Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Estimation and Inference of Heterogeneous Treatment Effects using Random Forests

Stefan Wager & Susan Athey

To cite this article: Stefan Wager & Susan Athey (2018) Estimation and Inference of Heterogeneous Treatment Effects using Random Forests, Journal of the American Statistical Association, 113:523, 1228-1242, DOI: 10.1080/01621459.2017.1319839

To link to this article: https://doi.org/10.1080/01621459.2017.1319839

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 523, 1228–1242, Theory and Methods
https://doi.org/10.1080/01621459.2017.1319839

Estimation and Inference of Heterogeneous Treatment Effects using Random Forests

Stefan Wager and Susan Athey

Stanford University, Stanford, CA

ABSTRACT

Many scientific and engineering challenges—ranging from personalized medicine to customized marketing recommendations—require an understanding of treatment effect heterogeneity. In this article, we develop a nonparametric *causal forest* for estimating heterogeneous treatment effects that extends Breiman's widely used random forest algorithm. In the potential outcomes framework with unconfoundedness, we show that causal forests are pointwise consistent for the true treatment effect and have an asymptotically Gaussian and centered sampling distribution. We also discuss a practical method for constructing asymptotic confidence intervals for the true treatment effect that are centered at the causal forest estimates. Our theoretical results rely on a generic Gaussian theory for a large family of random forest algorithms. To our knowledge, this is the first set of results that allows any type of random forest, including classification and regression forests, to be used for provably valid statistical inference. In experiments, we find causal forests to be substantially more powerful than classical methods based on nearest-neighbor matching, especially in the presence of irrelevant covariates.

ARTICLE HISTORY
Received December 2015
Revised March 2017

KEYWORDS
Adaptive nearest neighbors matching; Asymptotic normality; Potential outcomes; Unconfoundedness

1. Introduction

In many applications, we want to use data to draw inferences about the causal effect of a treatment: examples include medical studies about the effect of a drug on health outcomes, studies of the impact of advertising or marketing offers on consumer purchases, evaluations of the effectiveness of government programs or public policies, and "A/B tests" (large-scale randomized experiments) commonly used by technology firms to select algorithms for ranking search results or making recommendations. Historically, most datasets have been too small to meaningfully explore heterogeneity of treatment effects beyond dividing the sample into a few subgroups. Recently, however, there has been an explosion of empirical settings where it is potentially feasible to customize estimates for individuals.

An impediment to exploring heterogeneous treatment effects is the fear that researchers will iteratively search for subgroups with high treatment levels, and then report only the results for subgroups with extreme effects, thus highlighting heterogeneity that may be purely spurious (Assmann et al. 2000; Cook, Gebski, and Keech 2004). For this reason, protocols for clinical trials must specify in advance which subgroups will be analyzed, and other disciplines such as economics have instituted protocols for registering pre-analysis plans for randomized experiments or surveys. However, such procedural restrictions can make it difficult to discover strong but unexpected treatment effect heterogeneity. In this article, we seek to address this challenge by developing a powerful, nonparametric method for heterogeneous treatment effect estimation that yields valid asymptotic confidence intervals for the true underlying treatment effect.

Classical approaches to nonparametric estimation of heterogeneous treatment effects include nearest-neighbor matching, kernel methods, and series estimation; see, for example, Crump et al. (2008), Lee (2009), and Willke et al. (2012). These methods perform well in applications with a small number of covariates, but quickly break down as the number of covariates increases. In this article, we explore the use of ideas from the machine learning literature to improve the performance of these classical methods with many covariates. We focus on the family of random forest algorithms introduced by Breiman (2001a), which allow for flexible modeling of interactions in high dimensions by building a large number of regression trees and averaging their predictions. Random forests are related to kernels and nearest-neighbor methods in that they make predictions using a weighted average of "nearby" observations; however, random forests differ in that they have a data-driven way to determine which nearby observations receive more weight, something that is especially important in environments with many covariates or complex interactions among covariates.

Despite their widespread success at prediction and classification, there are important hurdles that need to be cleared before random forests are directly useful to causal inference. Ideally, an estimator should be consistent with a well-understood asymptotic sampling distribution, so that a researcher can use it to test hypotheses and establish confidence intervals. For example, when deciding to use a drug for an individual, we may wish to test the hypothesis that the expected benefit from the treatment is less than the treatment cost. Asymptotic normality results are especially important in the causal inference setting, both because many policy applications require confidence intervals for decision-making, and because it can be difficult to directly evaluate the model's performance using, for example, cross-validation, when estimating causal effects. Yet, the asymptotics

CONTACT Stefan Wager *swager@stanford.edu* Stanford Graduate School of Business, 655 Knight Way, Stanford, CA 94305.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.

of random forests have been largely left open, even in the standard regression or classification contexts.

This article addresses these limitations, developing a forest-based method for treatment effect estimation that allows for a tractable asymptotic theory and valid statistical inference. Following Athey and Imbens (2016), our proposed forest is composed of *causal trees* that estimate the effect of the treatment at the leaves of the trees; we thus refer to our algorithm as a *causal forest*.

In the interest of generality, we begin our theoretical analysis by developing the desired consistency and asymptotic normality results in the context of regression forests. We prove these results for a particular variant of regression forests that uses subsampling to generate a variety of different trees, while it relies on deeply grown trees that satisfy a condition we call "honesty" to reduce bias. An example of an honest tree is one where the tree is grown using one subsample, while the predictions at the leaves of the tree are estimated using a different subsample. We also show that the heuristically motivated infinitesimal jackknife for random forests developed by Efron (2014) and Wager, Hastie, and Efron (2014) is consistent for the asymptotic variance of random forests in this setting. Our proof builds on classical ideas from Efron and Stein (1981), Hájek (1968), and Hoeffding (1948), as well as the adaptive nearest neighbors interpretation of random forests of Lin and Jeon (2006). Given these general results, we next show that our consistency and asymptotic normality results extend from the regression setting to estimating heterogeneous treatment effects in the potential outcomes framework with unconfoundedness (Neyman 1923; Rubin 1974).

Although our main focus in this article is causal inference, we note that there are a variety of important applications of the asymptotic normality result in a pure prediction context. For example, Kleinberg et al. (2015) sought to improve the allocation of Medicare funding for hip or knee replacement surgery by detecting patients who had been prescribed such a surgery, but were in fact likely to die of other causes before the surgery would have been useful to them. Here, we need predictions for the probability that a given patient will survive for more than, say, one year that come with rigorous confidence statements; our results are the first that enable the use of random forests for this purpose.

Finally, we compare the performance of the causal forest algorithm against classical k-nearest neighbor matching using simulations, finding that the causal forest dominates in terms of both bias and variance in a variety of settings, and that its advantage increases with the number of covariates. We also examine coverage rates of our confidence intervals for heterogeneous treatment effects.

1.1. Related Work

There has been a longstanding understanding in the machine learning literature that prediction methods such as random forests ought to be validated empirically (Breiman 2001b): if the goal is prediction, then we should hold out a test set, and the method will be considered as good as its error rate is on this test set. However, there are fundamental challenges with applying a test set approach in the setting of causal inference. In the widely used potential outcomes framework we use to formalize our results (Neyman 1923; Rubin 1974), a treatment effect is understood as a difference between two potential outcomes, for example, would the patient have died if they received the drug versus if they did not receive it. Only one of these potential outcomes can ever be observed in practice, and so direct test-set evaluation is in general impossible.[1] Thus, when evaluating estimators of causal effects, asymptotic theory plays a much more important role than in the standard prediction context.

From a technical point of view, the main contribution of this article is an asymptotic normality theory enabling us to do statistical inference using random forest predictions. Recent results by Biau (2012), Meinshausen (2006), Mentch and Hooker (2016), Scornet, Biau, and Vert (2015), and others have established asymptotic properties of particular variants and simplifications of the random forest algorithm. To our knowledge, however, we provide the first set of conditions under which predictions made by random forests are both asymptotically unbiased and Gaussian, thus allowing for classical statistical inference; the extension to the causal forests proposed in this article is also new. We review the existing theoretical literature on random forests in more detail in Section 3.1.

A small but growing literature, including Green and Kern (2012), Hill (2011), and Hill and Su (2013), has considered the use of forest-based algorithms for estimating heterogeneous treatment effects. These articles use the Bayesian additive regression tree (BART) method of Chipman, George, and McCulloch (2010), and report posterior credible intervals obtained by Markov chain Monte Carlo (MCMC) sampling based on a convenience prior. Meanwhile, Foster, Taylor, and Ruberg (2011) used regression forests to estimate the effect of covariates on outcomes in treated and control groups separately, and then take the difference in predictions as data and project treatment effects onto units' attributes using regression or classification trees (in contrast, we modify the standard random forest algorithm to focus on directly estimating heterogeneity in causal effects). A limitation of this line of work is that, until now, it has lacked formal statistical inference results.

We view our contribution as complementary to this literature, by showing that forest-based methods need not only be viewed as black-box heuristics, and can instead be used for rigorous asymptotic analysis. We believe that the theoretical tools developed here will be useful beyond the specific class of algorithms studied in our article. In particular, our tools allow for a fairly direct analysis of variants of the method of Foster, Taylor, and Ruberg (2011). Using BART for rigorous statistical analysis may prove more challenging since, although BART is often successful in practice, there are currently no results guaranteeing posterior concentration around the true conditional mean function, or convergence of the MCMC sampler in polynomial time. Advances of this type would be of considerable interest.

Several papers use tree-based methods for estimating heterogeneous treatment effects. In growing trees to build our forest, we follow most closely the approach of Athey and Imbens (2016), who propose honest, causal trees, and obtain valid

[1] Athey and Imbens (2016) proposed indirect approaches to mimic test-set evaluation for causal inference. However, these approaches require an estimate of the true treatment effects and/or treatment propensities for all the observations in the test set, which creates a new set of challenges. In the absence of an observable ground truth in a test set, statistical theory plays a more central role in evaluating the noise in estimates of causal effects.

confidence intervals for average treatment effects for each of the subpopulations (leaves) identified by the algorithm. (Instead of personalizing predictions for each individual, this approach only provides treatment effect estimates for leaf-wise subgroups whose size must grow to infinity.) Other related approaches include those of Su et al. (2009) and Zeileis, Hothorn, and Hornik (2008), which build a tree for treatment effects in subgroups and use statistical tests to determine splits; however, these papers do not analyze bias or consistency properties.

