Solving Heterogeneous Estimating Equations with Gradient Forests

Susan Athey
Stanford University
athey@stanford.edu

Julie Tibshirani
Palantir Technologies
jtibs@cs.stanford.edu

Stefan Wager
Columbia University / Stanford University
swager@stanford.edu

Draft version October 2016

Abstract

Forest-based methods are being used in an increasing variety of statistical tasks, including causal inference, survival analysis, and quantile regression. Extending forest-based methods to these new statistical settings requires specifying tree-growing algorithms that are targeted to the task at hand, and the ad-hoc design of such algorithms can require considerable effort. In this paper, we develop a unified framework for the design of fast tree-growing procedures for tasks that can be characterized by heterogeneous estimating equations. The resulting *gradient forest* consists of trees grown by recursively applying a pre-processing step where we label each observation with gradient-based pseudo-outcomes, followed by a regression step that runs a standard CART regression split on these pseudo-outcomes. We apply our framework to two important statistical problems, non-parametric quantile regression and heterogeneous treatment effect estimation via instrumental variables, and we show that the resulting procedures considerably outperform baseline forests whose splitting rules do not take into account the statistical question at hand. Finally, we prove the consistency of gradient forests, and establish a central limit theorem. Our method will be available as an `R`-package, `gradientForest`, which draws from the `ranger` package for random forests.

1 Introduction

As the amount of data available to us increases, there is a growing interest in personalizing statistical analyses: In medicine, we seek the most appropriate treatment for each patient, while in economics we may want to model heterogeneous preferences across different agents. And, in several application areas, random forests [Breiman, 2001] have shown considerable promise as a tool for such personalized analyses. For example, forest-based algorithms have been found to perform well for hetereogeneous treatment effect estimation [Green and Kern,

We are grateful for helpful comments from several colleagues; in particular, we are indebted to Jerry Friedman for first suggesting we take a closer look at splitting rules for quantile regression forests, and to Will Fithian for drawing our attention to connections between our early ideas and gradient boosting.

2012, Hill, 2011, Wager and Athey, 2015, Taddy et al., 2016], person-specific survival analysis [Ishwaran et al., 2008, Ishwaran and Kogalur, 2010], and personalized treatment allocation [Kallus, 2016].

At a high level, forest-based methods for personalized statistical analysis operate in two steps: First, they use an ensemble of recursive partitioning trees [Breiman et al., 1984] to obtain an adaptive nearest-neighbor function, and then they apply the desired statistical analysis on these adaptive neighborhoods. For example, in order to estimate the effect of a medicine on a specific person of interest, forest-based methods would first construct an ensemble of trees to find people who ought to respond similarly to the treatment as our person of interest, and then for each tree in the forest, compute an average treatment effect on this personalized comparison group (that is, within that person's leaf in the tree).[1]

The success of any such forest-based analysis hinges on whether the adaptive neighborhood function obtained via partitioning adequately captures the heterogeneity in the underlying property we want to estimate. Even within the same class of statistical tasks, different types of questions can require different neighborhood functions. As a concrete example, suppose that two scientists are studying the effects of a new medical treatment: One wants to know how the treatment affects long-term survival, whereas the other is examining its effect on the length of hospital stays. It is entirely plausible that the neighborhood functions that are helpful in capturing the treatment heterogeneity in each setting would be based on completely different covariates, e.g., a patient's smoking habits for long-term survival, and the location and size of the hospital for the length of stay.

Thus, each time we apply random forests to a new scientific task, it is important to use tree-growing rules for recursive partitioning that are able to detect and highlight heterogeneity in the signal that researcher is interested in. Until now, such problem-specific rules have largely been designed by hand, a labor-intensive task. Although the CART rules of Breiman et al. [1984] have long been popular for classification and regression tasks, there has been a steady stream of papers proposing new splitting rules for other problems, including Athey and Imbens [2016] and Su et al. [2009] for treatment effect estimation, Beygelzimer and Langford [2009] and Kallus [2016] for personalized policy allocation, and Ciampi et al. [1986], Gordon and Olshen [1985], LeBlanc and Crowley [1992], Molinaro et al. [2004] as well as several others for survival analysis (see Bou-Hamad et al. [2011] for a review). Zeileis et al. [2008] proposes a method for constructing a single tree for general maximum likelihood problems, where the splitting rule is based on model goodness of fit.

Another challenge inherent in taking random forests to new scientific tasks is that the primary approach used in the literature to date is to estimate a model separtely in each leaf of the tree. For regression trees, estimating the model is simply estimating the sample mean outcome within the leaf; in Athey and Imbens [2016], this involves taking the difference of sample means for treated and control units, possibly weighted by the propensity score for treatment; in Zeileis et al. [2008], a more complex model might be estimated there. This approach presents a number of challenges if applied generally. First, the trees may need to be shallow (i.e., with large leaves) in order to estimate a more complex model reliably. Second, the subsample in each leaf may need to satisfy additional properties; for example, we may need to have sufficient numbers of both treated and control units for treatment effect estimation, and for instrumental variable estimation, more complex conditions must

[1] The causal forest algorithm of Wager and Athey [2015] proceeds in this way, using the causal trees of Athey and Imbens [2016] for recursive partitioning. The methods of Green and Kern [2012], Hill [2011], and Taddy et al. [2016] rely on more elaborate Bayesian heuristics for forest-based estimation, following Chipman et al. [2010].

be satisfied for estimation to be reliable and stable. However, both standard statistical practice following Breiman [2001] and existing results for statistical inference with random forests [Wager and Athey, 2015] rely on building deep trees with small leaves; forests based on shallow trees can yield estimates that are bias-dominated.

In this paper, we develop a method for forest-based estimation that deals with both challenges described here. First, we provide a general method for problem-specific splitting rules, one that is optimized for the primary objective of analyzing heterogeneity in a key parameter of interest. In the spirit of gradient boosting [Friedman, 2001], our recursive partitioning method begins by computing a linear, gradient-based approximation to the non-linear estimating equation we are trying to solve, and then uses this approximation to specify the tree-split point. Algorithmically, our procedure reduces to iteratively applying a labeling step where we generate pseudo-outcomes by computing gradients, and a regression step where we pass this labeled data to a standard CART regression routine. Thus, our approach lets us obtain high quality neighborhood functions while only using comparable computational resources to those required by the original classification or regression forests of Breiman [2001].

Second, at estimation time, we take a different approach than the majority of the existing literature on forest-based methods, and do not average parameter estimates obtained from different trees. Instead, we view the forest as generating weights for local generalized method of moments (or maximum likelihood) estimation. The forest is used to determine the relevance of each training sample for estimation at a specific point in features space: a given sample is weighted in proportion to the fraction of trees in which it is in the same leaf as the test point of interest. Thus, we can view the forest-based weighting function as an alternative to kernel-based weighting functions that have been proposed for local maximum likelihood or similar methods, as discussed in further detail below.

Our perspective enables us to immediately extend the set of problems for which we have forest-based algorithms. In the context of quantile regression, Meinshausen [2006] shows how to build consistent forest-based estimators, and in fact uses a similar kernel-based idea as us at estimation time. However, the splitting rule used in these quantile regression forests is based on a standard CART regression tree routine. For this reason the resulting method is not sensitive to quantile shifts that do not correspond to changes in the conditional mean function. Here, we show that the ability of gradient forests to specifically target changes in the conditional quantile function lets them outperform the baseline forests of Meinshausen [2006] in examples where the mean and quantile functions behave in divergent ways.

On the topic of heterogeneous treatment effect estimation, Wager and Athey [2015] studies the application and statistical behavior of random forests. The paper, however, only considers the case where the treatment assignment is effectively random conditional on the features, and so the proposed method cannot be used in economic applications where treatment assignment is endogenous. Gradient forests enable us to extend their results, making use of instrumental variables to identify causal effects. We emphasize that the methods advocated here are by no means a direct generalization of those studied by Wager and Athey [2015], and rather draw heavily from our gradient forest framework.

Finally, the most computationally intensive part of growing a CART-style tree is scanning over candidate covariates to find a good splitting point. Our algorithm naturally decomposes into an inexpensive label step, where we capture the structure of the specific statistical question, and a standard regression step. For this reason, we can make use of pre-existing, highly-optimized tree software to execute the regression step. In line with this approach, our package, `gradientForest` for `R` and `C++`, re-uses the regression splitting procedures from

the `ranger` implementation of random forests [Wright and Ziegler, 2015].

1.1 Related Work

The idea of local maximum likelihood estimation has a long history in statistics, with notable contributions due to Fan et al. [1998], Newey [1994], Staniswalis [1989], Stone [1977], Tibshirani and Hastie [1987] and others. In the economics literature, a popular application of these techniques has been to multinomial choice in a panel data setting [e.g., Honoré and Kyriazidou, 2000]. The basic idea is that when estimating parameters at a particular value of covariates, a kernel weighting function is used to place more weight on nearby observations in the covariate space. A challenge facing this approach is that if the covariate space has more than two or three dimensions, the "curse of dimensionality" implies that plain kernel-based methods may not perform well [e.g., Robins and Ritov, 1997].

