section, we discuss the case of unknown $|S|$ and obtain an algorithm with nearly matching guarantees.

Main Results

Linear Design

For our first results, we consider the simplest setting of linear models with uniform distribution of the inputs. Formally, we assume that there is a vector $\beta_k \in \mathbb{R}^p$ such that $f_k(x) = \beta_k \cdot x$ for all k. This is equivalent to considering the *linear regression* model $Y = \sum_k \beta_k X^k + W$. We further assume that each entry of the matrix $\mathbf{X}$ is an i.i.d draw from the uniform distribution on $[0, 1]$. This basic setting is important

from a theoretical perspective as it allows us to compare with existing results from the literature before extending to more general contexts. This initial result of Theorem 0.1, provides an upper bound on the failure probability for DSTUMP in this setting.

Theorem 0.1. *Assume that each entry of the input matrix X is sampled i.i.d from a uniform distribution on $[0, 1]$. Assume further that the output vectors satisfy the* linear regression *model $Y = \sum_k \beta_k X^k + W$ where W_i are sampled i.i.d from $N(0, \sigma_w^2)$. Algorithm 1 correctly recovers the active feature set S with probability at least*

$$1 - 4se^{-c\cdot n} - 4pe^{-c\cdot n \frac{\min_k \beta_k^2}{\|\beta\|_2^2 + \sigma_w^2}}.$$

for the median split, and with probability at least

$$1 - 4se^{-c\cdot n} - 4npe^{-c\cdot n \frac{\min_k \beta_k^2}{\|\beta\|_2^2 + \sigma_w^2}}.$$

for the optimal split.

Moreover, the above theorem provides a finite sample guarantee for the DSTUMP algorithm and does not make *any* assumptions on the parameters or their asymptotic relationship. In order to obtain a comparison with existing literature, (Kazemitabar et al. 2017; Klusowski and Tian 2021; Wainwright 2009a,b), we consider these bounds in the asymptotic regime $\min_{k\in S} \beta_k^2 \in \Omega(\frac{1}{s})$ and $\|\beta\|_2 \in \mathcal{O}(1)$.

Corollary 0.2. *In the same setting as Theorem 0.1, assume that $\|\beta\|_2^2 \in \mathcal{O}(1)$ and $\min_{k\in S} \beta_k^2 \in \Omega(\frac{1}{s})$. Then Algorithm 1 correctly recovers the active feature set S with high probability as long as $n \gtrsim s \log p$.*

The proof of the Corollary is presented in Appendix. The above result shows that DSTUMP is optimal for recovery when the data obeys a linear model. This is interesting considering tree based methods are known for their strength in capturing non-linear relationships and are not designed with a linearity assumption like LASSO. In the next section, we further extend our finite sample bound analysis to *non-linear models*.

Additive Design

We here consider the case of non-linear f_k and obtain theoretical guarantees for the original DSTUMP algorithm. Our main result is Theorem 0.3 stated below.

Theorem 0.3. *Assume that each entry of the input matrix X is sampled i.i.d from a uniform distribution on $[0, 1]$ and $Y = \sum_k f_k(X^k) + W$ where W_i are sampled i.i.d from $\mathcal{N}(0, \sigma_w^2)$. Assume further that each f_k is monotone and $f_k(\text{Unif}(0, 1))$ is sub-Gaussian with sub-Gaussian norm $\|f_k(\text{Unif}(0, 1))\|_{\psi_2}^2 \le \sigma_k^2$. For $k \in S$, define g_k as*

$$g_k := \mathbb{E}\left[f_k(\text{Unif}(\frac{1}{2}, 1))\right] - \mathbb{E}\left[f_k(\text{Unif}(0, \frac{1}{2}))\right] \quad (2)$$

and define σ^2 as $\sigma^2 = \sigma_w^2 + \sum_k \sigma_k^2$. Algorithm 1 correctly recovers the set S with probability at least

$$1 - 4se^{-cn} - 4pe^{-cn\frac{\min_k g_k^2}{\sigma^2}}$$

for the median split and

$$1 - 4se^{-cn} - 4npe^{-cn\frac{\min_k g_k^2}{\sigma^2}}$$

for the optimal split.