Finally, we note a growing literature on estimating heterogeneous treatment effects using different machine learning methods. Imai and Ratkovic (2013), Signorovitch (2007), Tian et al. (2014), and Weisberg and Pontes (2015) developed lasso-like methods for causal inference in a sparse high-dimensional linear setting. Beygelzimer and Langford (2009), Dudík, Langford, and Li (2011), and others discuss procedures for transforming outcomes that enable off-the-shelf loss minimization methods to be used for optimal treatment policy estimation. In the econometrics literature, Bhattacharya and Dupas (2012), Dehejia (2005), Hirano and Porter (2009), and Manski (2004) estimate parametric or semiparametric models for optimal policies, relying on regularization for covariate selection in the case of Bhattacharya and Dupas (2012). Taddy et al. (2016) used Bayesian nonparametric methods with Dirichlet priors to flexibly estimate the data-generating process, and then project the estimates of heterogeneous treatment effects down onto the feature space using regularization methods or regression trees to get low-dimensional summaries of the heterogeneity; but again, there are no guarantees about asymptotic properties.

2. Causal Forests

2.1. Treatment Estimation with Unconfoundedness

Suppose we have access to n independent and identically distributed training examples labeled $i = 1, \ldots, n$, each of which consists of a feature vector $X_i \in [0, 1]^d$, a response $Y_i \in \mathbb{R}$, and a treatment indicator $W_i \in \{0, 1\}$. Following the potential outcomes framework of Neyman (1923) and Rubin (1974) (see Imbens and Rubin 2015 for a review), we then posit the existence of potential outcomes $Y_i^{(1)}$ and $Y_i^{(0)}$ corresponding respectively to the response the ith subject would have experienced with and without the treatment, and define the treatment effect at x as

$$\tau(x) = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)} \,\middle|\, X_i = x\right]. \tag{1}$$

Our goal is to estimate this function $\tau(x)$. The main difficulty is that we can only ever observe one of the two potential outcomes $Y_i^{(0)}$, $Y_i^{(1)}$ for a given training example, and so cannot directly train machine learning methods on differences of the form $Y_i^{(1)} - Y_i^{(0)}$.

In general, we cannot estimate $\tau(x)$ simply from the observed data (X_i, Y_i, W_i) without further restrictions on the data-generating distribution. A standard way to make progress is to assume unconfoundedness (Rosenbaum and Rubin 1983), that is, that the treatment assignment W_i is independent of the potential outcomes for Y_i conditional on X_i:

$$\left\{Y_i^{(0)}, Y_i^{(1)}\right\} \perp\!\!\!\perp W_i \mid X_i. \tag{2}$$

The motivation behind this unconfoundedness is that, given continuity assumptions, it effectively implies that we can treat nearby observations in x-space as having come from a randomized experiment; thus, nearest-neighbor matching and other local methods will in general be consistent for $\tau(x)$.

An immediate consequence of unconfoundedness is that

$$\mathbb{E}\left[Y_i\left(\frac{W_i}{e(x)} - \frac{1 - W_i}{1 - e(x)}\right) \,\middle|\, X_i = x\right] = \tau(x), \quad \text{where}$$
$$e(x) = \mathbb{E}[W_i \mid X_i = x] \tag{3}$$

is the propensity of receiving treatment at x. Thus, if we knew $e(x)$, we would have access to a simple unbiased estimator for $\tau(x)$; this observation lies at the heart of methods based on propensity weighting (e.g., Hirano, Imbens, and Ridder 2003). Many early applications of machine learning to causal inference effectively reduce to estimating $e(x)$ using, for example, boosting, a neural network, or even random forests, and then transforming this into an estimate for $\tau(x)$ using (3) (e.g., McCaffrey, Ridgeway, and Morral 2004; Westreich, Lessler, and Funk 2010). In this article, we take a more indirect approach: we show that, under regularity assumptions, causal forests can use the unconfoundedness assumption (2) to achieve consistency without needing to explicitly estimate the propensity $e(x)$.[2]

2.2. From Regression Trees to Causal Trees and Forests

At a high level, trees and forests can be thought of as nearest neighbor methods with an adaptive neighborhood metric. Given a test point x, classical methods such as k-nearest neighbors seek the k closest points to x according to some prespecified distance measure, for example, Euclidean distance. In contrast, tree-based methods also seek to find training examples that are close to x, but now closeness is defined with respect to a decision tree, and the closest points to x are those that fall in the same leaf as it. The advantage of trees is that their leaves can be narrower along the directions where the signal is changing fast and wider along the other directions, potentially leading a to a substantial increase in power when the dimension of the feature space is even moderately large.

In this section, we seek to build causal trees that resemble their regression analogues as closely as possible. Suppose first that we only observe independent samples (X_i, Y_i), and want to build a CART regression tree. We start by recursively splitting the feature space until we have partitioned it into a set of leaves L, each of which only contains a few training samples. Then, given a test point x, we evaluate the prediction $\hat{\mu}(x)$ by identifying the leaf $L(x)$ containing x and setting

$$\hat{\mu}(x) = \frac{1}{|\{i : X_i \in L(x)\}|} \sum_{\{i : X_i \in L(x)\}} Y_i. \tag{4}$$

Heuristically, this strategy is well-motivated if we believe the leaf $L(x)$ to be small enough that the responses Y_i inside the leaf are roughly identically distributed. There are several procedures for how to place the splits in the decision tree; see, for example, Hastie, Tibshirani, and Friedman (2009).

[2] In follow-up work, Athey, Tibshirani, and Wager (2018) adapted the causal forest algorithm, enabling it to make use of propensity score estimates $\hat{e}(x)$ for improved robustness.

In the context of causal trees, we analogously want to think of the leaves as small enough that the $(Y_i,\ W_i)$ pairs corresponding to the indices i for which $i \in L(x)$ act as though they had come from a randomized experiment. Then, it is natural to estimate the treatment effect for any $x \in L$ as

$$\hat{\tau}(x) = \frac{1}{|\{i : W_i = 1,\ X_i \in L\}|} \sum_{\{i:W_i=1,\ X_i \in L\}}^{Y_i} - \frac{1}{|\{i : W_i = 0,\ X_i \in L\}|} \sum_{\{i:W_i=0,\ X_i \in L\}}^{Y_i.} . \quad (5)$$

In the following sections, we will establish that such trees can be used to grow causal forests that are consistent for $\tau(x)$.[3]

Finally, given a procedure for generating a single causal tree, a causal forest generates an ensemble of B such trees, each of which outputs an estimate $\hat{\tau}_b(x)$. The forest then aggregates their predictions by averaging them: $\hat{\tau}(x) = B^{-1} \sum_{b=1}^{B} \hat{\tau}_b(x)$. We always assume that the individual causal trees in the forest are built using random subsamples of s training examples, where $s/n \ll 1$; for our theoretical results, we will assume that $s \asymp n^{\beta}$ for some $\beta < 1$. The advantage of a forest over a single tree is that it is not always clear what the "best" causal tree is. In this case, as shown by Breiman (2001a), it is often better to generate many different decent-looking trees and average their predictions, instead of seeking a single highly-optimized tree. In practice, this aggregation scheme helps reduce variance and smooths sharp decision boundaries (Bühlmann and Yu 2002).

2.3. Asymptotic Inference with Causal Forests

Our results require some conditions on the forest-growing scheme: the trees used to build the forest must be grown on subsamples of the training data, and the splitting rule must not "inappropriately" incorporate information about the outcomes Y_i as discussed formally in Section 2.4. However, given these high level conditions, we obtain a widely applicable consistency result that applies to several different interesting causal forest algorithms.

Our first result is that causal forests are consistent for the true treatment effect $\tau(x)$. To achieve pointwise consistency, we need to assume that the conditional mean functions $\mathbb{E}[Y^{(0)} \,|\, X = x]$ and $\mathbb{E}[Y^{(1)} \,|\, X = x]$ are both Lipschitz continuous. To our knowledge, all existing results on pointwise consistency of regression forests (e.g., Biau 2012; Meinshausen 2006) require an analogous condition on $\mathbb{E}[Y \,|\, X = x]$. This is not particularly surprising, as forests generally have smooth response surfaces (Bühlmann and Yu 2002). In addition to continuity assumptions, we also need to assume that we have overlap, that is, for some $\varepsilon > 0$ and all $x \in [0,\ 1]^d$,

$$\varepsilon < \mathbb{P}[W = 1 \,|\, X = x] < 1 - \varepsilon. \quad (6)$$

This condition effectively guarantees that, for large enough n, there will be enough treatment and control units near any test point x for local methods to work.

Beyond consistency, to do statistical inference on the basis of the estimated treatment effects $\hat{\tau}(x)$, we need to understand their asymptotic sampling distribution. Using the potential nearest neighbors construction of Lin and Jeon (2006) and classical analysis tools going back to Hoeffding (1948) and Hájek (1968), we show that—provided the subsample size s scales appropriately with n—the predictions made by a causal forest are asymptotically Gaussian and unbiased. Specifically, we show that

$$(\hat{\tau}(x) - \tau(x)) \,/\, \sqrt{\mathrm{Var}[\hat{\tau}(x)]} \Rightarrow \mathcal{N}(0,\ 1) \quad (7)$$

under the conditions required for consistency, provided the subsample size s scales as $s \asymp n^{\beta}$ for some $\beta_{\min} < \beta < 1$

Moreover, we show that the asymptotic variance of causal forests can be accurately estimated. To do so, we use the infinitesimal jackknife for random forests developed by Efron (2014) and Wager, Hastie, and Efron (2014), based on the original infinitesimal jackknife procedure of Jaeckel (1972). This method assumes that we have taken the number of trees B to be large enough that the Monte Carlo variability of the forest does not matter; and only measures the randomness in $\hat{\tau}(x)$ due to the training sample.

To define the variance estimates, let $\hat{\tau}_b^*(x)$ be the treatment effect estimate given by the bth tree, and let $N_{ib}^* \in \{0,\ 1\}$ indicate whether or not the ith training example was used for the bth tree.[4] Then, we set

$$\widehat{V}_{IJ}(x) = \frac{n-1}{n} \left(\frac{n}{n-s}\right)^2 \sum_{i=1}^{n} \mathrm{Cov}_*[\hat{\tau}_b^*(x),\ N_{ib}^*]^2, \quad (8)$$

where the covariance is taken with respect to the set of all the trees $b = 1, \ldots, B$ used in the forest. The term $n(n-1)/(n-s)^2$ is a finite-sample correction for forests grown by subsampling without replacement; see Proposition 5. We show that this variance estimate is consistent, in the sense that $\widehat{V}_{IJ}(x) / \mathrm{Var}[\hat{\tau}(x)] \to_p 1$.