Our paper takes the approach of replacing the kernel weighting with "forest-based" weights, that is, weights derived from the fraction of trees in which an observation appears in the same leaf as the target value of the covariate vector. The original random forest algorithm for non-parametric classification and regression was proposed by [Breiman, 2001], building on insights from Amit and Geman [1997] and Breiman [1996]. The perspective we take on random forests as a form of adaptive nearest neighbor estimation, however, most closely builds on the proposal of Meinshausen [2006] for forest-based quantile regression. This adaptive nearest neighbors perspective also underlies several statistical analyses of random forests, including those of Arlot and Genuer [2014], Biau and Devroye [2010], and Lin and Jeon [2006].

Meanwhile, our gradient-based splitting scheme draws heavily from a long tradition in the statistics and econometrics literatures of using gradient-based test statistics to detect change points in likelihood models [Andrews, 1993, Hansen, 1992, Hjort and Koning, 2002, Nyblom, 1989, Ploberger and Krämer, 1992, Zeileis, 2005, Zeileis and Hornik, 2007]. In particular, Zeileis et al. [2008] consider the use of such methods for model-based recursive partitioning. Our problem setting differs from the above in that we are not focused on running a hypothesis test, but rather seek an adaptive nearest neighbor weighting that is as sensitive as possible to heterogeneity in our parameter of interest; we then rely on the random forest resampling mechanism to achieve statistical stability [Mentch and Hooker, 2016, Scornet et al., 2015, Wager and Athey, 2015]. In this sense, our approach is closely related to the gradient boosting algorithm of Friedman [2001], who uses similar gradient-based approximations to guide a greedy, heuristic, non-parametric regression procedure.

Our asymptotic theory relates to an extensive recent literature on the statistics of random forests, most of which focuses on the regression case [Arlot and Genuer, 2014, Biau, 2012, Biau et al., 2008, Biau and Scornet, 2016, Breiman, 2004, Bühlmann and Yu, 2002, Chipman et al., 2010, Denil et al., 2014, Efron, 2014, Geurts et al., 2006, Ishwaran and Kogalur, 2010, Lin and Jeon, 2006, Meinshausen, 2006, Mentch and Hooker, 2016, Samworth, 2012, Scornet et al., 2015, Sexton and Laake, 2009, Wager et al., 2014, Wager and Athey, 2015, Wager and Walther, 2015]. Our present paper complements this body work, by showing how methods developed to study regression forests can also be used understand solutions to heterogeneous estimating equations obtained via gradient forests.

Finally, we note that the problem we study, namely estimating how a parameter vector varies with covariates, where this relationship is non-parametric, is distinct from the problem of estimating a single, low-dimensional parameter—such as an average treatment effect—while controlling for a non-parametric or high-dimensional set of covariates [e.g., Athey

et al., 2016, Belloni et al., 2013, Robins et al., 1995, van der Laan and Rubin, 2006].

1.2 Outline

We begin by formalizing the problem settings in the language of heterogeneous estimating equations, then present an abstract version of our tree-growing scheme that can be used to solve any such estimating equation. The second part of the paper is devoted to discussing and evaluating the application of gradient forests to concrete problems of statistical and economic interest: Section 4 considers quantile regression while Section 5 studies causal inference with instrumental variables. Finally, in Section 6, we undertake a theoretical analysis of gradient forests, and prove consistency and asymptotic normality results.

2 Heterogeneous Estimating Equations

In the interest of generality, we frame our presentation in terms of the formalism of (generalized) estimating equations. Suppose that we have n independent and identically distributed subjects, and for each subject $i = 1, ..., n$. For each sample, we have access to an observable quantity O_i that encodes the information about the subject we are interested in, along with a set of auxiliary covariates X_i. In the case of non-parametric regression, this observable just consists of an outcomes $O_i = \{Y_i\}$ with $Y_i \in \mathbb{R}$; in general, however, it will contain richer information. For example, in the case of treatment effect estimation with exogenous treatment assignment, $O_i = \{Y_i, W_i\}$ also includes the treatment assignment W_i; while in the case of treatment effect estimation with instrumental variables, $O_i = \{Y_i, W_i, Z_i\}$, where W_i is the (endogeneous) treatment assignment and Z_i is an instrument used to identify causal effects (Z_i is correlated with W_i but not with potential outcomes [Imbens and Angrist, 1994]).

Given this kind of data, our goal is to solve an estimation equation of the form

$$\mathbb{E}\left[\psi_{\theta(x),\,\nu(x)}\left(O_i\right) \,\middle|\, X_i = x\right] = 0 \quad \text{for all} \quad x \in \mathcal{X}, \tag{1}$$

where $\theta(x)$ is the parameter we care about and $\nu(x)$ is an optional nuisance parameter. This setting is very general, and encompasses several important problems in statistics and econometrics. At the end of this section, we outline how our some important statistical tasks fit into this framework, while emphasizing that the setting of heterogeneous estimating equations covers many more cases than we can review here.

A popular approach to estimating $\theta(x)$ in a heterogeneous estimating equation is to first define similarity weights $\alpha_i(x)$ that measure the relevance of the i-th training example to fitting $\theta(\cdot)$ at x, and then fit the target of interest as the solution to the moment equation

$$\left(\hat{\theta}(x),\, \hat{\nu}(x)\right) \in \text{argmin}_{\theta,\,\nu} \left\{ \left| \sum_{i=1}^{n} \alpha_i(x)\, \psi_{\theta,\,\nu}\left(O_i\right) \right| \right\}. \tag{2}$$

When the above expression has a unique root, we can simply say that $\hat{\theta}(x)$, $\hat{\nu}(x)$ solves $\sum_{i=1}^{n} \alpha_i(x)\, \psi_{\hat{\theta}(x),\,\hat{\nu}(x)}\left(O_i\right) = 0$.

The weights $\alpha_i(x)$ used to specify the above solution to the heterogeneous estimating equation are often obtained via a deterministic kernel function, perhaps with an adaptively chosen bandwith parameter [Fan et al., 1998, Newey, 1994, Staniswalis, 1989, Stone, 1977, Tibshirani and Hastie, 1987]. Although methods of the above kind often work well in low

dimensions, they can be very sensitive to the curse of dimensionality. In this paper, our goal is to use forest-based algorithms to adaptively learn better, problem-specific, weights $\alpha_i(x)$ that can be used in conjunction with (2).

Before presenting our method, we first briefly review how our concrete problems of interest fit in the setting discussed above.

Least-square regression Classical regression forests, as introduced by Breiman [2001], can be understood as estimators for the conditional mean response function $\theta(x) = \mathbb{E}\left[Y \,\middle|\, X = x\right]$. This statistical objective can be encoded within the framework of (1) by using the moment function $\psi_{\theta(x)}(Y_i) = Y_i - \theta(x)$.

Quantile regression Another ubiquitous statistical task is that of quantile estimation, whereby we seek to recover quantiles of the conditional distribution of Y_i given $X_i = x$, i.e., $\theta(x) = F_x^{-1}(q)$ for some pre-specified $q \in (0, 1)$. This problem can again be cast into the language of estimating equations, by using the moment function $\psi_{\theta(x)}(Y_i) = q\mathbf{1}\left(\{Y_i > q\}\right) - (1-q)\mathbf{1}\left(\{Y_i \leq q\}\right)$; see Koenker [2005] for a review.

Instrumental variables regression Suppose we want to measure the causal effect of a treatment assignment W_i on an outcome Y_i, but cannot exclude the possibility of non-causal correlations between W_i and Y_i. For example, we may want to measure the effect of college education (W_i) on income (Y_i), but recognize that individuals who went to college may have had higher earning potential than those who didn't even if they hadn't gone to college. As discussed in, e.g., Angrist et al. [1996], a popular strategy in such situations is to rely on an instrument Z_i that is associated with the treatment W_i but is immune to spurious correlations: In the above example, a randomized incentive to attend college could act as such an instrument.

Formally, in the language of potential outcomes [Neyman, 1923, Rubin, 1974], let $Y_i(w)$ denote the counterfactual outcome we would have observed for the i-subject had they received a treatment w. The instrument Z_i then allows for identification of heterogeneous causal effects whenever the instrument is associated with treatment assignment, and the following unconfoundedness condition holds [Rosenbaum and Rubin, 1983]:

$$\left\{Y_i(w)\right\}_{w \in \mathcal{W}} \perp\!\!\!\perp Z_i \,\middle|\, X_i. \tag{3}$$

Given this setup, the causal effect $\tau(X_i)$ of W_i on Y_i is identified as an estimating equation (1) using the moment function [e.g., Angrist and Pischke, 2008]

$$\psi_{\tau(x),\,\mu(x)} = \begin{pmatrix} Z_i\left(Y_i - W_i\tau(x) - \mu(x)\right) \\ Y_i - W_i\tau(x) - \mu(x) \end{pmatrix}, \tag{4}$$

where the intercept $\mu(x)$ is a nuisance parameter.

As a side, we note that there has been considerable recent interest in using random forests for heterogeneous treatment effect estimation [Green and Kern, 2012, Hill, 2011, Wager and Athey, 2015]. However, all of the previous literature has focused on the simpler case where the treatment assignment is exogneous, i.e., where the unconfoundedness condition (3) still holds if we replace Z_i with W_i, and so we do not need to use an auxiliary instrument to identify causal effects.