Note that, by instantiating Theorem 0.3 for linear models, we obtain the same bound as in Theorem 0.1 implying the above bounds are tight in the linear setting.

The extension to all monotone functions in Theorem 0.1 has an important theoretical consequence: since the DSTUMP algorithm is invariant under monotone transformations of the input, we can obtain the same results for *any* distribution of $\mathbf{X}_{i,:}$. As a simple example, consider $\mathbf{X}_{ij} \sim \mathcal{N}(0, 1)$ and assume that we are interested in bounds for linear models. Define the matrix $\mathbf{Z}$ as $\mathbf{Z}_{ij} = F_{\mathcal{N}}(\mathbf{X}_{ij})$ where $F_{\mathcal{N}}(.)$ denotes the CDF of the Gaussian distribution. Since the CDF is an increasing function, running the DSTUMP algorithm with $(\mathbf{Z}, Y)$ produces the same output as running it with $(\mathbf{X}, Y)$. Furthermore, applying the CDF of a random variable to itself yields a uniform random variable. Therefore, $\mathbf{Z}_{ij}$ are i.i.d draws of the Unif$(0, 1)$ distribution. Setting $f_k(t) = \beta_k \cdot F_{\mathcal{N}}^{-1}(t)$, the results of Theorem 0.3 for $(\mathbf{Z}, Y)$ imply the same bound as Theorem 0.1. Notably, we can obtain the same sample complexity bound of $\mathcal{O}(s \log p)$ for the Gaussian distribution as well. In the appendix, we discuss a generalization of this idea, which effectively allows us to remove the uniform distribution condition in Theorem 0.3.

Unknown Sparsity Level

A drawback of the previous results are that they assume $|S|$ is given when, in general, this is not the case. Even if $|S|$ is not known however, Theorem 0.3 guarantees that the active features are ranked higher than non-active ones in τ, i.e, $\tau(k) < \tau(k')$ for all $k \in S, k' \notin S$. In order to recover S, it suffices to find a threshold γ such that $\max_{k\in S} \text{imp}_k \le \gamma \le \min_{k\notin S} \text{imp}_k$.

To solve this, we use the so called "permutation algorithm" which is a well known heuristic in the statistics literature (Barber and Candès 2015; Chung and Romano 2013, 2016; Fan, Feng, and Song 2011) and was discussed (without proof) by (Klusowski and Tian 2021) as well. Formally, we apply a random permutation σ on the rows of X, obtaining the matrix $\sigma(X)$ where $\sigma(X)_{ij} = X_{\sigma(i),j}$, We then rerun Algorithm 1 with $\sigma(X)$ and Y as input. The random permutation means that X and Y were "decoupled" from each other and effectively, *all* of the features are now inactive. We therefore expect $\min_{i,t} \text{imp}_i(\sigma(X), y)$ to be close to $\min_{k\notin S} \text{imp}_k(X, y)$. Since this estimate may be somewhat conservative, we repeat the sampling and take the minimum value across these repetitions. A formal pseudocode is provided in Algorithm 2. The STUMPSCORE method is the same algorithm as Algorithm 1, with the distinction that it returns imp in Line 12

Algorithm 2: Unknown $|S|$

Input: $\mathbf{X} \in \mathbb{R}^{n\times p}, Y \in \mathbb{R}^n$
Output: $\gamma \in [\max_{k\in S} \text{imp}_k, \min_{k\notin S} \text{imp}_k]$
1: **for** $t \leftarrow 1, \ldots, T-1$ **do**
2: $\quad \sigma^{(t)} \leftarrow$ Random permutation on $[n]$.
3: $\quad \text{imp}^{(t)} \leftarrow \textsc{StumpScore}(\sigma^{(t)}(X), y)$
return $\min_{i,t} \text{imp}_i^{(t)}$