2.4. Honest Trees and Forests

In our discussion so far, we have emphasized the flexible nature of our results: for a wide variety of causal forests that can be tailored to the application area, we achieve both consistency and centered asymptotic normality, provided the subsample size s scales at an appropriate rate. Our results do, however, require the individual trees to satisfy a fairly strong condition, which we call honesty: a tree is honest if, for each training example i, it only uses the response Y_i to estimate the within-leaf treatment effect τ using (5) or to decide where to place the splits, but not both. We discuss two causal forest algorithms that satisfy this condition.

Our first algorithm, which we call a double-sample tree, achieves honesty by dividing its training subsample into two halves $\mathcal{I}$ and $\mathcal{J}$. Then, it uses the $\mathcal{J}$-sample to place the splits, while holding out the $\mathcal{I}$-sample to do within-leaf estimation; see

[3] The causal tree algorithm presented above is a simplification of the method of Athey and Imbens (2016). The main difference between our approach and that of Athey and Imbens (2016) is that they seek to build a single well-tuned tree; to this end, they use fairly large leaves and apply a form propensity weighting based on (3) within each leaf to correct for variations in $e(x)$ inside the leaf. In contrast, we follow Breiman (2001a) and build our causal forest using deep trees. Since our leaves are small, we do not need to apply any additional corrections inside them.

[4] For double-sample trees defined in Procedure 1, $N_{ib}^* = 1$ if the ith example appears in either the $\mathcal{I}$-sample or the $\mathcal{J}$-sample.

Procedure 1 for details. In our experiments, we set the minimum leaf size to $k = 1$. A similar family of algorithms was discussed in detail by Denil, Matheson, and De Freitas (2014), who showed that such forests could achieve competitive performance relative to standard tree algorithms that do not divide their training samples. In the semiparametric inference literature, related ideas go back at least to the work of Schick (1986).

We note that sample splitting procedures are sometimes criticized as inefficient because they "waste" half of the training data at each step of the estimation procedure. However, in our case, the forest subsampling mechanism enables us to achieve honesty without wasting any data in this sense, because we rerandomize the $\mathcal{I}/\mathcal{J}$-data splits over each subsample. Thus, although no data point can be used for split selection and leaf estimation in a single tree, each data point will participate in both $\mathcal{I}$ and $\mathcal{J}$ samples of some trees, and so will be used for both specifying the structure and treatment effect estimates of the forest. Although our original motivation for considering double-sample trees was to eliminate bias and thus enable centered confidence intervals, we find that in practice, double-sample trees can improve upon standard random forests in terms of mean-squared error as well.

Procedure 1. Double-Sample Trees

Double-sample trees split the available training data into two parts: one half for estimating the desired response inside each leaf, and another half for placing splits.

Input: n training examples of the form $(X_i,\ Y_i)$ for regression trees or $(X_i,\ Y_i,\ W_i)$ for causal trees, where X_i are features, Y_i is the response, and W_i is the treatment assignment. A minimum leaf size k.

1. Draw a random subsample of size s from $\{1,\ \ldots,\ n\}$ without replacement, and then divide it into two disjoint sets of size $|\mathcal{I}| = \lfloor s/2 \rfloor$ and $|\mathcal{J}| = \lceil s/2 \rceil$.
2. Grow a tree via recursive partitioning. The splits are chosen using any data from the $\mathcal{J}$ sample and X- or W-observations from the $\mathcal{I}$ sample, but without using Y-observations from the $\mathcal{I}$-sample.
3. Estimate leafwise responses using only the $\mathcal{I}$-sample observations.

Double-sample *regression* trees make predictions $\hat{\mu}(x)$ using (4) on the leaf containing x, only using the $\mathcal{I}$-sample observations. The splitting criteria is the standard for CART regression trees (minimizing mean-squared error of predictions). Splits are restricted so that each leaf of the tree must contain k or more $\mathcal{I}$-sample observations.

Double-sample *causal* trees are defined similarly, except that for prediction we estimate $\hat{\tau}(x)$ using (5) on the $\mathcal{I}$ sample. Following Athey and Imbens (2016), the splits of the tree are chosen by maximizing the variance of $\hat{\tau}(X_i)$ for $i \in \mathcal{J}$; see Remark 1 for details. In addition, each leaf of the tree must contain k or more $\mathcal{I}$-sample observations of *each* treatment class.

Another way to build honest trees is to ignore the outcome data Y_i when placing splits, and instead first train a classification tree for the treatment assignments W_i (Procedure 2). Such propensity trees can be particularly useful in observational studies, where we want to minimize bias due to variation in $e(x)$. Seeking estimators that match training examples based on estimated propensity is a longstanding idea in causal inference, going back to Rosenbaum and Rubin (1983).[5]

Procedure 2. Propensity Trees

Propensity trees use only the treatment assignment indicator W_i to place splits, and save the responses Y_i for estimating τ.

Input: n training examples $(X_i,\ Y_i,\ W_i)$, where X_i are features, Y_i is the response, and W_i is the treatment assignment. A minimum leaf size k.

1. Draw a random subsample $\mathcal{I} \in \{1,\ \ldots,\ n\}$ of size $|\mathcal{I}| = s$ (no replacement).
2. Train a classification tree using sample $\mathcal{I}$ where the outcome is the treatment assignment, that is, on the $(X_i,\ W_i)$ pairs with $i \in \mathcal{I}$. Each leaf of the tree must have k or more observations of *each* treatment class.
3. Estimate $\tau(x)$ using (5) on the leaf containing x.

In step 2, the splits are chosen by optimizing, for example, the Gini criterion used by CART for classification (Breiman et al. 1984).

Remark 1. For completeness, we briefly outline the motivation for the splitting rule of Athey and Imbens (2016) we use for our double-sample trees. This method is motivated by an algorithm for minimizing the squared-error loss in regression trees. Because regression trees compute predictions $\hat{\mu}$ by averaging training responses over leaves, we can verify that

$$\sum_{i \in \mathcal{J}} (\hat{\mu}(X_i) - Y_i)^2 = \sum_{i \in \mathcal{J}} Y_i^2 - \sum_{i \in \mathcal{J}} \hat{\mu}(X_i)^2. \tag{9}$$

Thus, finding the squared-error minimizing split is equivalent to maximizing the variance of $\hat{\mu}(X_i)$ for $i \in \mathcal{J}$; note that $\sum_{i \in \mathcal{J}} \hat{\mu}(X_i) = \sum_{i \in \mathcal{J}} Y_i$ for all trees, and so maximizing variance is equivalent to maximizing the sum of the $\hat{\mu}(X_i)^2$. In Procedure 1, we emulate this algorithm by picking splits that maximize the variance of $\hat{\tau}(X_i)$ for $i \in \mathcal{J}$.[6]

Remark 2. In Appendix B, we present evidence that adaptive forests with small leaves can overfit to outliers in ways that make them inconsistent near the edges of sample space. Thus, the forests of Breiman (2001a) need to be modified in some way to get pointwise consistency results; here, we use honesty following, for example, Wasserman and Roeder (2009). We note that there have been some recent theoretical investigations of non-honest forests, including Scornet, Biau, and Vert (2015) and Wager and Walther (2015). However, Scornet, Biau, and Vert (2015) do not consider pointwise properties of forests;

[5] While this article was in press, we became aware of work by Wang et al. (2015), who use what we call propensity forests for average treatment effect estimation.

[6] Athey and Imbens (2016) also considered "honest splitting rules" that anticipate honest estimation, and correct for the additional sampling variance in small leaves using an idea closely related to the C_p penalty of Mallows (1973). Although it could be of interest for further work, we do not study the effect of such splitting rules here.

whereas Wager and Walther (2015) showed consistency of adaptive forests with larger leaves, but their bias bounds decay slower than the sampling variance of the forests and so cannot be used to establish centered asymptotic normality.

3. Asymptotic Theory for Random Forests

To use random forests to provide formally valid statistical inference, we need an asymptotic normality theory for random forests. In the interest of generality, we first develop such a theory in the context of classical regression forests, as originally introduced by Breiman (2001a). In this section, we assume that we have training examples $Z_i = (X_i, Y_i)$ for $i = 1, \ldots, n$, a test point x, and we want to estimate true conditional mean function

$$\mu(x) = \mathbb{E}[Y \mid X = x]. \tag{10}$$

We also have access to a regression tree T which can be used to get estimates of the conditional mean function at x of the form $T(x; \xi, Z_1, \ldots, Z_n)$, where $\xi \sim \Xi$ is a source of auxiliary randomness. Our goal is to use this tree-growing scheme to build a random forest that can be used for valid statistical inference about $\mu(x)$.

We begin by precisely describing how we aggregate individual trees into a forest. For us, a random forest is an average of trees trained over all possible size-s subsamples of the training data, marginalizing over the auxiliary noise ξ. In practice, we compute such a random forest by Monte Carlo averaging, and set

$$\mathrm{RF}(x; Z_1, \ldots, Z_n) \approx \frac{1}{B} \sum_{b=1}^{B} T(x; \xi_b^*, Z_{b1}^*, \ldots, Z_{bs}^*), \tag{11}$$

where $\{Z_{b1}^*, \ldots, Z_{bs}^*\}$ is drawn without replacement from $\{Z_1, \ldots, Z_n\}$, ξ_b^* is a random draw from Ξ, and B is the number of Monte Carlo replicates we can afford to perform. The formulation (12) arises as the $B \to \infty$ limit of (11); thus, our theory effectively assumes that B is large enough for Monte Carlo effects not to matter. The effects of using a finite B are studied in detail by Mentch and Hooker (2016); see also Wager, Hastie, and Efron (2014), who recommend taking B on the order of n.

Definition 1. The *random forest* with base learner T and subsample size s is

$$\mathrm{RF}(x; Z_1, \ldots, Z_n) = \binom{n}{s}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_s \le n} \mathbb{E}_{\xi \sim \Xi} \times \left[T(x; \xi, Z_{i_1}, \cdots, Z_{i_s}) \right]. \tag{12}$$

Next, as described in Section 2, we require that the trees T in our forest be honest. Double-sample trees, as defined in Procedure 1, can always be used to build honest trees with respect to the $\mathcal{I}$-sample. In the context of causal trees for observational studies, propensity trees (Procedure 2) provide a simple recipe for building honest trees without sample splitting.