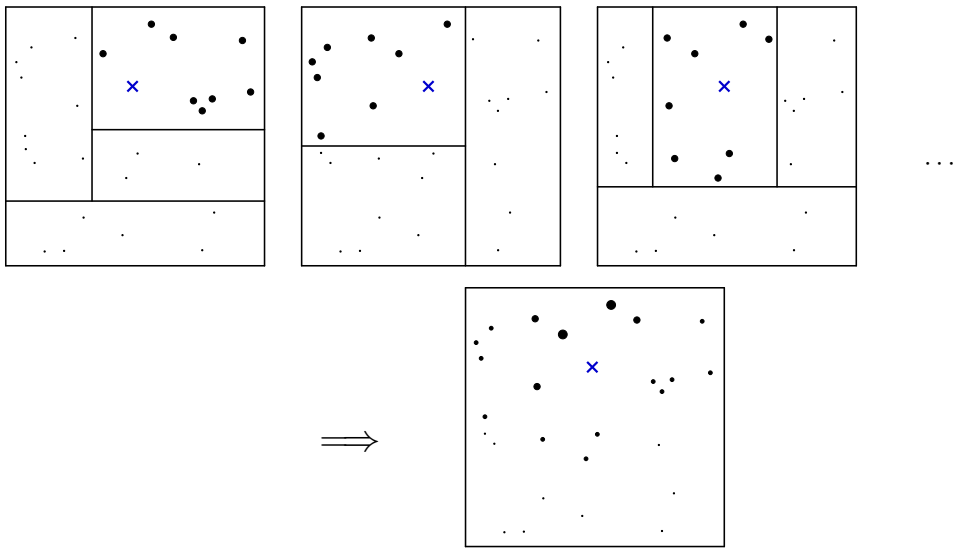

Figure 1: Illustration of the random forest weighting function. Each tree starts by giving equal (positive) weight to the training examples in the same leaf as our test point x of interest, and zero weight to all the other training examples. Then, the forest averages all these tree-based weightings, and effectively measures how often each training example falls into the same leaf as x.

3 Solving Estimating Equations with Forests

3.1 Forests as Weighted Neighborhood Estimation

We now return to our main topic of interest, that is, personalized estimation via random forests in generic statistical problems characterized by estimating equations. Following Meinshausen [2006], we view random forests a method producing a data-adaptive weighting function $\alpha_i(x)$ that quantifies the importance of the i-th training example for understanding a test point x; we then estimate $\theta(x)$ using (2) with the forest weights.

A random forest produces these weights $\alpha_i(x)$ by building an ensemble of B trees indexed by $b = 1, ..., B$, and for each such tree defining $L_b(x)$ as the set of training examples falling in the same "leaf" as x. Then, we define weights $\alpha_i(x)$ that capture the frequency with which the i-th training example falls into the same leaf as x:

$$\alpha_{bi}(x) = \frac{\mathbf{1}\left(\{X_i \in L_b(x)\}\right)}{|L_b(x)|}, \quad \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{bi}(x). \tag{5}$$

These weights sum to 1, and define the forest-based adaptive neighborhood of x; this weighting function is illustrated in Figure 1. Then, following (2), the random forest solves the heterogeneous estimating equation as $\sum_{i=1}^{n} \frac{1}{B}\sum_{b=1}^{B} \alpha_{bi}(x)\, \psi_{\hat{\theta}(x),\, \hat{\nu}(x)}\left(O_i\right) = 0$.

We note that, in the case of linear regression, this weighting-based definition of random forests is equivalent to the standard "average of trees" perspective taken in Breiman

[2001]. Specifically, suppose we want to estimate the conditional mean function $\theta(x) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x\right]$ which, as discussed above, is identified in (1) using the moment function $\psi_{\theta(x)}(Y_i) = Y_i - \theta(x)$. Then, we can use simple algebra to verify that

$$\sum_{i=1}^{n} \frac{1}{B} \sum_{b=1}^{B} \alpha_{bi}(x) \left(Y_i - \hat{\theta}(x)\right) = 0 \quad \Longleftrightarrow \quad \hat{\theta}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b(x), \tag{6}$$

where $\hat{\theta}_b(x) = \sum_{\{i: X_i \in L_b(x)\}} Y_i \,/\, |L_b(x)|$ is the prediction made by a single CART regression tree.

3.2 Splitting to Maximize Heterogeneity

Given this setup, our goal is to build trees that, when combined into a forest, induce weights $\alpha_i(x)$ that lead to good estimates of $\theta(x)$. In our search for good splits, we proceed greedily, i.e., our goal is that each split immediately improves the quality of the tree fit as much as possible. Every split starts with a parent node $P \in \mathcal{X}$; this parent node is characterized by a solution $\hat{\theta}_P$ to the estimating equation, namely

$$\sum_{\{i: X_i \in P\}} \psi_{\hat{\theta}_P}(O_i) = 0, \tag{7}$$

along with a mean-squared error

$$\text{err}(P) = \mathbb{E}\left[\left(\hat{\theta}_P - \theta(X)\right)^2 \,\middle|\, X \in P\right]. \tag{8}$$

We would like to divide P into two children C_1, $C_2 \in \mathcal{X}$ using an axis-aligned cut such as to improve the accuracy of our θ-estimates as much as possible or, in other words, to minimize the resulting squared error

$$\text{err}(C_1, C_2) = \sum_{j=1,2} \mathbb{P}\left[X \in C_j \,\middle|\, X \in P\right] \mathbb{E}\left[\left(\hat{\theta}_{C_j} - \theta(X)\right)^2 \,\middle|\, X \in C_j\right]. \tag{9}$$

In the least-squares regression case, we can simply use the prediction error of the tree over the two leaves C_1 and C_2 to get nearly-unbiased estimates of the error criterion (9), up to a shift parameter that does not depend on the split under consideration. Many standard regression tree implementations, such as CART, choose their splits by just minimizing the prediction error of the tree.

In our setting, however, this kind of direct loss minimization is not an option: If $\theta(x)$ is only identified through a moment condition, then we do not in general have access to unbiased, model-free estimates of the criterion (9). To side-step this issue, we rely on the following more abstract characterization of our target criterion.

Proposition 1. *Suppose that basic assumptions detailed in Section 6 hold, and that the parent node P has a radius smaller than r for some value $r > 0$. We write $n_P = |\{i : X_i \in P\}|$ for the number of observations in the parent, and suppose that $n_P \gg r^{-2}$. Define*

$$\Delta(C_1, C_2) := n_{C_1} n_{C_2} \,/\, n_P^2 \left(\hat{\theta}_{C_1} - \hat{\theta}_{C_2}\right)^2, \tag{10}$$

where $\hat{\theta}_{C_1}$ and $\hat{\theta}_{C_2}$ are solutions to the estimating equation computed in the children, following (7). Then, for any split of the parent node P into two children C_1 and C_2, we have

$$\text{err}\left(C_1,\, C_2\right) = K(P) - \Delta(C_1,\, C_2) + o_P\left(r^2\right), \tag{11}$$

where $K(P)$ is a deterministic term that measures the purity of the parent node that does not depend on how the parent is split.

Motivated by this observation, we consider splits that make the above Δ-criterion (10) large. A special case of the above idea also underlies the splitting rule for treatment effect estimation proposed by Athey and Imbens [2016]. At a heuristic level, we can think of this Δ-criterion as favoring splits that increase the heterogeneity of the in-sample θ-estimates as fast as possible.

Finally, we note that the dominant error term in (11) is due to the sampling variance of regression trees, and is the same term that appears in the analysis of Athey and Imbens [2016]. Including this error term in the splitting criterion may stabilize the construction of the tree, and further it can prevent the splitting criterion from favoring splits that make the model difficult to estimate, for example, splits where there is not sufficient variation in the data to estimate the model parameters within the resulting child leaves.

3.3 The Gradient Tree Algorithm

The above discussion provides some helpful conceptual guidance on how to pick good splits. However, from a computational perspective, actually optimizing the criterion $\Delta(C_1,\, C_2)$ over all possible axis-aligned splits while explicitly solving for $\hat{\theta}_{C_1}$ and $\hat{\theta}_{C_2}$ in each candidate child using an analogue to (7) may be quite expensive.

To avoid this issue, we instead optimize an approximate criterion $\widetilde{\Delta}(C_1,\, C_2)$ built using gradient-based approximations for $\hat{\theta}_{C_1}$ and $\hat{\theta}_{C_2}$: For each child C, we use $\tilde{\theta}_C \approx \hat{\theta}_C$ with

$$\tilde{\theta}_C = \hat{\theta}_P - \frac{1}{\left|\{i : X_i \in C\}\right|} \sum_{\{i : X_i \in C\}} \xi^\top A_P^{-1}\, \psi_{\hat{\theta}_P,\, \hat{\nu}_p}\left(O_i\right), \tag{12}$$

$$A_P = \frac{1}{\left|\{i : X_i \in P\}\right|} \sum_{\{i : X_i \in P\}} \nabla \psi_{\hat{\theta}_P,\, \hat{\nu}_P}\left(O_i\right), \tag{13}$$

where $\hat{\theta}_P$ and $\hat{\nu}_P$ are obtained by solving (7) once in the parent node, and ξ is a vector that picks out the θ-coordinate from the $(\theta,\, \nu)$ vector. Similar gradient-based approximations also underlie other popular statistical algorithms, including gradient boosting [Friedman, 2001] and the tree-building algorithm of Zeileis et al. [2008].