Assuming we have used T repetitions in the algorithm, the probability that $\min_{k\notin S} \text{imp}_k(X, y) \le \gamma$ is at most $\frac{1}{T}$. While we provide a formal proof in the appendix, the main intuition behind the result is that $\text{imp}_i^{(t)}$ and $\text{imp}_{k'}$ for $k' \notin S$ are all the impurities corresponding to inactive features. Thus, the probability that the maximum across all of these impurities falls in $[p]\backslash S$ is at most $\frac{p-s}{Tp} \le \frac{1}{T}$. Ensuring that $\max_{k\in S} \text{imp}_k(X, y) \le \gamma$ involves treating the extra T impurity scores as $(T-1)p$ extra inactive features. This means that we can use the same results of Theorem 0.3 with p set to Tp since our sample complexity bound is *logarithmic* in p. The formal result follows with proof in the appendix.

Theorem 0.4. *In the same setting as Theorem 0.3, let γ be the output of Algorithm 2 and set $\widehat{S}$ to be $\{k : \text{imp}_k \le \gamma\}$. The probability that $\widehat{S} = S$ is at least*

$$1 - T^{-1} - 4se^{-cn} - 4Tpe^{-cn\frac{\min_k g_k^2}{\sigma^2}}$$

for the median split and at least

$$1 - T^{-1} - 4se^{-cn} - 4nTpe^{-cn\frac{\min_k g_k^2}{\sigma^2}}$$

for the optimal split.

We note that if we set $T = n^c$ for some constant $c > 0$, we obtain the same $\mathcal{O}(s \log p)$ as before.

Proofs

In this section, we prove Theorem 0.3 as it is the more general version of Theorem 0.1, and defer with remainder of the proofs to the appendix.

To prove that the algorithm succeeds, we need to show that $\text{imp}_k < \text{imp}_{k'}$ for all $k \in S, k' \notin S$. We proceed by first proving an upper bound imp_k for all $k \in S$

Lemma 0.5. *In the setting of Theorem 0.3, for any active feature $k \in S$, $\text{imp}_k \le \widehat{Var}(Y) - \frac{\min_k g_k^2}{720}$ with probability at least*

$$1 - 4e^{-c\cdot n} - 4e^{-c\cdot n\cdot\frac{\min_k g_k^2}{\sigma^2}}.$$

Subsequently, we need to prove an analogous lower bound on $\text{imp}_{k'}$ for all $k' \notin S$.

Lemma 0.6. *In the setting of Theorem 0.3, for any inactive feature $k' \notin S$, $\text{imp}_{k'} > \widehat{Var}(Y) - \frac{\min_k g_k^2}{720}$ with probability at least*

$$1 - 4e^{-c\cdot n\cdot\frac{\min_k g_k^2}{\sigma^2}}$$

for the median split and

$$1 - 4ne^{-c\cdot n\cdot\frac{\min_k g_k^2}{\sigma^2}}$$

for the optimal split.

Taking the union bound over all k, k', Lemmas 0.5 and 0.6 prove the theorem as they show that $\text{imp}_k < \text{imp}_{k'}$ for all $k \in S, k' \notin S$ with the desired probabilities.

We now focus on proving Lemma 0.5. We assume without loss of generality that f_k is increasing [1]. We further assume that $\mathbb{E}\left[f_k(X_i^k)\right] = 0$ as DSTUMP is invariant under constant shifts of the output. Finally, we assume that $n > 5$, as for $n \le 5$, the theorem's statement can be made vacuous by choosing large c.

We will assume $\{n_L, n_R\} = \{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil\}$ throughout the proof; as such our results will hold for both the optimal and median splitting criteria. As noted before, a key point for obtaining a tight bound is considering *both* sub-trees in the analysis instead of considering them individually. Formally, while the impurity is usually defined via variance as in (1), it has the following equivalent definition.

$$\text{imp}_k = \widehat{\text{Var}}(Y) - \frac{n_L \cdot n_R}{n^2} \cdot \left(\widehat{\mathbb{E}}\left[Y_L^k\right] - \widehat{\mathbb{E}}\left[Y_R^k\right]\right)^2. \quad (3)$$

The above identity is commonly used for the analysis of decision trees and their properties (Breiman et al. 1983; Li et al. 2019; Klusowski 2020; Klusowski and Tian 2021). From an analytic perspective, the key difference between (3) and (1) is that the empirical averaging is calculated *before* taking the square, allowing us to more simply analyze the *first* moment rather than the second.