Definition 2. A tree grown on a training sample $(Z_1 = (X_1, Y_1), \ldots, Z_s = (X_s, Y_s))$ is *honest* if (a) (*standard case*) the tree does not use the responses $Y_1, \ldots, Y_s$ in choosing where to place its splits; or (b) (*double sample case*) the tree does not use the $\mathcal{I}$-sample responses for placing splits.

To guarantee consistency, we also need to enforce that the leaves of the trees become small in *all* dimensions of the feature space as n gets large.[7] Here, we follow Meinshausen (2006), and achieve this effect by enforcing some randomness in the way trees choose the variables they split on: at each step, each variable is selected with probability at least π/d for some $0 < \pi \le 1$ (e.g., we could satisfy this condition by completely randomizing the splitting variable with probability π). Formally, the randomness in how to pick the splitting features is contained in the auxiliary random variable ξ.

Definition 3. A tree is a *random-split* tree if at every step of the tree-growing procedure, marginalizing over ξ, the probability that the next split occurs along the jth feature is bounded below by π/d for some $0 < \pi \le 1$, for all $j = 1, \ldots, d$.

The remaining definitions are more technical. We use regularity to control the shape of the tree leaves, while symmetry is used to apply classical tools in establishing asymptotic normality.

Definition 4. A tree predictor grown by recursive partitioning is *α-regular* for some $\alpha > 0$ if either (a) (*standard case*) each split leaves at least a fraction α of the available training examples on each side of the split and, moreover, the trees are fully grown to depth k for some $k \in \mathbb{N}$, that is, there are between k and $2k - 1$ observations in each terminal node of the tree; or (b) (*double sample case*) if the predictor is a double-sample tree as in Procedure 1, the tree satisfies part (a) for the $\mathcal{I}$ sample.

Definition 5. A predictor is *symmetric* if the (possibly randomized) output of the predictor does not depend on the order ($i = 1, 2, \ldots$) in which the training examples are indexed.

Finally, in the context of classification and regression forests, we estimate the asymptotic variance of random forests using the original infinitesimal jackknife of Wager, Hastie, and Efron (2014), that is,

$$\widehat{V}_{IJ}(x) = \frac{n-1}{n} \left(\frac{n}{n-s} \right)^2 \sum_{i=1}^{n} \mathrm{cov}_*[\hat{\mu}_b^*(x), N_{ib}^*]^2, \tag{13}$$

where $\hat{\mu}_b^*(x)$ is the estimate for $\mu(x)$ given by a single regression tree. We note that the finite-sample correction $n(n-1)/(n-s)^2$ did not appear in Wager, Hastie, and Efron (2014), as their article focused on subsampling with replacement, whereas this correction is only appropriate for subsampling without replacement.

Given these preliminaries, we can state our main result on the asymptotic normality of random forests. As discussed in Section 2.3, we require that the conditional mean function $\mu(x) = \mathbb{E}[Y \mid X = x]$ be Lipschitz continuous. The asymptotic normality result requires for the subsample size s to scale within the bounds given in (14). If the subsample size grows slower than

[7] Biau (2012) and Wager and Walther (2015) considered the estimation of low-dimensional signals embedded in a high-dimensional ambient space using random forests; in this case, the variable selection properties of trees also become important. We leave a study of asymptotic normality of random forests in high dimensions to future work.

this, the forest will still be asymptotically normal, but the forest may be asymptotically biased. For clarity, we state the following result with notation that makes the dependence of $\hat{\mu}_n(x)$ and s_n on n explicit; in most of the article, however, we drop the subscripts to $\hat{\mu}_n(x)$ and s_n when there is no risk of confusion.

Theorem 3.1. Suppose that we have n independent and identically distributed training examples $Z_i = (X_i,\ Y_i) \in [0,\ 1]^d \times \mathbb{R}$. Suppose moreover that the features are independently and uniformly distributed[8] $X_i \sim U([0,\ 1]^d)$, that $\mu(x) = \mathbb{E}[Y \,|\, X = x]$ and $\mu_2(x) = \mathbb{E}[Y^2 \,|\, X = x]$ are Lipschitz-continuous, and finally that $\mathrm{Var}[Y \,|\, X = x] > 0$ and $\mathbb{E}[|Y - \mathbb{E}[Y \,|\, X = x]|^{2+\delta} \,|\, X = x] \leq M$ for some constants $\delta,\ M > 0$, uniformly over all $x \in [0,\ 1]^d$. Given this data-generating process, let T be an honest, α-regular with $\alpha \leq 0.2$, and symmetric random-split tree in the sense of Definitions 2–5, and let $\hat{\mu}_n(x)$ be the estimate for $\mu(x)$ given by a random forest with base learner T and a subsample size s_n. Finally, suppose that the subsample size s_n scales as

$$s_n \asymp n^{\beta} \text{ for some } \beta_{\min} := 1 - \left(1 + \frac{d}{\pi} \frac{\log(\alpha^{-1})}{\log((1-\alpha)^{-1})}\right)^{-1} < \beta < 1. \tag{14}$$

Then, random forest predictions are asymptotically Gaussian:

$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \Rightarrow \mathcal{N}(0,\ 1) \quad \text{for a sequence } \sigma_n(x) \to 0. \tag{15}$$

Moreover, the asymptotic variance σ_n can be consistently estimated using the infinitesimal jackknife (8):

$$\widehat{V}_{IJ}(x) \big/ \sigma_n^2(x) \to_p 1. \tag{16}$$

Remark 3 (binary classification). We note that Theorem 3.1 also holds for binary classification forests with leaf size $k = 1$, as is default in the R package `randomForest` (Liaw and Wiener 2002). Here, we treat the output $\mathrm{RF}(x)$ of the random forests as an estimate for the probability $\mathbb{P}[Y = 1 \,|\, X = x]$; Theorem 3.1 then lets us construct valid confidence intervals for this probability. For classification forests with $k > 1$, the proof of Theorem 3.1 still holds if the individual classification trees are built by *averaging* observations within a leaf, but not if they are built by *voting*. Extending our results to voting trees is left as further work.

The proof of this result is organized as follows. In Section 3.2, we provide bounds for the bias $\mathbb{E}[\hat{\mu}_n(x) - \mu(x)]$ of random forests, while Section 3.3 studies the sampling distributions of $\hat{\mu}_n(x) - \mathbb{E}[\hat{\mu}_n(x)]$ and establishes Gaussianity. Given a subsampling rate satisfying (14), the bias decays faster than the variance, thus allowing for (15). Before beginning the proof, however, we relate our result to existing results about random forests in Section 3.1.

3.1. Theoretical Background

There has been considerable work in understanding the theoretical properties of random forests. The convergence and consistency properties of trees and random forests have been studied by, among others, Biau (2012), Biau, Devroye, and Lugosi (2008), Breiman (2004), Breiman et al. (1984), Meinshausen (2006), Scornet, Biau, and Vert (2015), Wager and Walther (2015), and Zhu, Zeng, and Kosorok (2015). Meanwhile, their sampling variability has been analyzed by Duan (2011), Lin and Jeon (2006), Mentch and Hooker (2016), Sexton and Laake (2009), and Wager, Hastie, and Efron (2014). However, to our knowledge, our Theorem 3.1 is the first result establishing conditions under which predictions made by random forests are asymptotically unbiased and normal.

Probably the closest existing result is that of Mentch and Hooker (2016), who showed that random forests based on subsampling are asymptotically normal under substantially strong conditions than us: they require that the subsample size s grows slower than $\sqrt{n}$, that is, that $s_n/\sqrt{n} \to 0$. However, under these conditions, random forests will not in general be asymptotically unbiased. As a simple example, suppose that $d = 2$, that $\mu(x) = \|x\|_1$, and that we evaluate an honest random forest at $x = 0$. A quick calculation shows that the bias of the random forest decays as $1/\sqrt{s_n}$, while its variance decays as s_n/n. If $s_n/\sqrt{n} \to 0$, the squared bias decays slower than the variance, and so confidence intervals built using the resulting Gaussian limit distribution will not cover $\mu(x)$. Thus, although the result of Mentch and Hooker (2016) may appear qualitatively similar to ours, it cannot be used for valid asymptotic statistical inference about $\mu(x)$.

The variance estimator $\widehat{V}_{IJ}$ was studied in the context of random forests by Wager, Hastie, and Efron (2014), who showed empirically that the method worked well for many problems of interest. Wager, Hastie, and Efron (2014) also emphasized that, when using $\widehat{V}_{IJ}$ in practice, it is important to account for Monte Carlo bias. Our analysis provides theoretical backing to these results, by showing that $\widehat{V}_{IJ}$ is in fact a consistent estimate for the variance $\sigma_n^2(x)$ of random forest predictions. The earlier work on this topic (Efron 2014; Wager, Hastie, and Efron 2014) had only motivated the estimator $\widehat{V}_{IJ}$ by highlighting connections to classical statistical ideas, but did not establish any formal justification for it.

Instead of using subsampling, Breiman originally described random forests in terms of bootstrap sampling, or bagging (Breiman 1996). Random forests with bagging, however, have proven to be remarkably resistant to classical statistical analysis. As observed by Buja and Stuetzle (2006), Chen and Hall (2003), Friedman and Hall (2007) and others, estimators of this form can exhibit surprising properties even in simple situations; meanwhile, using subsampling rather than bootstrap sampling has been found to avoid several pitfalls (e.g., Politis, Romano, and Wolf 1999). Although they are less common in the literature, random forests based on subsampling have also been occasionally studied and found to have good practical and theoretical properties (e.g., Bühlmann and Yu 2002; Mentch and Hooker 2016; Scornet, Biau, and Vert 2015; Strobl et al. 2007).

Finally, an interesting question for further theoretical study is to understand the optimal scaling of the subsample size s_n for minimizing the mean-squared error of random forests. For subsampled nearest-neighbors estimation, the optimal rate for s_n is $s_n \asymp n^{1-(1+d/4)^{-1}}$ (Biau, Cérou, and Guyader 2010; Samworth 2012). Here, our specific value for $\beta_{\min}$ depends on the upper bounds for bias developed in the following section. Now, as shown by Biau (2012), under some sparsity assumptions

[8] The result also holds with a density that is bounded away from 0 and infinity; however, we assume uniformity for simpler exposition.

on $\mu(x)$, it is possible to get substantially stronger bounds for the bias of random forests; thus, it is plausible that under similar conditions we could push back the lower bound $\beta_{\min}$ on the growth rate of the subsample size.