Algorithmically, our recursive partitioning scheme reduces to alternatively applying the following two steps. First, in a **labeling step**, we compute $\hat{\theta}_P$, $\hat{\nu}_P$, and the derivative matrix A_P^{-1} on the parent data as in (7), and use them to get pseudo-outcomes

$$\widetilde{Y}_i = -\xi^\top A_P^{-1}\, \psi_{\hat{\theta}_P,\, \hat{\nu}_P}\left(O_i\right) \in \mathbb{R}. \tag{14}$$

Next, in a **regression step**, we run a standard CART regression split on the pseudo-outcomes $\widetilde{Y}_i$. Specifically, we split P into two axis-aligned children C_1 and C_2 such as to

Algorithm 1 Gradient forests with honesty and subsampling

Note: All tuning parameters, such as the total number of trees B and the sub-sampling rate used in SUBSAMPLE, are taken as pre-specified.

procedure GRADIENTFOREST(set of examples $\mathcal{S}$, test point x)
  weight vector $\alpha \leftarrow 0$
  for $b = 1$ to total number of trees B **do**
    set of examples $\mathcal{I} \leftarrow$ SUBSAMPLE($\mathcal{S}$)
    sets of examples $\mathcal{J}_1$, $\mathcal{J}_2 \leftarrow$ SPLITSAMPLE($\mathcal{I}$)
    tree $\mathcal{T} \leftarrow$ GRADIENTTREE($\mathcal{J}_1$) ▷ Grows a tree by recursive partitioning, alternating the steps (14) and (15).
    $\mathcal{N} \leftarrow$NEIGHBORS(x, $\mathcal{T}$, $\mathcal{J}_2$) ▷ Returns those elements of $\mathcal{J}_2$ that fall into the same leaf as x in the tree $\mathcal{T}$.
    for all example $e \in \mathcal{N}$ **do**
      α[SAMPLEINDEX(e)] += $1/|\mathcal{N}|$
  output $\hat{\theta}(x)$, the solution to (2) with weights α/B

maximize the criterion

$$\widetilde{\Delta}(C_1, C_2) = \sum_{j=1}^{2} \frac{-1}{|\{i : X_i \in C_j\}|} \left(\sum_{\{i:X_i \in C_j\}} \widetilde{Y}_i \right)^2. \tag{15}$$

As discussed earlier, the most computationally intensive step in this algorithm is the regression step; and, thanks to our two-step structuring of our method, this step can be executed using standard optimized software for CART regression trees. Finally, once we have executed the regression step, we solve the estimating equation exactly in each child separately and then proceed to relabel the observations, etc.

As theoretical justification for our gradient-based approximation, we can verify that the error from using the approximate criterion (15) instead of the exact Δ-criterion (10) is within the tolerance used to motivate the Δ-criterion in Proposition 1.

Proposition 2. *Under the conditions of Proposition 1,*

$$\widetilde{\Delta}(C_1, C_2) = \Delta(C_1, C_2) + o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right). \tag{16}$$

As a final sanity check on our procedure, we note that in the case of least-squares regression, i.e., with $\psi_\theta(x)(Y) = Y - \theta(x)$, the labeling step (14) does not change the problem away from standard CART : $\widetilde{Y}_i = Y_i - \overline{Y}_p$, where $\overline{Y}_p$ is the mean outcome in the parent. Thus, in the case of regression, our gradient trees are equivalent to growing a standard CART regression tree.

Remark: Non-differentiable estimating equations. In some cases of interest, such a quantile regression, the derivative matrix A_P in (13) may not be well-defined. In such cases, we can replace A_p in the definition of the pseudo-outcomes (14) with $\widehat{A}_p$, a consistent estimate for the gradient of the expectation of the ψ-function, i.e., $\nabla \mathbb{E}[\psi_{\hat{\theta}_P, \hat{\nu}_P}(O) \,|\, X = x]$.

3.4 Building a Forest: Consistency, Honesty, and Subsampling

Now, given a practical splitting scheme for growing individual trees, our goal is to grow a forest that allows for consistent estimation of $\theta(x)$ using (2) paired with the forest weights (5). At a high level, we expect each tree to provide small, relevant neighborhoods for x that give us noisy estimates of $\theta(x)$. However, if every tree has different small, relevant neighborhoods for x, we may hope that forest-based aggregation will provide a single larger but still relevant neighborhood for x that yields stable estimates $\hat{\theta}(x)$.

To ensure that forest-based aggregation succeeds in providing the kind of stability discussed above, we rely on two conceptual ideas that have proven to be successful in the literature on forest-based least-squares regression: Training trees on subsamples of the training data [Mentch and Hooker, 2016, Scornet et al., 2015, Wager and Athey, 2015], and a subsample splitting technique that we call honesty [Biau, 2012, Denil et al., 2014, Wager and Athey, 2015]. Our final algorithm for forest-based solutions to heterogeneous estimating equations is given as Algorithm 1; we refer to Wager and Athey [2015] for a more in-depth discussion of subsampling and honesty in the context of forests.

As we will show in the theoretical analysis in Section 6, assuming regularity conditions, the estimates $\hat{\theta}(x)$ obtained using a gradient forest as described in Algorithm 1 are consistent for $\theta(x)$. Moreover, given appropriate subsampling rates, we can extend the analysis of Wager and Athey [2015] to show asymptotic normality of the resulting forest estimates $\hat{\theta}(x)$. Before discussing theory, however, we first review some concrete instantiations of our gradient forests.

4 Application: Quantile Regression Forests

As a first application gradient forests, we consider the problem of quantile regression with random forests. This problem has also been considered in detail by Meinshausen [2006], who proposed a consistent forest-based quantile regression algorithm; his method also fits into the paradigm of solving estimating equations (2) using random forest weights (5). However, unlike us, Meinshausen [2006] does not propose a splitting rule that is tailored to the quantile regression context, and instead builds his forests using plain CART regression splits. Thus, a comparison of our method with that of Meinshausen [2006] provides a perfect opportunity for evaluating the value of our splitting scheme.

Recall that, in the language of estimating equations, the q-th quantile $\theta_q(x)$ of the distribution of Y conditionally on $X = x$ is identified via (1), using the moment function

$$\psi_{\theta(x)}(Y_i) = q\mathbf{1}\left(\{Y_i > q\}\right) - (1-q)\mathbf{1}\left(\{Y_i \leq q\}\right). \tag{17}$$

Plugging this moment function into our splitting scheme, (14) gives us pseudo-outcomes

$$\widetilde{Y}_i = \mathbf{1}\left(\left\{Y_i > \hat{\theta}_{q,\,P}\right\}\right) \text{ where } \hat{\theta}_{q,\,P} \text{ is the } q\text{-th quantile of the parent } P, \tag{18}$$

up to a scaling and re-centering that do not affect the subsequent regression split on these pseudo-outcomes. In other words, gradient-based quantile regression trees simply try to separate observations that fall above the q-th quantile of the parent from those below it.

We compare our method to that of Meinshausen [2006] in Figure 2. In the left panel, we have a mean shift in the distribution of Y_i conditional on X_i at $(X_i)_1 = 0$, and both methods are able to pick it up as expected. However, in the right panel, the mean of Y

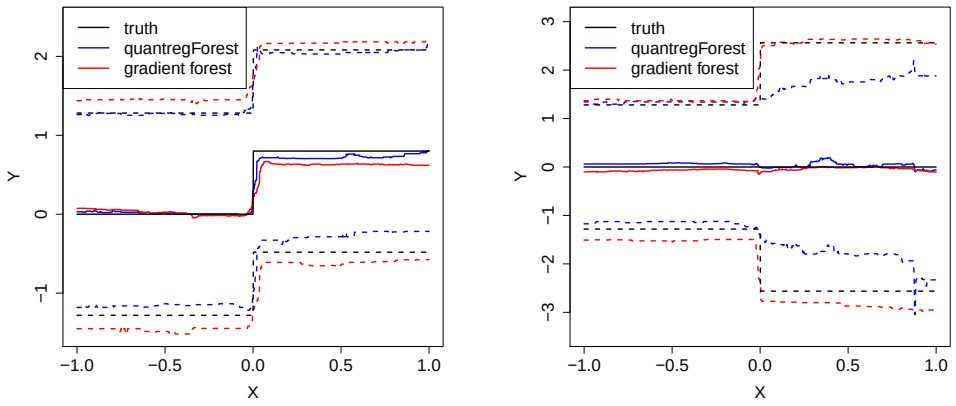

Figure 2: Comparison of quantile regression using our gradient forests and the `quantregForest` package of Meinshausen [2006]. In both cases, we have $n = 20,000$ independent and identically distributed examples where X_i is uniformly distributed over $[-1, 1]^p$ with $p = 40$, and Y_i is Gaussian conditionally on $(X_i)_1$. The other 39 covariates are noise. We estimate the quantiles at $q = 0.1, 0.5, 0.9$.

given X is constant, but there is a scale shift at $(X_i)_1 = 0$. Here, our method still performs well, as our splitting rule targets changes in the quantiles of the Y-distribution. However, the method of Meinshausen [2006] breaks down completely, as it relies on CART regression splits that are only sensitive to changes in the conditional mean of Y given X.