Intuitively, we want to use concentration inequalities to show that $\widehat{\mathbb{E}}\left[Y_L^k\right]$ and $\widehat{\mathbb{E}}\left[Y_R^k\right]$ concentrate around their expectations and lower bound $|\mathbb{E}\left[Y_L^k\right] - \mathbb{E}\left[Y_R^k\right]|$. This is challenging however as concentration results typically require an i.i.d assumption but, as we will see, $Y_{L,i}^k$ are not i.i.d. More formally, for each k, define the random variable $X_L^k \in \mathbb{R}^{n_L}$ as $(X^k_{\tau^k(1)}, \ldots, X^k_{\tau^k(n_L)})^T$ and thus $Y_{L,i}^k = f_k(X_{L,i}^k) + \sum_{j\neq k} f_j(X^j_{\tau^k(i)}) + W_{\tau^k(i)}$. While the random vectors $X^{j\neq k}$ and W have i.i.d entries, X_L^k was obtained by sorting the coordinates of X^k. Thus, its coordinates are non-identically distributed and dependent. To solve the first problem, observe that the empirical mean is invariant under permutation and we can thus randomly shuffle the elements of Y_L^k in order to obtain a vector with identically distributed coordinates. Furthermore, by De Finetti's Theorem, any random vector with coordinates that are identically distributed (but not necessarily independent), can be viewed as a *mixture* of i.i.d vectors, effectively solving the second problem. Formally, the following result holds.

Lemma 0.7 (Lemma 1 in (Kazemitabar et al. 2017))**.** *let $\tilde{\tau} : [n_L] \to [n_L]$ be a random permutation on $[n_L]$ independent of $(\mathbf{X}, W)$ and define $\widetilde{X}_L^k \in \mathbb{R}^{n_L}$ as $\widetilde{X}_{L,i}^k := X_{L,\tilde{\tau}(i)}^k$. The random vector $\widetilde{X}_L^k$ is distributed as a mixture of uniform i.i.d uniform vectors of size n_L on $[0, \Theta]$ with Θ sampled from Beta$(n_L + 1, n_R)$.*

Defining $\widetilde{Y}_L^k \in \mathbb{R}^{n_L}$ as $\widetilde{Y}_{L,i}^k := Y_{L,\tilde{\tau}(i)}^k$, it is clear that $\widehat{\mathbb{E}}\left[\widetilde{Y}_L^k\right] = \widehat{\mathbb{E}}\left[Y_L^k\right]$ and therefore we can analyze $\widetilde{Y}_L^k$ instead

[1] The case of decreasing f_k follows by a symmetric argument or by mapping $X^k \to -X^k$

of Y_L^k as

$$\widetilde{Y}_{L,i}^k := f_k(\widetilde{X}_{L,i}^k) + \sum_{j\neq k} f_j(X^j_{\tau^k\circ\tilde{\tau}(i)}) + W_{\tau^k\circ\tilde{\tau}(i)}$$

which, given Lemma 0.7, can be seen as a mixture of i.i.d random variables.

Lemma 0.7 shows that there are two sources of randomness in the distribution of $\widetilde{Y}_{L,i}^k$: the mixing variable Θ and the sub-Gaussian randomness of $\widetilde{X}_L^k|\Theta$ and $(X^{j\neq k}, W)$. For the second source, it is possible to use standard concentration inequalities to show that conditioned on Θ, $\widehat{\mathbb{E}}\left[\widetilde{Y}_L^k\right]$ concentrates around $\mathbb{E}\left[\widetilde{Y}_{L,1}^k|\Theta\right]$. We will formally do this in Lemma 0.9. Before we do this however, we focus on the first source and how Θ affects the distribution of $\widetilde{Y}_{L,i}^k$.