3.2. Bias and Honesty

We start by bounding the bias of regression trees. Our approach relies on showing that as the sample size s available to the tree gets large, its leaves get small; Lipschitz-continuity of the conditional mean function and honesty then let us bound the bias. To state a formal result, define the *diameter* $\text{diam}(L(x))$ of a leaf $L(x)$ as the length of the longest segment contained inside $L(x)$, and similarly let $\text{diam}_j(L(x))$ denote the length of the longest such segment that is parallel to the jth axis. The following lemma is a refinement of a result of Meinshausen (2006), who showed that $\text{diam}(L(x)) \to_p 0$ for regular trees.

Lemma 1. Let T be a regular, random-split tree and let $L(x)$ denote its leaf containing x. Suppose that $X_1, \ldots, X_s \sim U([0, 1]^d)$ independently. Then, for any $0 < \eta < 1$, and for large enough s,

$$\mathbb{P}\left[\text{diam}_j\left(L(x)\right) \geq \left(\frac{s}{2k-1}\right)^{-\frac{0.99\,(1-\eta)\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}\right] \leq \left(\frac{s}{2k-1}\right)^{-\frac{\eta^2}{2}\frac{1}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}.$$

This lemma then directly translates into a bound on the bias of a single regression tree. Since a forest is an average of independently-generated trees, the bias of the forest is the same as the bias of a single tree.

Theorem 3.2. Under the conditions of Lemma 1, suppose moreover that $\mu(x)$ is Lipschitz continuous and that the trees T in the random forest are honest. Then, provided that $\alpha \leq 0.2$, the bias of the random forest at x is bounded by

$$\left|\mathbb{E}[\hat{\mu}(x)] - \mu(x)\right| = \mathcal{O}\left(s^{-\frac{1}{2}\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{d}}\right);$$

the constant in the $\mathcal{O}$-bound is given in the proof.

3.3. Asymptotic Normality of Random Forests

Our analysis of the asymptotic normality of random forests builds on ideas developed by Hoeffding (1948) and Hájek (1968) for understanding classical statistical estimators such as U-statistics. We begin by briefly reviewing their results to give some context to our proof. Given a predictor T and independent training examples $Z_1, \ldots, Z_n$, the Hájek projection of T is defined as

$$\mathring{T} = \mathbb{E}[T] + \sum_{i=1}^{n}\left(\mathbb{E}[T \mid Z_i] - \mathbb{E}[T]\right). \tag{17}$$

In other words, the Hájek projection of T captures the first-order effects in T. Classical results imply that $\text{var}[\mathring{T}] \leq \text{var}[T]$, and further:

$$\lim_{n\to\infty} \text{var}[\mathring{T}]/\text{var}[T] = 1 \text{ implies that } \lim_{n\to\infty} \mathbb{E} \times \left[\left\|\mathring{T} - T\right\|_2^2\right]/\text{var}[T] = 0. \tag{18}$$

Since the Hájek projection $\mathring{T}$ is a sum of independent random variables, we should expect it to be asymptotically normal under weak conditions. Thus, whenever the ratio of the variance of $\mathring{T}$ to that of T tends to 1, the theory of Hájek projections almost automatically guarantees that T will be asymptotically normal.[9]

If T is a regression tree, however, the condition from (18) does not apply, and we cannot use the classical theory of Hájek projections directly. Our analysis is centered around a weaker form of this condition, which we call ν-incrementality. With our definition, predictors T to which we can apply the argument (18) directly are 1-incremental.

Definition 6. The predictor T is $\nu(s)$-incremental at x if

$$\text{var}[\mathring{T}(x;\, Z_1,\, \ldots,\, Z_s)]/\text{var}[T(x;\, Z_1,\, \ldots,\, Z_s)] \gtrsim \nu(s),$$

where $\mathring{T}$ is the Hájek projection of T (17). In our notation,

$$f(s) \gtrsim g(s) \text{ means that } \liminf_{s\to\infty} f(s)/g(s) \geq 1.$$

Our argument proceeds in two steps. First, we establish lower bounds for the incrementality of regression trees in Section 3.3.1. Then, in Section 3.3.2 we show how we can turn weakly incremental predictors T into 1-incremental ensembles by subsampling (Lemma 4), thus bringing us back into the realm of classical theory. We also establish the consistency of the infinitesimal jackknife for random forests. Our analysis of regression trees is motivated by the "potential nearest neighbors" model for random forests introduced by Lin and Jeon (2006); the key technical device used in Section 3.3.2 is the ANOVA decomposition of Efron and Stein (1981). The discussion of the infinitesimal jackknife for random forest builds on results of Efron (2014) and Wager, Hastie, and Efron (2014).

3.3.1. Regression Trees and Incremental Predictors

Analyzing specific greedy tree models such as CART trees can be challenging. We thus follow the lead of Lin and Jeon (2006), and analyze a more general class of predictors—potential nearest neighbors predictors—that operate by doing a nearest-neighbor search over rectangles; see also Biau and Devroye (2010). The study of potential (or layered) nearest neighbors goes back at least to Barndorff-Nielsen and Sobel (1966).

Definition 7. Consider a set of points $X_1, \ldots, X_s \in \mathbb{R}^d$ and a fixed $x \in \mathbb{R}^d$. A point X_i is a *potential nearest neighbor* (PNN) of x if the smallest axis-aligned hyperrectangle with vertices x and X_i contains no other points X_j. Extending this notion, a *PNN k-set* of x is a set of points $\Lambda \subseteq \{X_1, \ldots, X_s\}$ of size $k \leq |L| < 2k-1$ such that there exists an axis aligned hyperrectangle L containing x, Λ, and no other training points. A training example X_i is called a k-PNN of x if there exists a PNN k-set of x containing X_i. Finally, a predictor T is a *k-PNN predictor* over $\{Z\}$ if,

[9] The moments defined in (17) depend on the data-generating process for the Z_i, and so cannot be observed in practice. Thus, the Hájek projection is mostly useful as an abstract theoretical tool. For a review of classical projection arguments, see Chapter 11 of Van der Vaart (2000).

given a training set

$$\{Z\} = \{(X_1,\ Y_1),\ \ldots,\ (X_s,\ Y_s)\} \in \{\mathbb{R}^d \times \mathcal{Y}\}^s$$

and a test point $x \in \mathbb{R}^d$, T always outputs the average of the responses Y_i over a k-PNN set of x.

This formalism allows us to describe a wide variety of tree predictors. For example, as shown by Lin and Jeon (2006), any decision tree T that makes axis-aligned splits and has leaves of size between k and $2k-1$ is a k-PNN predictor. In particular, the base learners originally used by Breiman (2001a), namely CART trees grown up to a leaf size k (Breiman et al. 1984), are k-PNN predictors. Predictions made by k-PNN predictors can always be written as

$$T(x;\ \xi,\ Z_1,\ \ldots,\ Z_s) = \sum_{i=1}^{s} S_i Y_i, \tag{19}$$

where S_i is a selection variable that takes the value $1/|\{i : X_i \in L(x)\}|$ for indices i in the selected leaf-set $L(x)$ and 0 for all other indices. If the tree is honest, we know in addition that, for each i, S_i is independent of Y_i conditional on X_i.

An important property of k-PNN predictors is that we can often get a good idea about whether S_i is nonzero even if we only get to see Z_i; more formally, as we show below, the quantity $s\,\mathrm{var}[\mathbb{E}[S_1 \mid Z_1]]$ cannot get too small. Establishing this fact is a key step in showing that k-PNNs are incremental. In the following result, T can be an arbitrary symmetric k-PNN predictor.

Lemma 3.2. Suppose that the observations $X_1,\ X_2,\ \ldots$ are independent and identically distributed on $[0,\ 1]^d$ with a density f that is bounded away from infinity, and let T be any symmetric k-PNN predictor. Then, there is a constant $C_{f,\,d}$ depending only on f and d such that, as s gets large,

$$s\,\mathrm{var}[\mathbb{E}[S_1 \mid Z_1]] \gtrsim \frac{1}{k} C_{f,\,d} / \log(s)^d, \tag{20}$$

where S_i is the indicator for whether the observation is selected in the subsample. When f is uniform over $[0,\ 1]^d$, the bound holds with $C_{f,\,d} = 2^{-(d+1)}(d-1)!$.

When $k = 1$ we see that, marginally, $S_1 \sim \mathrm{Bernoulli}(1/s)$ and so $s\,\mathrm{var}[S_1] \sim 1$; more generally, a similar calculation shows that $1/(2k-1) \lesssim s\,\mathrm{var}[S_1] \lesssim 1/k$. Thus, (20) can be interpreted as a lower bound on how much information Z_1 contains about the selection event S_1.

Thanks to this result, we are now ready to show that all honest and regular random-split trees are incremental. Notice that any symmetric k-regular tree following Definition 4 is also a symmetric k-PNN predictor.

Theorem 3.3. Suppose that the conditions of Lemma 3.2 hold and that T is an honest k-regular symmetric tree in the sense of Definitions 2 (part a), 4 (part a), and 5. Suppose moreover that the conditional moments $\mu(x)$ and $\mu_2(x)$ are both Lipschitz continuous at x. Finally, suppose that $\mathrm{var}[Y \mid X = x] > 0$. Then T is $\nu(s)$-incremental at x with

$$\nu(s) = C_{f,\,d} / \log(s)^d, \tag{21}$$

where $C_{f,\,d}$ is the constant from Lemma 3.2.

Finally, the result of Theorem 3.3 also holds for double-sample trees of the form described in Procedure 1. To establish the following result, we note that a double-sample tree is an honest, symmetric k-PNN predictor with respect to the $\mathcal{I}$-sample, while all the data in the $\mathcal{J}$-sample can be folded into the auxiliary noise term ξ; the details are worked out in the proof.

Corollary 3. Under the conditions of Theorem 3.3, suppose that T is instead a double-sample tree (Procedure 1) satisfying Definitions 2 (part b), 4 (part b), and 5. Then, T is ν-incremental, with $\nu(s) = C_{f,\,d}/(4\log(s)^d)$.

3.3.2. Subsampling Incremental Base Learners

In the previous section, we showed that decision trees are ν-incremental, in that the Hájek projection $\mathring{T}$ of T preserves at least some of the variation of T. In this section, we show that randomly subsampling ν-incremental predictors makes them 1-incremental; this then lets us proceed with a classical statistical analysis. The following lemma, which flows directly from the ANOVA decomposition of Efron and Stein (1981), provides a first motivating result for our analysis.