Remark: Estimating many quantiles. In many cases, we want to estimate multiple quantiles at the same time; for example, in Figure 2, we sought to get $q = 0.1, 0.5, 0.9$ at the same time. Estimating different forests for each quantile separately would be undesirable for many reasons: It would be computationally expensive and, moreover, there is a risk that quantile estimates might cross in finite samples due to statistical noise. Thus, we need to build a forest using a splitting scheme that is sensitive to changes at any of our quantiles of interests. Here, we use a simple heuristic inspired by the relabling transformation (18). Given a set of quantiles of interest $q_1 < ... < q_k$, we first evaluate all these quantiles $\hat{\theta}_{q_1, P} \leq ... \leq \hat{\theta}_{q_k, P}$ in the parent node, and label i-th point by the interval $[\hat{\theta}_{q_{j-1}, P}, \hat{\theta}_{q_j, P})$ it falls into. Then, we choose the split point using a multiclass classification rule that classifies each observation into one of the intervals.

5 Application: Heterogeneous Treatment Effect Estimation via Instrumental Variables

In many economics applications, we want to measure the causal effect of an intervention on some outcome, all while recognizing that the intervention and the outcome may also be tied together through non-causal pathways. A popular approach in this situation is to rely

on instrumental variables (IV) regression, where we find an auxiliary source of randomness that can be used to identify causal effects.

As a concrete example, suppose we want to measure the causal effect of child rearing on a mother's labor-force participation. It is well known that, in the United States, mothers with more children are less likely to work. But how much of this link causal, i.e., some mothers work less because they are busy raising children, and how much of it is merely due to confounding factors, e.g. some mothers have preferences that both lead them to raise more children and be less likely to participate in the labor force? Understanding effects like this may be helpful in predicting the value of programs like subsidized daycare that assist mothers' labor force participation while they have young children.

To study this question, Angrist and Evans [1998] found a source of auxiliary randomness that can be used to distinguish causal versus correlational effects: They found that, in the United States, parents who already have two children of mixed sexes, i.e., one boy and one girl, will have fewer kids in the future than parents whose first two children were of the same sex. Assuming that the sexes of the first two children in a family are effectively random, this observed preference for having children of both sexes provides an exogenous source of variation in family size that can be used to identify causal effects: if the mixed sex indicator is unrelated to the mother's propensity to work for a fixed number of children, then the effect of the mixed sex indicator on the observed propensity to work can be attributed to its effect on family size. The instrumental variable estimator normalizes this effect by the effect of mixed sex on family size, so that the normalized estimate is a consistent estimate of the treatment effect of family size on work. Other classical uses of instrumental variables regression include measuring the impact of military service on lifetime income by using the Vietnam draft lottery as an instrument [Angrist, 1990], and measuring the extent to which 401(k) savings programs crowd out other savings, using eligibility for 401(k) savings programs as an instrument [Abadie, 2003, Poterba et al., 1996].

5.1 A Forest for Instrumental Variables Regression

Classical approaches to instrumental variables regression only seek a global understanding of the treatment effect: for example, on average over the whole US population, does having more children reduce the labor force participation of women? Here, we seek to use forests to answer a more ambitious question, and estimate a heterogeneous treatment effect: we might ask how the causal effect of child rearing varies with a mother's age and socioeconomic status.

Suppose that we observe $i = 1, ..., n$ independent and identically distributed subjects, each of whom has features $X_i \in \mathcal{X}$, an outcome $Y_i \in \mathbb{R}$, a treatment assignment $W_i \in \mathbb{R}$, and an instrument $Z_i \in \mathbb{R}$. We believe that the outcomes Y_i and treatment assignment W_i are related via a structural model[2]

$$Y_i = \mu\left(X_i\right) + \tau\left(X_i\right) W_i + \varepsilon_i, \tag{19}$$

where $\tau(X_i)$ is understood to be the causal effect of W_i on Y_i, and ε_i is a noise term that may be positively correlated with W_i. Because ε_i is correlated with W_i, standard regression analyses will not in general be consistent for $\tau(X_i)$. This is where we need to use the

[2]If we are not willing to assume that every individual i with features $X_i = x$ has the same treatment effect $\tau(x)$, then heterogeneous instrumental variables regression allows us to estimate a (conditional) local average treatment effect [Imbens and Angrist, 1994]; see, e.g., Abadie [2003]. Here, however, we use the additive structure (19) for simplicity of exposition.

instrument Z_i. Suppose we know that Z_i is independent of ε_i conditionally on X_i. Then, provided that Z_i has an influence on the treatment assignment W_i, i.e., that the covariance of Z_i and W_i conditionally on $X_i = x$ is non-zero, we can verify that the treatment effect $\tau(x)$ is identified via

$$\tau(x) = \text{Cov}\left[Y_i,\, Z_i \,\middle|\, X_i = x\right] \,/\, \text{Cov}\left[W_i,\, Z_i \,\middle|\, X_i = x\right]. \tag{20}$$

We can use the above moment-based identification to estimate $\tau(x)$ in practice by solving an estimating equation (1) with a moment function (4) [e.g., Angrist and Pischke, 2008].

We can again use our gradient-based formalism to derive a forest that is targeted towards estimating causal effects identified via (20). When growing the forest, the gradient-based labeling (14) gives us pseudo-outcomes

$$\widetilde{Y}_i = \left(Z_i - \overline{Z}_P\right)\left(\left(Y_i - \overline{Y}_P\right) - \left(W_i - \overline{W}_P\right)\hat{\tau}_P\right), \tag{21}$$

where $\overline{Y}_P$, $\overline{W}_P$, $\overline{Z}_P$ are moments in the parent node, and $\hat{\tau}_P$ is a solution to the estimating equation with moments (4) in the parent. Then, given these pseudo-outcomes, the tree executes a CART regression split on the $\widetilde{Y}_i$ as usual. Finally, we obtain personalized treatment effect estimates $\hat{\tau}(x)$ by solving the estimation equation (2) with forest weights (5).

Remark: Causal forests with exogenous treatment There has been considerable recent interest in estimating heterogeneous treatment effects with forests with the treatment assignment W_i is exogenous or unconfounded, i.e., W_i is independent of ε_i conditionally on X_i in (19) [Green and Kern, 2012, Hill, 2011, Wager and Athey, 2015]. Formally, this is a special case of our instrumental variables setup where we use the treatment itself as an instrument, i.e., we set $Z_i := W_i$. In this case, our forest makes splits using pseudo-outcomes

$$\widetilde{Y}_i = \left(W_i - \overline{W}_P\right)\left(\left(Y_i - \overline{Y}_P\right) - \left(W_i - \overline{W}_P\right)\hat{\tau}_P\right); \tag{22}$$

and we can verify that running CART regression splits using the the above $\widetilde{Y}_i$ is closely related to the causal splitting rule advocated by Athey and Imbens [2016]. We find it reassuring that, in simple cases, our gradient trees use splitting rules that resemble those that were motivated with more direct arguments.

5.2 Simulation Study

We illustrate the behavior of IV forests in Figure 3 using two simple simulation designs. In both examples, X is uniformly spread over a cube, $X_i \sim [-1,\, 1]^p$, but the causal effect $\tau(X_i)$ only depends on the first coordinate $(X_i)_1$. In both panels of Figure 3, we show estimates of $\tau(x)$ produced by different methods, where we vary x_1 and set all other coordinates to 0.

In the first panel, we illustrate the importance of using an IV forests when the treatment assignment may be endogenous. We consider a case where the true causal effect of has a single jump, $\tau(X_i) = 2 \times \mathbf{1}\left(\{(X_i)_1 > -1/3\}\right)$. Meanwhile, at $(X_i)_1 = +1/3$, there is a change in the correlation structure between W_i and ε_i that leads to a spurious (i.e., non-causal) jump in the correlation between W_i and Y_i. As expected, our IV forest correctly picks out the first jump while ignoring the second one. Conversely, a plain causal forest following Wager and Athey [2015] that assumes that the treatment assignment W_i is exogenous will mistakenly also pick out the second spurious jump in the correlation structure of W_i and Y_i.

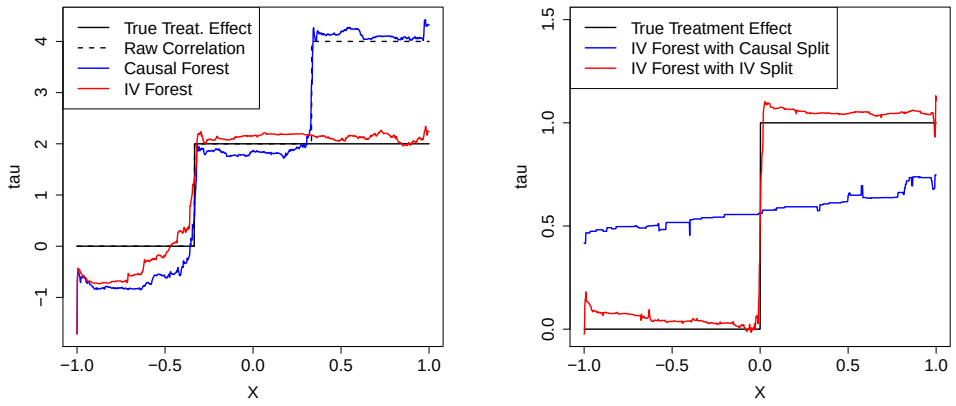

Figure 3: In both panels, we generate data as $X_i \sim [-1,\, 1]^p$, with $n = 10,000$ and $p = 20$.