Since Θ is sampled from Beta(n_L+1, n_R), we can use standard techniques to show that it concentrates around $\frac{1}{2}$. More formally, we can use the following lemma, the proof of which is in the appendix.

Lemma 0.8. *If $n \geq 5$, we have $\Theta \in [\frac{1}{4}, \frac{3}{4}]$ with probability at least $1 - 2e^{-cn}$.*

Given the above result, we can analyze $\widetilde{Y}_L^k$ assuming $\Theta \in [1/4, 3/4]$. In this case, we can use concentration inequalities to show that with high probability, $\widehat{\mathbb{E}}\left[\widetilde{Y}_L^k\right]$ concentrates around $\mathbb{E}\left[f_k(\text{Unif}(0,\Theta)]\right)$. Since f_k was assumed to be increasing, this can be further bounded by $\mathbb{E}\left[f_k(\text{Unif}(0,\frac{3}{4})\right]$. Formally, we obtain the following result.

Lemma 0.9. *Let $k \in S$ be an active feature. For any $\theta \in [\frac{1}{4}, \frac{3}{4}]$,*

$$\Pr\left[\widehat{\mathbb{E}}\left[\widetilde{Y}_L^k\right] - \mathbb{E}\left[\widetilde{Y}_{L,1}^k|\Theta=\theta\right] \geq t|\Theta=\theta\right] \leq 2e^{-cn\frac{t^2}{\sigma^2}}.$$

Furthermore, letting $\overline{f}_{a,b}^k$ denote $\mathbb{E}\left[f_k(\text{Unif}(a,b))\right]$,

$$\Pr\left[\widehat{\mathbb{E}}\left[Y_L^k\right] \geq \overline{f}_{0,\frac{3}{4}}^k + g_k/8\right] \leq 2e^{-c\cdot n} + 2e^{-c\cdot n\cdot\frac{g_k^2}{\sigma^2}}. \quad (4)$$

Proof. For ease of notation, we will fix $\theta \in [\frac{1}{4}, \frac{3}{4}]$ and let the random variables $\widehat{X}_L^k$ and $\widehat{Y}_L^k$ denote $\widetilde{X}_L^k|\Theta=\theta$ and $\widetilde{Y}_L^k|\Theta=\theta$ respectively. Recall that for all j, $f_k(X_i^j)$ was sub-Gaussian with parameter σ_j by assumption. It is straightforward to show (see the Appendix) that this means $f_k(\widehat{X}_{L,i}^k) - \mathbb{E}\left[f(\widehat{X}_{L,i}^k)\right]$ is also sub-Gaussian with norm at most $C\cdot\sigma_j^2$. Thus,

$$\begin{aligned}
\left\|\widehat{Y}_{L,i}^k\right\|_{\psi_2} &= \left\|f_k(\widehat{X}_{L,i}^k) + \sum_{j\neq k} f_j(X^j_{\tau^k\circ\tilde{\tau}(i)}) + W_{\tau^k\circ\tilde{\tau}(i)}\right\|_{\psi_2}^2 \\
&\overset{(i)}{=} \left\|f_k(\widehat{X}_{L,i}^k) + \sum_{j\neq k} f_j(X_i^j) + W_i\right\|_{\psi_2}^2 \\
&\overset{(ii)}{\leq} \left\|f_k(\widehat{X}_{L,i}^k)\right\|_{\psi_2}^2 + \sum_{j\neq k}\left\|f_j(X_i^j)\right\|_{\psi_2}^2 + \|W_i\|_{\psi_2}^2 \\
&\leq C\cdot\sigma_k^2 + \sum_{j\neq k}\sigma_j^2 + \sigma_w^2 \leq C\cdot\sigma^2.
\end{aligned}$$

In the above analysis, (i) follows from the independence assumption of $(\widehat{X}_L^k, X^{j\neq k}, W)$ together with the i.i.d assumption on $(X_i^{j\neq k}, W_i)$. As for (ii), it follows from (**P2**) together with the independence assumption of $(\widehat{X}_L^k, X^{j\neq k}, W)$. Property (**P3**) further implies that $\left\|\widehat{Y}_{L,i}^k - \mathbb{E}\left[\widehat{Y}_{L,i}^k\right]\right\|_{\psi_2}^2$ is upper bounded by $C\cdot\sigma^2$, proving the first Equation in the Lemma.