Lemma 3.3. Let $\hat{\mu}(x)$ be the estimate for $\mu(x)$ generated by a random forest with base learner T as defined in (12), and let $\mathring{\hat{\mu}}$ be the Hájek projection of $\hat{\mu}$ (17). Then

$$\mathbb{E}[(\hat{\mu}(x) - \mathring{\hat{\mu}}(x))^2] \leq \left(\frac{s}{n}\right)^2 \mathrm{var}[T(x;\ \xi,\ Z_1,\ \ldots,\ Z_s)]$$

whenever the variance $\mathrm{var}[T]$ of the base learner is finite.

This technical result paired with Theorem 3.3 or Corollary 3 leads to an asymptotic Gaussianity result; from a technical point of view, it suffices to check Lyapunov-style conditions for the central limit theorem.

Theorem 3.4. Let $\hat{\mu}(x)$ be a random forest estimator trained according the conditions of Theorem 3.3 or Corollary 3. Suppose, moreover, that the subsample size s_n satisfies

$$\lim_{n\to\infty} s_n = \infty \text{ and } \lim_{n\to\infty} s_n \log(n)^d / n = 0,$$

and that $\mathbb{E}[|Y - \mathbb{E}[Y \mid X = x]|^{2+\delta} \mid X = x] \leq M$ for some constants $\delta,\ M > 0$, uniformly over all $x \in [0,\ 1]^d$. Then, there exists a sequence $\sigma_n(x) \to 0$ such that

$$\frac{\hat{\mu}_n(x) - \mathbb{E}[\hat{\mu}_n(x)]}{\sigma_n(x)} \Rightarrow \mathcal{N}(0,\ 1), \tag{22}$$

where $\mathcal{N}(0,\ 1)$ is the standard normal distribution.

Moreover, as we show below, it is possible to accurately estimate the variance of a random forest using the infinitesimal jackknife for random forests (Efron 2014; Wager, Hastie, and Efron 2014).

Theorem 3.5. Let $\widehat{V}_{IJ}(x;\ ,\ Z_1,\ \ldots,\ Z_n)$ be the infinitesimal jackknife for random forests as defined in (8). Then, under the conditions of Theorem 3.4,

$$\widehat{V}_{IJ}(x;\ Z_1,\ \ldots,\ Z_n) / \sigma_n^2(x) \to_p 1. \tag{23}$$

Finally, we end this section by motivating the finite sample correction $n(n-1)/(n-s)^2$ appearing in (13) by considering the simple case where we have trivial trees that do not make any splits: $T(x;\, \xi,\, Z_{i_1},\, \ldots,\, Z_{i_s}) = s^{-1}\sum_{j=1}^{s} Y_{i_j}$. In this case, we can verify that the full random forest is nothing but $\hat{\mu} = n^{-1}\sum_{i=1}^{n} Y_i$, and the standard variance estimator

$$\widehat{V}_{\text{simple}} = \frac{1}{n\,(n-1)}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,\quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is well-known to be unbiased for $\text{Var}[\hat{\mu}]$. We show below that, for trivial trees $\widehat{V}_{IJ} = \widehat{V}_{\text{simple}}$, implying that our correction makes $\widehat{V}_{IJ}$ exactly unbiased in finite samples for trivial trees. Of course, $n(n-1)/(n-s)^2 \to 1$, and so Theorem 3.5 would hold even without this finite-sample correction; however, we find it to substantially improve the performance of our method in practice.

Proposition 5. For trivial trees $T(x;\, \xi,\, Z_{i_1},\, \ldots,\, Z_{i_s}) = s^{-1}\sum_{j=1}^{s} Y_{i_j}$, the variance estimate $\widehat{V}_{IJ}$ (13) is equivalent to the standard variance estimator $\widehat{V}_{\text{simple}}$, and $\mathbb{E}[\widehat{V}_{IJ}] = \text{var}[\hat{\mu}]$.

4. Inferring Heterogeneous Treatment Effects

We now return to our main topic, namely estimating heterogeneous treatment effects using random forests in the potential outcomes framework with unconfoundedness, and adapt our asymptotic theory for regression forests to the setting of causal inference. Here, we again work with training data consisting of tuples $Z_i = (X_i,\, Y_i,\, W_i)$ for $i = 1,\, \ldots,\, n$, where X_i is a feature vector, Y_i is the response, and W_i is the treatment assignment. Our goal is to estimate the conditional average treatment effect $\tau(x) = \mathbb{E}[Y^{(1)} - Y^{(0)} \,|\, X = x]$ at a pre-specified test point x. By analogy to Definition 1, we build our causal forest CF by averaging estimates for τ obtained by training causal trees Γ over subsamples:

$$\text{CF}(x;\, Z_1,\, \ldots,\, Z_n) = \binom{n}{s}^{-1}\sum_{1\le i_1<i_2<\ldots<i_s\le n}\mathbb{E}_{\xi\sim\Xi} \times[\Gamma(x;\, \xi,\, Z_{i_1},\, \cdots,\, Z_{i_s})]. \quad (24)$$

We seek an analogue to Theorem 3.1 for such causal forests.

Most of the definitions used to state Theorem 3.1 apply directly to this context; however, the notions of honesty and regularity need to be adapted slightly. Specifically, an honest causal tree is not allowed to look at the responses Y_i when making splits but can look at the treatment assignments W_i. Meanwhile, a regular causal tree must have at least k examples from both treatment classes in each leaf; in other words, regular causal trees seek to act as fully grown trees for the rare treatment assignment, while allowing for more instances of the common treatment assignment.

Definition 2b. A causal tree grown on a training sample $(Z_1 = (X_1,\, Y_1,\, W_1),\, \ldots,\, Z_s = (X_s,\, Y_s,\, W_s))$ is *honest* if (a) (*standard case*) the tree does not use the responses $Y_1,\, \ldots,\, Y_s$ in choosing where to place its splits; or (b) (*double sample case*) the tree does not use the $\mathcal{I}$-sample responses for placing splits.

Definition 4b. A causal tree grown by recursive partitioning is *α-regular* at x for some $\alpha > 0$ if either: (a) (*standard case*) (1) each split leaves at least a fraction α of the available training examples on each side of the split, (2) the leaf containing x has at least k observations from each treatment group ($W_i \in \{0,\, 1\}$) for some $k \in \mathbb{N}$, and (3) the leaf containing x has either less than $2k-1$ observations with $W_i = 0$ or $2k-1$ observations with $W_i = 1$; or (b) (*double-sample case*) for a double-sample tree as defined in Procedure 1, (a) holds for the $\mathcal{I}$ sample.

Given these assumptions, we show a close analogue to Theorem 3.1, given below. The main difference relative to our first result about regression forests is that we now rely on unconfoundedness and overlap to achieve consistent estimation of $\tau(x)$. To see how these assumptions enter the proof, recall that an honest causal tree uses the features X_i and the treatment assignments W_i in choosing where to place its splits, but not the responses Y_i. Writing $\mathcal{I}^{(1)}(x)$ and $\mathcal{I}^{(0)}(x)$ for the indices of the treatment and control units in the leaf around x, we then find that after the splitting stage

$$\begin{aligned}
&\mathbb{E}[\Gamma(x)|\, X,\, W] \\
&\quad= \frac{\sum_{\{i\in\mathcal{I}^{(1)}(x)\}}\mathbb{E}[Y^{(1)}\,|\,X = X_i,\, W = 1]}{|\mathcal{I}^{(1)}(x)|} \\
&\qquad- \frac{\sum_{\{i\in\mathcal{I}^{(0)}(x)\}}\mathbb{E}[Y^{(0)}\,|\,X = X_i,\, W = 0]}{|\mathcal{I}^{(0)}(x)|} \\
&\quad= \frac{\sum_{\{i\in\mathcal{I}^{(1)}(x)\}}\mathbb{E}[Y^{(1)}\,|\,X = X_i]}{|\mathcal{I}^{(1)}(x)|} \\
&\qquad- \frac{\sum_{\{i\in\mathcal{I}^{(0)}(x)\}}\mathbb{E}[Y^{(0)}\,|\,X = X_i]}{|\mathcal{I}^{(0)}(x)|},
\end{aligned} \quad (25)$$

where the second equality follows by unconfoundedness (2). Thus, it suffices to show that the two above terms are consistent for estimating $\mathbb{E}[Y^{(0)}\,|\,X = x]$ and $\mathbb{E}[Y^{(1)}\,|\,X = x]$. To do so, we can essentially emulate the argument leading to Theorem 3.1, provided we can establish an analogue to Lemma 3.1 and give a fast enough decaying upper bound to the diameter of $L(x)$; this is where we need the overlap assumption. A proof of Theorem 4.1 is given in the Appendix.

Theorem 4.1. Suppose that we have n independent and identically distributed training examples $Z_i = (X_i,\, Y_i,\, W_i) \in [0,\, 1]^d \times \mathbb{R} \times \{0,\, 1\}$. Suppose, moreover, that the treatment assignment is unconfounded (2) and has overlap (6). Finally, suppose that both potential outcome distributions $(X_i,\, Y_i^{(0)})$ and $(X_i,\, Y_i^{(1)})$ satisfy the same regularity assumptions as the pair $(X_i,\, Y_i)$ did in the statement of Theorem 3.1. Given this data-generating process, let Γ be an honest, α-regular with $\alpha \le 0.2$, and symmetric random-split causal forest in the sense of Definitions 2b, 3, 4b, and 5, and let $\hat{\tau}(x)$ be the estimate for $\tau(x)$ given by a causal forest with base learner Γ and a subsample size s_n scaling as in (14). Then, the predictions $\hat{\tau}(x)$ are consistent and asymptotically both Gaussian and centered, and the variance of the causal forest can be consistently estimated using the infinitesimal jackknife for random forests, that is, (7) holds.

Remark 4. (Testing at many points) We note that it is not in general possible to construct causal trees that are regular in the sense of Definition 4b for all x simultaneously. As a simple example, consider the situation where $d = 1$, and $W_i = 1(\{X_i \ge 0\})$; then, the tree can have at most 1 leaf for which it is regular. In the

proof of Theorem 4.1, we avoided this issue by only considering a single test point x, as it is always possible to build a tree that is regular at a single given point x. In practice, if we want to build a causal tree that can be used to predict at many test points, we may need to assign different trees to be valid for different test points. Then, when predicting at a specific x, we treat the set of trees that were assigned to be valid at that x as the relevant forest and apply Theorem 4.1 to it.

5. Simulation Experiments

In observational studies, accurate estimation of heterogeneous treatment effects requires overcoming two potential sources of bias. First, we need to identify neighborhoods over which the actual treatment effect $\tau(x)$ is reasonably stable and, second, we need to make sure that we are not biased by varying sampling propensities $e(x)$. The simulations here aim to test the ability of causal forests to respond to both of these factors.