Meanwhile, in the second panel, we test our splitting rule. We have a simulation design where there is a jump in the true causal effect, $\tau(X_i) = \mathbf{1}\left(\{(X_i)_1 > 0\}\right)$. However this causal effect is masked by a change in the correlation of W_i and ε_i, such that the joint distribution of W_i and Y_i does not depend on X_i. Here, the IV forest described in the previous section again performs well. However, if we try to use the simpler causal tree splitting rule of Athey and Imbens [2016] that was not designed for IV regression instead of our proposed splitting rule with pseudo-outcomes (21), then the forest fails to detect any signal.

6 Theoretical Analysis

6.1 Basic Assumptions

We begin by listing the basic assumptions underlying all of our theoretical results. First, we assume that the covariate space and the parameter space are both subsets of Euclidean space; specifically, we assume that $\mathcal{X} = [0,\, 1]^p$ and $(\theta,\, \nu) \in \mathbb{R}^k$ for some $p,\, k > 0$. Moreover, we assume that the features X have a density that is bounded away from 0 and ∞; as argued in, e.g., Wager and Walther [2015], this is equivalent to imposing a weak dependence condition on the individual features $(X_i)_j$ because trees and forests are invariant to monotone rescaling of the features.

Regularity of moments We make assumptions about how the expected moment functions change both when we vary x with $(\theta,\, \nu)$ fixed and when we vary $(\theta,\, \nu)$ with x fixed. Throughout, we will write

$$M_{\theta,\,\nu}(x) := \mathbb{E}\left[\psi_{\theta,\,\nu}(O) \,\middle|\, X = x\right] \tag{23}$$

for the expectation of the ψ-function. When we hold $(\theta,\, \nu)$ fixed, we assume that $M_{\theta,\,\nu}(x)$ is Lipschitz continuous in x. Meanwhile, when x is fixed, we assume that this M-function is continuously differentiable in $(\theta,\, \nu)$, and that its derivative $V(x)$ at the true parameter

value $(\theta(x),\, \nu(x))$,

$$V(x) = \frac{\partial}{\partial(\theta,\, \nu)} M_{\theta,\, \nu}(x) \,\big|_{\,\theta(x),\ \nu(x)}, \tag{24}$$

is invertible. Finally, we assume that for any fixed parameter value $(\theta,\, \nu)$, the second moments $\mathbb{E}[\psi_{\theta,\, \nu}^{\otimes 2}(O) \,\big|\, X = x]$ are uniformly bounded over all $x \in \mathcal{X}$.

Smoothness of the estimating equation In addition to assuming that the moment function defined above is regular, we also require that the ψ-functions used to define the estimating equation themselves be smooth. Specifically, we assume that $\psi_{\theta,\, \nu}(O)$ is twice continuously differentiable in θ, ν for any fixed O, and that $\text{Var}\left[\nabla \psi_{\theta,\, \nu}(O) \,\big|\, X = x\right]$ is bounded. We also assume that, with probability tending to 1, the estimating equation (2) has a unique solution $\sum_{i=1}^n \alpha_i(x)\, \psi_{\hat{\theta}(x),\, \hat{\nu}(x)}(O_i) = 0$. In its current form, this set of assumptions is rather restrictive, and in particular does not apply to quantile regression. However, we believe that these assumptions can be considerably weakened while still preserving the validity of our results, and intend to pursue this line of work.

Assumptions about the forest Our consistency and Gaussianty results also require some control on the behavior of the trees comprising the forest. To do so, we follow Wager and Athey [2015]. We assume that our trees are symmetric, in that their output is invariant to permuting the indices of the training examples. We also assume that the tree makes balanced splits, in the sense that every split puts at least a fraction ω of the observations in the parent node into each child, for some $\omega > 0$. Finally, we take the tree to be randomized in such a way that, at every split, the probability that the tree splits on the j-th feature is bounded from below by some $\pi > 0$.

6.2 Approximating Gradient Forests with Regression Forests

Our proof strategy is built using the method of influence functions [Hampel, 1974]. In our context, similar technical ideas also underlie the analysis of Newey [1994]. Here, the influence function heuristic motivates a way to approximate gradient forests with a class of regression forests. The upshot is that we can analyze the approximating regression forests using tools developed in Wager and Athey [2015], and then a use coupling result to derive conclusions about gradient forests.

To construct such an approximating forest, let $\widetilde{Y}_i^*(x)$ denote the influence function of the i-th observation with respect to the true parameter value $\theta(x)$:

$$\widetilde{Y}_i^*(x) = -\xi^\top V(x)^{-1} \psi_{\theta(x),\, \nu(x)}(O_i). \tag{25}$$

These quantities are closely related to the pseudo-outcomes (14) used in our gradient tree splitting rule; the main difference is that, here, the quantities $\widetilde{Y}_i^*(x)$ depend on the unknown true parameter values at x and are thus inaccessible in practice. We use the $*$-superscript to remind ourselves of this fact.

Now, given any set of forest weights $\alpha_i(x)$ used to define the gradient forest estimate $\hat{\theta}(x)$ by solving (2), we can also define a pseudo-forest as follows:

$$\tilde{\theta}^*(x) = \theta(x) + \sum_{i=1}^n \alpha_i \widetilde{Y}_i^*(x). \tag{26}$$

Formally, this pseudo-forest estimate $\tilde{\theta}^*(x)$ is equivalent to the output of a regression forest with weights $\alpha_i(x)$ and outcomes $\theta(x) + \widetilde{Y}_i^*(x)$. The following result establishes a coupling between the gradient forest output we want to study, $\hat{\theta}(x)$, and our pseudo-forest approximation $\tilde{\theta}^*(x)$.

Lemma 3. *Given our basic assumptions, suppose that the gradient forest estimator $\hat{\theta}(x)$ is consistent for $\theta(x)$. Then $\hat{\theta}(x)$ and $\tilde{\theta}^*(x)$ are coupled,*

$$\tilde{\theta}^*(x) - \hat{\theta}(x) = o_P\left(\left\|\sum_{i=1}^n \alpha_i(x)\,\psi_{\theta(x),\,\nu(x)}\left(O_i\right)\right\|_2\right). \tag{27}$$

6.3 Asymptotic Theory for Gradient Forests

We can now build on the results of Wager and Athey [2015] to provide a Gaussian asymptotic characterization of $\hat{\theta}^*(x)$. Our first task is to establish consistency of $\hat{\theta}(x)$, to activate the guarantees from Lemma 3.

Lemma 4. *Maintain the basic assumptions detailed above, and in addition, assume that we obtain $(\hat{\theta}(x),\, \hat{\nu}(x))$ by solving (2) with forest weights (5), where our forest is an honest, subsampled forest with subsample size s satisfying $s/n \to 0$ and $s \to \infty$. Then, $(\hat{\theta}(x),\, \hat{\nu}(x))$ converges in probability to $(\theta(x),\, \nu(x))$ as $n \to \infty$.*

Given the above results, it now remains to study the asymptotic behavior of the pseudo-forest $\tilde{\theta}^*(x)$. The key idea is that, because $\tilde{\theta}^*(x)$ is a linear function of the pseudo-outcomes $\widetilde{Y}_i^*(x)$, we can write it as an average of tree predictions

$$\tilde{\theta}^*(x) = \frac{1}{B}\sum_{b=1}^B \tilde{\theta}_b^*(x), \quad \tilde{\theta}_b^*(x) = \sum_{i=1}^n \alpha_{ib}\left(\theta(x) + \widetilde{Y}_i^*(x)\right). \tag{28}$$

Given this representation, we see that $\tilde{\theta}^*(x)$ is a U-statistic, and so can be decomposed via the machinery of Efron and Stein [1981], as in the analysis of Wager and Athey [2015]. Pursuing this approach, we can show that whenever trees are grown on subsamples of size s scaling as $s = n^\beta$ for some $\beta_{\min} < \beta < 1$, $\tilde{\theta}^*(x)$ is asymptotically normal.

Theorem 5. *Under the conditions of Lemma 4, suppose moreover that trees are grown on subsamples of size s with*

$$s = n^\beta \quad \textit{for some} \quad \beta_{\min} := 1 - \left(1 + \frac{1}{\pi}\frac{\log\left(\omega^{-1}\right)}{\log\left((1-\omega)^{-1}\right)}\right)^{-1} < \beta < 1, \tag{29}$$

where π and ω are constants defined when stating basic assumptions about the forest. Finally, suppose that $\mathrm{Var}[\widetilde{Y}_i^*(x) \,|\, X_i = x] > 0$. Then, there is a sequence $\sigma_n(x)$ for which

$$\left(\hat{\theta}(x) - \theta(x)\right) \,/\, \sigma_n(x) \Rightarrow \mathcal{N}\left(0,\, 1\right), \quad \lim_{n\to\infty} \sigma_n(x) = 0. \tag{30}$$

7 Appendix: Proofs

The proofs are presented in order of logical dependence. Many of our results use the shorthand

$$\Psi_{\alpha(x)}(\theta, \nu) := \sum_{i=1}^{n} \alpha_i(x)\psi_{\theta,\nu}(O_i). \tag{31}$$

where the $\alpha_i(x)$ are our forest weights (5).