Now, using Hoeffding's inequality, we obtain

$$\Pr\left[\widehat{\mathbb{E}}\left[\widehat{Y}_L^k\right] - \mathbb{E}\left[\widehat{Y}_{L,i}^k\right] \geq g_k/8\right] \leq 2e^{-c\cdot n\cdot\frac{g_k^2}{\sigma^2}}.$$

Using Lemma 0.8 with $\Pr(A) \leq \Pr(B) + \Pr(A|B^C)$ for any two events A, B, we obtain

$$\Pr\left[\widehat{\mathbb{E}}\left[Y_L^k\right] - \mathbb{E}\left[\widehat{Y}_{L,i}^k\right] \geq g_k/8\right] \leq 2e^{-c\dot{n}} + 2e^{-c\cdot n\cdot\frac{g_k^2}{\sigma^2}}.$$

Note however that $\mathbb{E}\left[\widehat{Y}_{L,i}^k\right] = \overline{f}_{0,\theta}^k$ which as we show in the appendix, can further be upper bounded by $\overline{f}_{0,\frac{3}{4}}^k$, concluding the proof. □

Using the symmetry of the decision tree algorithm, we can further obtain that

$$\Pr\left[\widehat{\mathbb{E}}\left[Y_R^k\right] \geq \overline{f}_{\frac{1}{4},1}^k - g_k/8\right] \leq 2e^{-c\cdot n} + 2e^{-c\cdot n\cdot\frac{g_k^2}{\sigma^2}} \quad (5)$$

from (4) with the change of variable $X^k \to -X^k$ and $f_k = -f_k$. Taking union bound over (4) and (5), it follows that with probability at least $1 - 4e^{-c\cdot n} - 4e^{-c\cdot n\cdot\frac{g_k^2}{\sigma^2}}$,

$$\widehat{\mathbb{E}}\left[Y_R^k\right] - \widehat{\mathbb{E}}\left[Y_L^k\right] \geq \overline{f}_{\frac{1}{4},1}^k - \overline{f}_{0,\frac{3}{4}}^k - g_k/4.$$

As we show in the appendix however, a simple application of conditional expectations implies $\overline{f}_{\frac{1}{4},1}^k - \overline{f}_{0,\frac{3}{4}}^k \geq g_k/3$. Therefore, with probability at least $1 - 4e^{-c\cdot n} - 4e^{-c\cdot n\cdot\frac{g_k^2}{\sigma^2}}$, we have $\widehat{\mathbb{E}}\left[Y_R^k\right] - \widehat{\mathbb{E}}\left[Y_L^k\right] \geq \frac{g_k}{12}$. Assuming $n \geq 5$, we can further conclude that $\frac{n_L\cdot n_R}{n^2} \geq \frac{1}{5}$ which together with (3), proves the lemma.

Experimental Results

In this section, we provide further justification of our theoretical results in the form of simulations on the finite sample count for active feature recovery under different regimes, as well as the predictive power of a single sub-tree as compared to the full tree. We additionally contrast DSTUMP with the widely studied optimal LASSO .