Since causal forests are adaptive nearest neighbor estimators, it is natural to use a nonadaptive nearest neighborhood method as our baseline. We compare our method to the standard k nearest neighbors (k-NN) matching procedure, which estimates the treatment effect as

$$\hat{\tau}_{\text{KNN}}(x) = \frac{1}{k} \sum_{i \in \mathcal{S}_1(x)} Y_i - \frac{1}{k} \sum_{i \in \mathcal{S}_0(x)} Y_i, \tag{26}$$

where $\mathcal{S}_1$ and $\mathcal{S}_0$ are the k nearest neighbors to x in the treatment ($W = 1$) and control ($W = 0$) samples respectively. We generate confidence intervals for the k-NN method by modeling $\hat{\tau}_{\text{KNN}}(x)$ as Gaussian with mean $\tau(x)$ and variance $(\widehat{V}(\mathcal{S}_0) + \widehat{V}(\mathcal{S}_1))/(k(k-1))$, where $\widehat{V}(\mathcal{S}_{0/1})$ is the sample variance for $\mathcal{S}_{0/1}$.

The goal of this simulation study is to verify that forest-based methods can be used build rigorous, asymptotically valid confidence intervals that improve over nonadaptive methods like k-NN in finite samples. The fact that forest-based methods hold promise for treatment effect estimation in terms of predictive error has already been conclusively established elsewhere; for example, BART methods following Hill (2011) won the recent Causal Inference Data Analysis Challenge at the 2016 Atlantic Causal Inference Conference. We hope that the conceptual tools developed in this article will prove to be helpful in analyzing a wide variety of forest-based methods.

5.1. Experimental Setup

We describe our experiments in terms of the sample size n, the ambient dimension d, as well as the following functions:

$$\begin{aligned}
&\text{main effect: } m(x) = 2^{-1}\,\mathbb{E}[Y^{(0)} + Y^{(1)} \mid X = x],\\
&\text{treatment effect: } \tau(x) = \mathbb{E}[Y^{(1)} - Y^{(0)} \mid X = x],\\
&\text{treatment propensity: } e(x) = \mathbb{P}[W = 1 \mid X = x].
\end{aligned}$$

In all our examples, we respect unconfoundedness (2), use $X \sim U([0, 1]^d)$, and have homoscedastic noise $Y^{(0/1)} \sim \mathcal{N}(\mathbb{E}[Y^{(0/1)} \,|\, X], 1)$. We evaluate performance in terms of expected mean-squared error for estimating $\tau(X)$ at a random test example X, as well as expected coverage of $\tau(X)$ with a target coverage rate of 0.95.

In our first experiment, we held the treatment effect fixed at $\tau(x) = 0$, and tested the ability of our method to resist bias due to an interaction between $e(x)$ and $m(x)$. This experiment is intended to emulate the problem that in observational studies, a treatment assignment is often correlated with potential outcomes, creating bias unless the statistical method accurately adjusts for covariates. k-NN matching is a popular approach for performing this adjustment in practice. Here, we set

$$e(X) = \frac{1}{4}(1 + \beta_{2,4}(X_1)), \quad m(X) = 2X_1 - 1, \tag{27}$$

where $\beta_{a,b}$ is the β-density with shape parameters a and b. We used $n = 500$ samples and varied d between 2 and 30. Since our goal is accurate propensity matching, we use propensity trees (Procedure 2) as our base learner; we grew $B = 1000$ trees with $s = 50$.

For our second experiment, we evaluated the ability of causal forests to adapt to heterogeneity in $\tau(x)$, while holding $m(x) = 0$ and $e(x) = 0.5$ fixed. Thanks to unconfoundedness, the fact that $e(x)$ is constant means that we are in a randomized experiment. We set τ to be a smooth function supported on the first two features:

$$\tau(x) = \varsigma(X_1)\,\varsigma(X_2), \quad \varsigma(x) = 1 + \frac{1}{1 + e^{-20(x-1/3)}}. \tag{28}$$

We took $n = 5000$ samples, while varying the ambient dimension d from 2 to 8. For causal forests, we used double-sample trees with the splitting rule of Athey and Imbens (2016) as our base learner (Procedure 1). We used $s = 2500$ (i.e., $|\mathcal{I}| = 1250$) and grew $B = 2000$ trees.

One weakness of nearest neighbor approaches in general, and random forests in particular, is that they can fill the valleys and flatten the peaks of the true $\tau(x)$ function, especially near the edge of feature space. We demonstrate this effect using an example similar to the one studied above, except now $\tau(x)$ has a sharper spike in the $x_1,\ x_2 \approx 1$ region:

$$\tau(x) = \varsigma(X_1)\,\varsigma(X_2), \quad \varsigma(x) = \frac{2}{1 + e^{-12(x-1/2)}}. \tag{29}$$

We used the same training method as with (28), except with $n = 10,000$, $s = 2000$, and $B = 10,000$.

We implemented our simulations in R, using the packages `causalTree` (Athey and Imbens 2016) for building individual trees, `randomForestCI` (Wager, Hastie, and Efron 2014) for computing $\widehat{V}_{IJ}$, and `FNN` (Beygelzimer et al. 2013) for k-NN regression. All our trees had a minimum leaf size of $k = 1$. Software replicating the above simulations is available from the authors.

5.2. Results

In our first setup (27),causal forests present a striking improvement over k-NN matching; see Table 1. Causal forests succeed in maintaining a mean-squared error of 0.02 as d grows from 2 to 30, while 10-NN and 100-NN do an order of magnitude worse. We note that the noise of k-NN due to variance in Y after conditioning on X and W is already $2/k$, implying that k-NN with $k \leq 100$ cannot hope to match the performance of causal forests. Here, however, 100-NN is overwhelmed by bias,

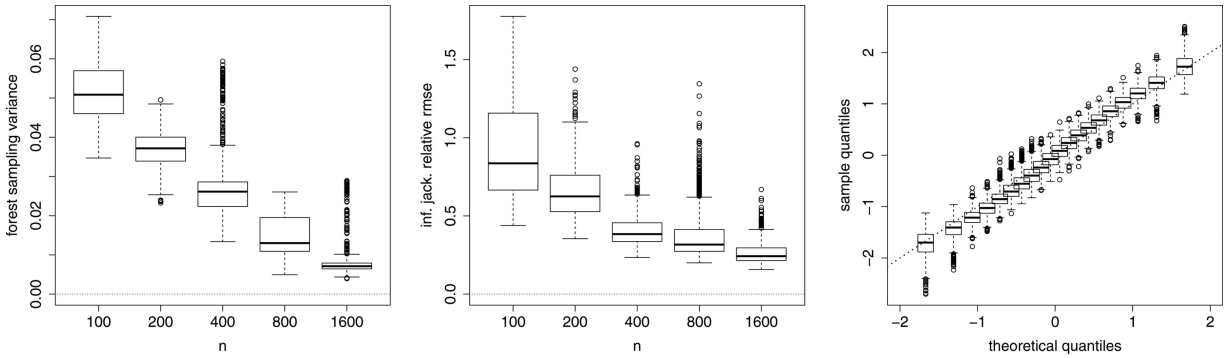

Figure 1. Graphical diagnostics for causal forests in the setting of (27). The first two panels evaluate the sampling error of causal forests and our infinitesimal jackknife estimate of variance over 1,000 randomly draw test points, with $d = 20$. The right-most panel shows standardized Gaussian *QQ*-plots for predictions at the same 1000 test points, with $n = 800$ and $d = 20$. The first two panels are computed over 50 randomly drawn training sets, and the last one over 20 training sets.

Table 1. Comparison of the performance of a causal forests (CF) with that of the k-nearest neighbors (k-NN) estimator with $k = 10$, 100, on the setup (27). The numbers in parentheses indicate the (rounded) standard sampling error for the last printed digit, obtained by aggregating performance over 500 simulation replications

	Mean-squared error			Coverage		
d	CF	10-NN	100-NN	CF	10-NN	100-NN
2	0.02 (0)	0.21 (0)	0.09 (0)	0.95 (0)	0.93 (0)	0.62 (1)
5	0.02 (0)	0.24 (0)	0.12 (0)	0.94 (1)	0.92 (0)	0.52 (1)
10	0.02 (0)	0.28 (0)	0.12 (0)	0.94 (1)	0.91 (0)	0.51 (1)
15	0.02 (0)	0.31 (0)	0.13 (0)	0.91 (1)	0.90 (0)	0.48 (1)
20	0.02 (0)	0.32 (0)	0.13 (0)	0.88 (1)	0.89 (0)	0.49 (1)
30	0.02 (0)	0.33 (0)	0.13 (0)	0.85 (1)	0.89 (0)	0.48 (1)

even with $d = 2$. Meanwhile, in terms of uncertainty quantification, our method achieves nominal coverage up to $d = 10$, after which the performance of the confidence intervals starts to decay. The 10-NN method also achieves decent coverage; however, its confidence intervals are much wider than ours as evidenced by the mean-squared error.

Figure 1 offers some graphical diagnostics for causal forests in the setting of (27). In the left panel, we observe how the causal forest sampling variance $\sigma_n^2(x)$ goes to zero with n; while the center panel depicts the decay of the relative root-mean squared error of the infinitesimal jackknife estimate of variance, that is, $\mathbb{E}[(\hat{\sigma}_n^2(x) - \sigma_n^2(x))^2]^{1/2}/\sigma_n^2(x)$. The boxplots display aggregate results for 1,000 randomly sampled test points x. Finally, the right-most panel evaluates the Gaussianity of the forest predictions. Here, we first drew 1,000 random test points x, and computed $\hat{\tau}(x)$ using forests grown on many different training sets. The plot shows standardized Gaussian *QQ*-plots aggregated over all these x; that is, for each x, we plot Gaussian theoretical quantiles against sample quantiles of $(\hat{\tau}(x) - \mathbb{E}(\hat{\tau}(x)))/\sqrt{\mathrm{Var}[\hat{\tau}(x)]}$.

In our second setup (28), causal forests present a similar improvement over k-NN matching when $d > 2$, as seen in Table 2.[10] Unexpectedly, we find that the performance of causal forests *improves* with d, at least when d is small. To understand this phenomenon, we note that the variance of a forest depends on the product of the variance of individual trees times the correlation between different trees (Breiman 2001a; Hastie, Tibshirani, and Friedman 2009). Apparently, when d is larger, the individual trees have more flexibility in how to place their splits, thus reducing their correlation and decreasing the variance of the full ensemble.