Proof of Lemma 3

By taking a Taylor expansion of ψ around $(\theta(x), \nu(x))$ we see that

$$\begin{aligned} 0 = \Psi_{\alpha(x)}\left(\hat{\theta}(x), \hat{\nu}(x)\right) &= \Psi_{\alpha(x)}(\theta(x), \nu(x)) \\ &+ \nabla\Psi_{\alpha(x)}(\theta(x), \nu(x)) \begin{pmatrix} \hat{\theta}(x) - \theta(x) \\ \hat{\nu}(x) - \nu(x) \end{pmatrix} + H \begin{pmatrix} \hat{\theta}(x) - \theta(x) \\ \hat{\nu}(x) - \nu(x) \end{pmatrix}, \end{aligned} \tag{32}$$

for some matrix $H(x)$, where $\|H\|_2 \to_p 0$ by consistency of $\hat{\theta}(x)$. Now, because $\nabla\psi$ has a bounded variance, we can use the same argument as in the proof of Lemma 4 to verify that $\nabla\Psi_{\alpha(x)}(\theta(x), \nu(x)) \to_p V(x)$, where $V(x)$ is the expected derivative matrix (24). Thus, we can re-arrange the above expression as

$$\begin{aligned} \begin{pmatrix} \hat{\theta}(x) - \theta(x) \\ \hat{\nu}(x) - \nu(x) \end{pmatrix} &= (V(x) + o_P(1))^{-1} \Psi_{\alpha(x)}(\theta(x), \nu(x)) \\ &= V(x)^{-1}\Psi_{\alpha(x)}(\theta(x), \nu(x)) + o_P\left(\left\|\Psi_{\alpha(x)}(\theta(x), \nu(x))\right\|_2\right). \end{aligned} \tag{33}$$

When restricted to only the first coordinate θ, this in turn becomes

$$\hat{\theta}(x) = \tilde{\theta}^*(x) + o_P\left(\left\|\Psi_{\alpha(x)}(\theta(x), \nu(x))\right\|_2\right). \tag{34}$$

Lemma 6. *Under the conditions of Lemma 4, the quantity $\Psi_{\alpha(x)}(\theta(x), \nu(x))$ defined as in (31) above satisfies the following moment bounds*

$$\mathbb{E}\left[\Psi_{\alpha(x)}(\theta(x), \nu(x))\right] = \mathcal{O}\left(s^{-\frac{\pi}{2}\frac{\log\left((1-\omega)^{-1}\right)}{\log(\omega^{-1})}}\right) \tag{35}$$

$$\text{Var}\left[\Psi_{\alpha(x)}(\theta(x), \nu(x))\right] = \mathcal{O}(s/n). \tag{36}$$

Proof. We start by expanding Ψ as

$$\Psi_{\alpha(x)}(\theta, \nu) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} \alpha_{bi}(x)\psi_{\theta,\nu}(O_i), \tag{37}$$

where the α_{bi} are the individual tree weights used to build the forest weights in (5). Now, $\Psi_{\alpha(x)}(\theta, \nu)$ is nothing but the output of a regression forest with response $\psi_{\theta,\nu}(O_i)$. Thus, given our assumptions about the moments of $\psi_{\theta,\nu}(O_i)$ and the fact that our trees are built via honest subsampling, (35) follows immediately from Theorem 3 of Wager and Athey [2015]. Meanwhile, because individual trees are grown on subsamples, we can verify that

$$\frac{n}{s}\text{Var}\left[\Psi_{\alpha(x)}(\theta(x), \nu(x))\right] \leq \text{Var}\left[\sum_{i=1}^{n} \alpha_{bi}(x)\psi_{\theta,\nu}(O_i)\right] = \mathcal{O}(1), \tag{38}$$

where the first inequality results from classical results about U-statistics going back to Hoeffding [1948], while the second inequality follows from second-moment bounds on ψ along with the fact that our trees are grown on honest subsamples. □

Proof of Lemma 4

Lemma 6 in particular implies that $\Psi_{\alpha(x)}\left(\theta(x),\, \nu(x)\right) \to_p 0$. Thus, thanks to our smoothness assumptions on the estimating equation, we can use a version the inverse function theorem as stated in, e.g., Theorem 4.2 of Wang [1999] to conclude that there exists a sequence $(\hat{\theta}(x),\, \hat{\nu}(x))$ with

$$\lim_{n\to\infty} \mathbb{P}\left[\Psi_{\alpha(x)}\left(\hat{\theta}(x),\, \hat{\nu}(x)\right) = 0\right] = 1 \text{ such that } \left(\hat{\theta}(x),\, \hat{\nu}(x)\right) \to_p \left(\theta(x),\, \nu(x)\right). \tag{39}$$

Moreover, because the estimating equation has a unique root with high probability, we conclude that the forest estimates $(\hat{\theta}(x),\, \hat{\nu}(x))$ must match the consistent estimators produced above.

Proof of Theorem 5

As argued in Section 6.3, Theorem 5 of Wager and Athey [2015] immediately implies that, given the assumptions made by hypothesis,

$$\left(\tilde{\theta}^*(x) - \theta(x)\right) \,/\, \sigma_n(x) \Rightarrow \mathcal{N}\left(0,\, 1\right), \quad \sigma_n^2(x) = \widetilde{\Theta}\left(\frac{s}{n}\right). \tag{40}$$

Combining this result with Lemmas 3, 4 to establish a coupling between $\hat{\theta}(x)$ and $\tilde{\theta}^*(x)$, and Lemma 6 to control the tightness of the coupling, we obtain the desired result.

Proof of Proposition 2

Our goal is to couple the actual solution $\hat{\theta}_{C_j}$ of the estimating equation over the leaf C_j with the gradient-based approximation $\tilde{\theta}_{C_j}$ obtained by taking a single gradient step from the parent. Here, instead of directly establishing a relationship between these two quantities, we couple the both to the average of the influence functions $\widetilde{Y}_i^*(x)$ averaged over C_j, namely

$$\tilde{\theta}^*_{C_j}(x) = \theta(x) + \frac{1}{|C_j|}\sum_{i\in C_j} \widetilde{Y}_i^*(x). \tag{41}$$

Because the leaf C_j is considered fixed, we can use second-moment bounds on ψ to verify that $\mathrm{Var}[\tilde{\theta}^*_{C_j}(x)] = \mathcal{O}\left(1/n_{C_j}\right)$; meanwhile, by Lipschitz-continuity of the M-function (23), we see that $\mathbb{E}[\tilde{\theta}^*_{C_j}(x) - \theta(x)] = \mathcal{O}\left(r\right)$, where r is the radius of the leaf. Finally, given assumptions made so far about the estimating equation, it is straight-forward to show that $\hat{\theta}_{C_j}$ is consistent for $\theta(x)$ is a limit where $r \to 0$ and $n_{C_j} \to \infty$. Thus, a direct analogue to our result, Lemma 3, implies that

$$\tilde{\theta}^*_{C_j}(x) - \hat{\theta}_{C_j} = o_P\left(r,\, 1/\sqrt{n_{C_j}}\right). \tag{42}$$

Next, in order to couple $\tilde{\theta}_{C_j}(x)$ and $\tilde{\theta}^*_{C_j}(x)$, we note that

$$\begin{aligned}\tilde{\theta}_{C_j} - \tilde{\theta}^*_{C_j}(x) = \hat{\theta}_P - \theta(x) \\ + \xi^\top A_P^{-1} \sum_{i \in C_j} \left(\psi_{\hat{\theta}_P,\, \hat{\nu}_P}(O_i) - \psi_{\theta(x),\, \nu(x)}(O_i) \right) \\ + \xi^\top \left(A_P^{-1} - V(x)^{-1} \right) \sum_{i \in C_j} \psi_{\theta(x),\, \nu(x)}(O_i)\,;\end{aligned} \tag{43}$$

our goal is then to bound all three summands at the desired rate. The first term is already bounded by the same argument as $\tilde{\theta}^*_{C_j}(x) - \hat{\theta}_{C_j}(x)$. The middle summand is bounded by $\mathcal{O}(r^2)$ by smoothness of the ψ-function as we change θ and ν. Finally, the last summand can be bounded by showing that $A_P^{-1} \to_p V(x)^{-1}$, and that $\sum_{i \in C_j} \psi_{\theta(x),\, \nu(x)}(O_i) = \mathcal{O}_P\left(1/\sqrt{n_{C_j}}\right)$. Everything we have showed so far implies that

$$\tilde{\theta}_{C_j} - \hat{\theta}_{C_j} = o_P\left(r,\, 1/\sqrt{n_{C_j}}\right), \text{ for } j = 1,\, 2. \tag{44}$$

Finally, it is straight-forward to check that

$$\tilde{\theta}_{C_2} - \tilde{\theta}_{C_1} = O_P\left(r,\, 1/\sqrt{n_{C_1}},\, 1/\sqrt{n_{C_2}}\right), \tag{45}$$

which implies the desired for the coupling of $\Delta(C_1,\, C_2)$ and $\widetilde{\Delta}(C_1,\, C_2)$.