We first validate the result of Theorem 0.1 and consider the linear design with $p = 200$ and design matrix entries sampled i.i.d. from $U(0,1)$ with additive Gaussian noise $\mathcal{N}(0,.1)$. Concretely, we examine the optimal number of samples required to recover approximately 95% of the active features s. This is achieved by conducting a binary search on the number of samples to find the minimal such value that recovers the desired fraction of the active feature set, averaged across 25 independent replications. In the leftmost plot of Figure 1, we plot the sample count as a function of varying

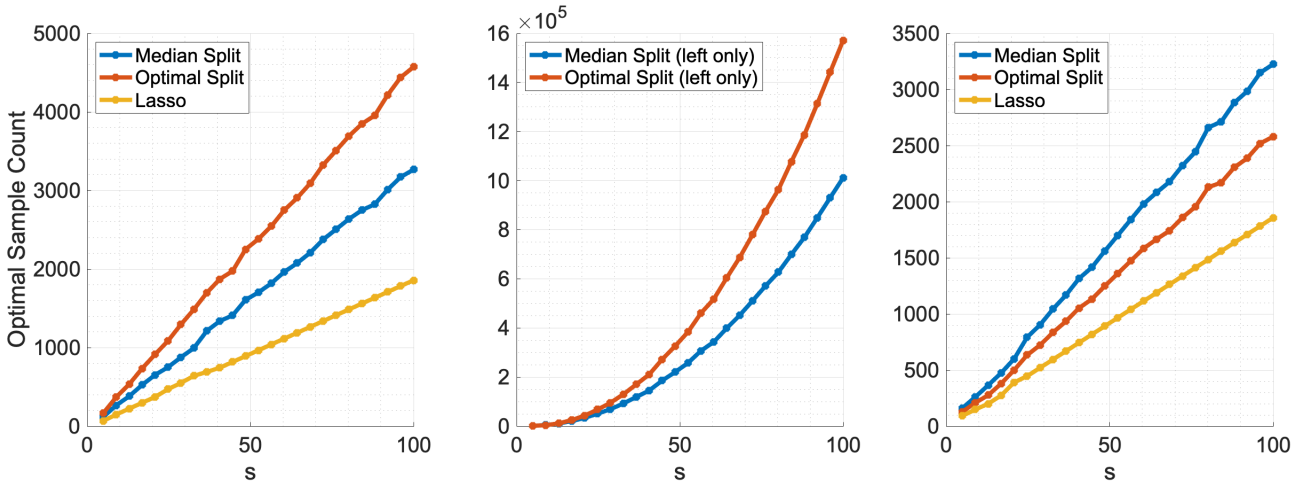

Figure 1: Optimal sample count to recover 95% of the active features where design matrix samples i.i.d from $U(-1,1)$ or $\mathcal{N}(0,1)$ with additive Gaussian noise $\mathcal{N}(0,0.1)$, comparing three methods: DSTUMP with optimal split, DSTUMP with median split, and LASSO .

sparsity levels $s \in [5, 100]$ for DSTUMP with a median split, DSTUMP with the optimal split, as well as LASSO for benchmarking (with penalty parameter selected through standard cross-validation). By fixing p, we are evaluating the dependence of n on the sparsity level s alone. The results here validate the theoretical $\mathcal{O}(s \log p)$ bound that nearly matches the optimal LASSO . Also of note, the number of samples required by the median splitting is less than that of the optimal. Thus, in the linear setting, we see that DSTUMP with median splitting is both more simplistic and computationally inexpensive. This optimal bound result is repeated with Gaussian data samples in the right most plot of Figure 1. Notably, in this setting the optimal split decision stumps perform better than the median as it demonstrates their varied utility under different problem contexts.

We additionally reiterate that the prior work of (Kazemitabar et al. 2017) attempted to simplify the analysis of the sparse recovery problem using DSTUMP by examining only the left sub-tree, which produced the non-optimal $\mathcal{O}(s^2 \log p)$ finite sample bound. To analyze the effect of this choice, the middle plot of Figure 1 presents the optimal sample recovery count when using only the left sub-tree subject to the additive model of Theorem 0.1. In accordance with our expectation and previous literature's analysis, we see a clear quadratic relationship between n and s when fixing the feature count p.

Overall, these simulations further validate the practicality and predictive power of decisions stumps. Benchmarked against the optimal LASSO , we see a slight decrease in performance but a computational reduction and analytic simplification.

Conclusion

In this paper, we presented a simple and consistent feature selection algorithm in the regression case with single-depth decision trees, and derived the finite-sample performance guarantees in a high-dimensional sparse system. Our theoretical results demonstrate that this very simple class of weak learners is nearly optimal compared to the gold standard LASSO . We have provided strong theoretical evidence for the success of binary decision tree based methods in practice and provided a framework on which to extend the analysis of these structures to arbitrary height, a potential direction for future work.

Acknowledgements

The work is partially support by DARPA QuICC, NSF AF:Small #2218678, and NSF AF:Small # 2114269. Max Springer was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1840340. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Barber, R. F.; and Candès, E. J. 2015. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5): 2055–2085.

Breiman, L. 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3): 199 – 231.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1983. Classification and Regression Trees.

Bureau, A.; Dupuis, J.; Falls, K.; Lunetta, K.; Hayward, B.; Keith, T.; and Eerdewegh, P. 2005. Identifying SNPs predictive of phenotype using random forests. *Genetic epidemiology*, 28: 171–82.

Chung, E.; and Romano, J. P. 2013. Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2): 484–507.

Chung, E.; and Romano, J. P. 2016. Multivariate and multiple permutation tests. *Journal of econometrics*, 193(1): 76–91.

Duroux, Roxane; and Scornet, Erwan. 2018. Impact of subsampling and tree depth on random forests. *ESAIM: PS*, 22: 96–128.

Fan, J.; Feng, Y.; and Song, R. 2011. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494): 544–557.

Fan, J.; and Lv, J. 2006. Sure independence screening for ultrahigh dimensional feature space. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 70: 849–911.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29: 1189–1232.

Han, C.; Rao, N. S.; Sorokina, D.; and Subbian, K. 2020. Scalable Feature Selection for (Multitask) Gradient Boosted Trees. *ArXiv*, abs/2109.01965.

Huynh-Thu, V. A.; Irrthum, A.; Wehenkel, L.; and Geurts, P. 2012. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLOS ONE*, 5(9): 1–10.

Kazemitabar, S. J.; Amini, A. A.; Bloniarz, A.; and Talwalkar, A. S. 2017. Variable Importance Using Decision Trees. In *NIPS*.

Klusowski, J. 2021. Sharp Analysis of a Simple Model for Random Forests. In Banerjee, A.; and Fukumizu, K., eds., *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 757–765. PMLR.

Klusowski, J. M. 2020. Sparse learning with CART. *ArXiv*, abs/2006.04266.

Klusowski, J. M.; and Tian, P. M. 2021. Nonparametric Variable Screening with Optimal Decision Stumps. In *AISTATS*.

Lafferty, J. D.; and Wasserman, L. A. 2008. Rodeo: Sparse, greedy nonparametric regression. *Annals of Statistics*, 36: 28–63.

Li, L.; and Wang, J. 2018. Research on feature importance evaluation of wireless signal recognition based on decision tree algorithm in cognitive computing. *Cognitive Systems Research*, 52: 882–890.

Li, X.; Wang, Y.; Basu, S.; Kumbier, K.; and Yu, B. 2019. A Debiased mDI Feature Importance Measure for Random Forests. *ArXiv*, abs/1906.10845.

Lunetta, K. L.; Hayward, L. B.; Segal, J.; and Eerdewegh, P. V. 2004. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics*, 5(1): 32.

Qian, H.; Wang, B.; Yuan, M.; Gao, S.; and Song, Y. 2022. Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree. *Expert Systems with Applications*, 190: 116202.

Vershynin, R. 2018. High-Dimensional Probability: An Introduction with Applications in Data Science.

Wainwright, M. J. 2009a. Information-Theoretic Limits on Sparsity Recovery in the High-Dimensional and Noisy Setting. *IEEE Transactions on Information Theory*, 55: 5728–5741.

Wainwright, M. J. 2009b. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*.

Xia, Y.; Liu, C.; Li, Y.; and Liu, N. 2017. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, 78: 225–241.

Xu, Z.; Huang, G.; Weinberger, K. Q.; and Zheng, A. X. 2014. Gradient boosted feature selection. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 522–531.