Proof of Proposition 1

First, we show that we can replace $\hat{\theta}_{C_j}$ with the influence-based approximation $\tilde{\theta}^*_{C_j}(x)$ when computing the error function $\text{err}(C_j)$, as follows

$$\begin{aligned}\text{err}(C_j) = \mathbb{E}\left[\left(\hat{\theta}_{C_j} - \theta(X) \right)^2 \,\middle|\, X \in C_j \right] = \mathbb{E}\left[\left(\tilde{\theta}^*_{C_j}(x) - \theta(X) \right)^2 \,\middle|\, X \in C_j \right] \\ + \underbrace{\left(\hat{\theta}_{C_j} - \tilde{\theta}^*_{C_j}(x) \right)^2}_{o_P\left(r^2,\, 1/n_{C_j}\right)} + 2 \underbrace{\left(\hat{\theta}_{C_j} - \tilde{\theta}^*_{C_j}(x) \right)}_{o_P\left(r,\, 1/\sqrt{n_{C_j}}\right)} \underbrace{\mathbb{E}\left[\tilde{\theta}^*_{C_j}(x) - \theta(X) \,\middle|\, X \in C_j \right]}_{\mathcal{O}(r^2)}.\end{aligned} \tag{46}$$

Using the above expansion, we find that

$$\text{err}\left(C_1,\, C_2\right) = \sum_{j=1}^{2} \frac{n_{C_j}}{n_P} \mathbb{E}\left[\left(\tilde{\theta}^*_{C_j}(x) - \theta(X) \right)^2 \,\middle|\, X \in C_j \right] + o_P\left(r^2,\, \frac{1}{n_{C_1}},\, \frac{1}{n_{C_2}} \right) \tag{47}$$

Next, as in all our results, we can invoke the fact that $\tilde{\theta}^*_{C_j}(x)$ behaves like a regression tree; in this case, we can now follow the argument of Athey and Imbens [2016]. We see that

$$\mathbb{E}\left[\left(\tilde{\theta}^*_{C_j}(x) - \theta(X) \right)^2 \,\middle|\, X \in C_j \right] = \text{Var}\left[\theta(X) \,\middle|\, X \in C_j \right] + \text{Var}\left[\tilde{\theta}^*_{C_j}(x) \right] + \mathcal{O}\left(r^4\right), \tag{48}$$

and so

$$\begin{aligned}
\operatorname{err}(C_1, C_2) &= \sum_{j=1}^{2} \frac{n_{C_j}}{n_P} \left(\operatorname{Var}\left[\theta(X) \,\middle|\, X \in C_j\right] + \operatorname{Var}\left[\tilde{\theta}^*_{C_j}(x)\right] \right) + o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right) \\
&= \operatorname{Var}\left[\theta(X) \,\middle|\, X \in P\right] + \sum_{j=1}^{2} \frac{n_{C_j}}{n_P} \operatorname{Var}\left[\tilde{\theta}^*_{C_j}(x)\right] + o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right) \\
&\quad - \frac{n_{C_1} n_{C_2}}{n_P} \left(\mathbb{E}\left[\theta(X) \,\middle|\, X \in C_2\right] - \mathbb{E}\left[\theta(X) \,\middle|\, X \in C_1\right] \right)^2 \\
&= \operatorname{Var}\left[\theta(X) \,\middle|\, X \in P\right] - \frac{n_{C_1} n_{C_2}}{n_P} \left(\tilde{\theta}^*_{C_2}(x) - \tilde{\theta}^*_{C_1}(x) \right)^2 \\
&\quad + \frac{n_{C_1} n_{C_2}}{n_P} \left(\left(\tilde{\theta}^*_{C_2}(x) - \tilde{\theta}^*_{C_1}(x) \right)^2 - \mathbb{E}\left[\tilde{\theta}^*_{C_2}(x) - \tilde{\theta}^*_{C_1}(x) \right]^2 \right) \\
&\quad + \sum_{j=1}^{2} \frac{n_{C_j}}{n_P} \operatorname{Var}\left[\tilde{\theta}^*_{C_j}(x)\right] + o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right).
\end{aligned}$$

Now, to parse this expression, note that, by the proof of Proposition 2,

$$\Delta(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P} \left(\tilde{\theta}^*_{C_2}(x) - \tilde{\theta}^*_{C_1}(x) \right)^2 + o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right). \tag{49}$$

Thus, writing

$$K(P) := \operatorname{Var}\left[\theta(X) \,\middle|\, X \in P\right] \tag{50}$$

as the split-independent error term, all that remains is a term

$$\begin{aligned}
\mathcal{E} := &\sum_{j=1}^{2} \frac{n_{C_j}}{n_P} \operatorname{Var}\left[\tilde{\theta}^*_{C_j}(x)\right] \\
&+ \frac{n_{C_1} n_{C_2}}{n_P} \left(\left(\tilde{\theta}^*_{C_2}(x) - \tilde{\theta}^*_{C_1}(x) \right)^2 - \mathbb{E}\left[\tilde{\theta}^*_{C_2}(x) - \tilde{\theta}^*_{C_1}(x) \right]^2 \right)
\end{aligned} \tag{51}$$

that captures the effect of sampling noise in estimating $\tilde{\theta}^*_{C_j}(x)$. This last term scales as

$$\mathcal{E} = \mathcal{O}_P\left(\frac{r}{\sqrt{n_{C_1}}}, \frac{r}{\sqrt{n_{C_2}}}, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right),$$

and so can be ignored since we assume that $n_P \gg r^{-2}$.

References

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003.

Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, 1997.

Donald WK Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, pages 821–856, 1993.

Joshua D Angrist. Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review*, pages 313–336, 1990.

Joshua D Angrist and William N Evans. Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, pages 450–477, 1998.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2008.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*, 2014.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey, Guido W Imbens, and Stefan Wager. Efficient inference of average treatment effects in high dimensions via approximate residual balancing. *arXiv preprint arXiv:1604.07125*, 2016.

Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Chris Hansen. Program evaluation with high-dimensional data. *arXiv preprint arXiv:1311.2645*, 2013.

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–138. ACM, 2009.

Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13:1063–1095, 2012.

Gérard Biau and Luc Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518, 2010.

Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *The Journal of Machine Learning Research*, 9:2015–2033, 2008.

Imad Bou-Hamad, Denis Larocque, and Hatem Ben-Ameur. A review of survival trees. *Statistics Surveys*, 5:44–71, 2011.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman. Consistency for a simple model of random forests. *Statistical Department, University of California at Berkeley. Technical Report*, (670), 2004.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and Regression Trees*. CRC press, 1984.

Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

Hugh A Chipman, Edward I George, and Robert E McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Antonio Ciampi, Johanne Thiffault, Jean-Pierre Nakache, and Bernard Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational Statistics & Data Analysis*, 4(3):185–204, 1986.

Misha Denil, David Matheson, and Nando De Freitas. Narrowing the gap: Random forests in theory and in practice. In *Proceedings of The 31st International Conference on Machine Learning*, pages 665–673, 2014.

Bradley Efron. Estimation and accuracy after model selection (with discussion). *Journal of the American Statistical Association*, 109(507), 2014.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596, 1981.

Jianqing Fan, Mark Farmen, and Irene Gijbels. Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60 (3):591–608, 1998.

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

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

Louis Gordon and Richard A Olshen. Tree-structured survival analysis. *Cancer Treatment Reports*, 69(10):1065–1069, 1985.

Donald P Green and Holger L Kern. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511, 2012.

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Bruce E Hansen. Testing for parameter instability in linear models. *Journal of Policy Modeling*, 14(4):517–533, 1992.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 2011.

Nils Lid Hjort and Alexander Koning. Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, 14(1-2):113–132, 2002.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Bo E Honoré and Ekaterini Kyriazidou. Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68(4):839–874, 2000.

Guido W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Hemant Ishwaran and Udaya B Kogalur. Consistency of random survival forests. *Statistics & Probability Letters*, 80(13):1056–1064, 2010.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

Nathan Kallus. Learning to personalize from observational data. *arXiv preprint arXiv:1608.08925*, 2016.

Roger Koenker. *Quantile regression*. Number 38. Cambridge University Press, 2005.

Michael LeBlanc and John Crowley. Relative risk trees for censored survival data. *Biometrics*, pages 411–425, 1992.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

Nicolai Meinshausen. Quantile regression forests. *The Journal of Machine Learning Research*, 7:983–999, 2006.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26):1–41, 2016.

Annette M Molinaro, Sandrine Dudoit, and Mark J Van der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1):154–177, 2004.

Whitney K Newey. Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, 10(02):1–21, 1994.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

Jukka Nyblom. Testing for the constancy of parameters over time. *Journal of the American Statistical Association*, 84(405):223–230, 1989.

Werner Ploberger and Walter Krämer. The CUSUM test with OLS residuals. *Econometrica: Journal of the Econometric Society*, pages 271–285, 1992.

James M Poterba, Steven F Venti, and David A Wise. How retirement saving programs increase saving. *The Journal of Economic Perspectives*, 10(4):91–112, 1996.

James M Robins and Yaacov Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16, 1997.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Richard J Samworth. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

Joseph Sexton and Petter Laake. Standard errors for bagged and random forest estimators. *Computational Statistics & Data Analysis*, 53(3):801–811, 2009.

Joan G Staniswalis. The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84(405):276–283, 1989.

Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research*, 10: 141–158, 2009.

Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. A nonparametric Bayesian analysis of heterogeneous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, (just-accepted), 2016.

Robert Tibshirani and Trevor Hastie. Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567, 1987.

Mark J van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*, 2015.

Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15, 2014.

Xinghua Wang. Convergence of Newton's method and inverse function theorem in Banach space. *Mathematics of Computation of the American Mathematical Society*, 68(225):169–186, 1999.

Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*, 2015.

Achim Zeileis. A unified approach to structural change tests based on ML scores, F statistics, and OLS residuals. *Econometric Reviews*, 24(4):445–466, 2005.

Achim Zeileis and Kurt Hornik. Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, 61(4):488–508, 2007.

Achim Zeileis, Torsten Hothorn, and Kurt Hornik. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514, 2008.