Finally, in the setting (29), Table 3 shows that causal forests still achieve an order of magnitude improvement over k-NN in terms of mean-squared error when $d > 2$, but struggle more in terms of coverage. This appears to largely be a bias effect: especially as d gets larger, the random forest is dominated by bias instead of variance and so the confidence intervals are not centered. Figure 2 illustrates this phenomenon: although the causal forest faithfully captures the qualitative aspects of the true τ-surface, it does not exactly match its shape, especially in the upper-right corner where τ is largest. Our theoretical results guarantee that this effect will go away as $n \to \infty$. Figure 2 also helps us understand why k-NN performs so poorly in terms of mean-squared error: its predictive surface is both badly biased and noticeably "grainy," especially for $d = 20$. It suffers from bias not only at the boundary where the treatment effect is the

Table 2. Comparison of the performance of a causal forests (CF) with that of the k-nearest neighbors (k-NN) estimator with $k = 7$, 50, on the setup (28). The numbers in parentheses indicate the (rounded) standard sampling error for the last printed digit, obtained by aggregating performance over 25 simulation replications

	Mean-squared error			Coverage		
d	CF	7-NN	50-NN	CF	7-NN	50-NN
2	0.04 (0)	0.29 (0)	0.04 (0)	0.97 (0)	0.93 (0)	0.94 (0)
3	0.03 (0)	0.29 (0)	0.05 (0)	0.96 (0)	0.93 (0)	0.92 (0)
4	0.03 (0)	0.30 (0)	0.08 (0)	0.94 (0)	0.93 (0)	0.86 (1)
5	0.03 (0)	0.31 (0)	0.11 (0)	0.93 (1)	0.92 (0)	0.77 (1)
6	0.02 (0)	0.34 (0)	0.15 (0)	0.93 (1)	0.91 (0)	0.68 (1)
8	0.03 (0)	0.38 (0)	0.21 (0)	0.90 (1)	0.90 (0)	0.57 (1)

Table 3. Comparison of the performance of a causal forests (CF) with that of the k-nearest neighbors (k-NN) estimator with $k = 10$, 100, on the setup (29). The numbers in parentheses indicate the (rounded) standard sampling error for the last printed digit, obtained by aggregating performance over 40 simulation replications

	Mean-squared error			Coverage		
d	CF	10-NN	100-NN	CF	10-NN	100-NN
2	0.02 (0)	0.20 (0)	0.02 (0)	0.94 (0)	0.93 (0)	0.94 (0)
3	0.02 (0)	0.20 (0)	0.03 (0)	0.90 (0)	0.93 (0)	0.90 (0)
4	0.02 (0)	0.21 (0)	0.06 (0)	0.84 (1)	0.93 (0)	0.78 (1)
5	0.02 (0)	0.22 (0)	0.09 (0)	0.81 (1)	0.93 (0)	0.67 (0)
6	0.02 (0)	0.24 (0)	0.15 (0)	0.79 (1)	0.92 (0)	0.58 (0)
8	0.03 (0)	0.29 (0)	0.26 (0)	0.73 (1)	0.90 (0)	0.45 (0)

[10] When $d = 2$, we do not expect causal forests to have a particular advantage over k-NN since the true τ also has 2-dimensional support; our results mirror this, as causal forests appear to have comparable performance to 50-NN.

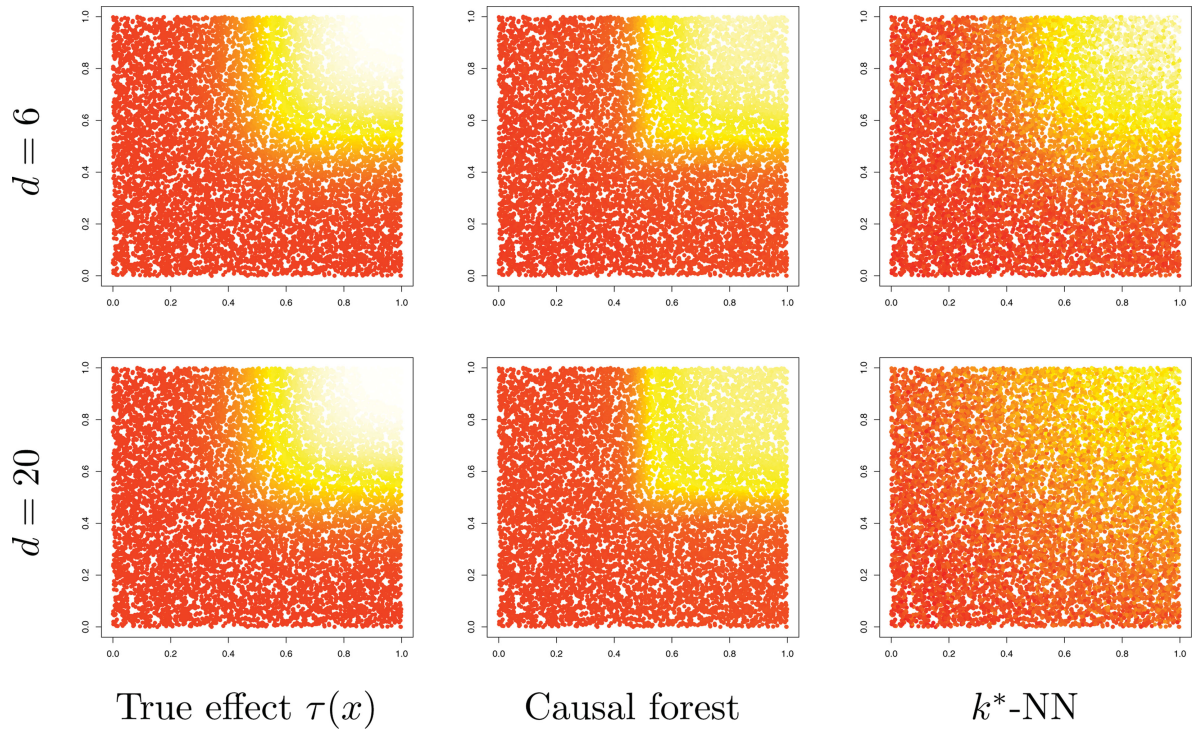

Figure 2. The true treatment effect $\tau(X_i)$ at 10,000 random test examples X_i, along with estimates $\hat{\tau}(X_i)$ produced by a causal forest and optimally-tuned k-NN, on data drawn according to (29) with $d = 6,\ 20$. The test points are plotted according to their first two coordinates; the treatment effect is denoted by color, from dark (low) to light (high). On this simulation instance, causal forests and k^*-NN had a mean-squared error of 0.03, and 0.13 respectively, for $d = 6$, and of 0.05 and 0.62, respectively, for $d = 20$. The optimal tuning choices for k-NN were $k^* = 39$ for $d = 6$, and $k^* = 24$ for $d = 20$.

largest, but also where the slope of the treatment effect is high in the interior.

These results highlight the promise of causal forests for accurate estimation of heterogeneous treatment effects, all while emphasizing avenues for further work. An immediate challenge is to control the bias of causal forests to achieve better coverage. Using more powerful splitting rules is a good way to reduce bias by enabling the trees to focus more closely on the coordinates with the greatest signal. The study of splitting rules for trees designed to estimate causal effects is still in its infancy and improvements may be possible.

A limitation of the present simulation study is that we manually chose whether to use double-sample forests or propensity forests, depending on which procedure seemed more appropriate in each problem setting. An important challenge for future work is to design splitting rules that can automatically choose which characteristic of the training data to split on. A principled and automatic rule for choosing s would also be valuable.

We present additional simulation results in the supplementary material. Appendix A has extensive simulations in the setting of Table 2 while varying both s and n; and also considers a simulation setting, where the signal is spread out over many different features, meaning that forests have less upside over baseline methods. Finally, in Appendix B, we study the effect of honesty versus adaptivity on forest predictive error.

6. Discussion

This article proposed a class of nonparametric methods for heterogeneous treatment effect estimation that allow for data-driven feature selection all while maintaining the benefits of classical methods, that is, asymptotically normal and unbiased point estimates with valid confidence intervals. Our causal forest estimator can be thought of as an adaptive nearest neighbor method, where the data determine which dimensions are most important to consider in selecting nearest neighbors. Such adaptivity seems essential for modern large-scale applications with many features.

In general, the challenge in using adaptive methods as the basis for valid statistical inference is that selection bias can be difficult to quantify; see Berk et al. (2013), Chernozhukov, Hansen, and Spindler (2015), Taylor and Tibshirani (2015), and references therein for recent advances. In this article, pairing "honest" trees with the subsampling mechanism of random forests enabled us to accomplish this goal in a simple yet principled way. In our simulation experiments, our method provides dramatically better mean-squared error than classical methods while achieving nominal coverage rates in moderate sample sizes.

A number of important extensions and refinements are left open. Our current results only provide pointwise confidence intervals for $\tau(x)$; extending our theory to the setting of global functional estimation seems like a promising avenue for further work. Another challenge is that nearest-neighbor nonparametric estimators typically suffer from bias at the boundaries of the support of the feature space. A systematic approach to trimming at the boundaries, and possibly correcting for bias, would improve the coverage of the confidence intervals. In general, work can be done to identify methods that produce accurate variance estimates even in more challenging circumstances, for example, with small samples or a large number of covariates, or to identify when variance estimates are unlikely to be reliable.

Acknowledgment

We are grateful for helpful feedback from Brad Efron, Trevor Hastie, Guido Imbens, Guenther Walther, as well as the associate editor, two anonymous referees, and seminar participants at Atlantic Causal Inference Conference, Berkeley Haas, Berkeley Statistics, California Econometrics Conference, Cambridge, Carnegie Mellon, COMPSTAT, Cornell, Columbia Business School, Columbia Statistics, CREST, EPFL, ISAT/DARPA Machine Learning for Causal Inference, JSM, London Business School, Microsoft Conference on Digital Economics, Merck, MIT IDSS, MIT Sloan, Northwestern, SLDM, Stanford GSB, Stanford Statistics, University College London, University of Bonn, University of Chicago Booth, University of Chicago Statistics, University of Illinois Urbana–Champaign, University of Pennsylvania, University of Southern California, University of Washington, Yale Econometrics, and Yale Statistics. S. W. was partially supported by a B. C. and E. J. Eaves Stanford Graduate Fellowship. Part of the results developed in this article were made available as an earlier technical report *"Asymptotic Theory for Random Forests"*, available at http://arxiv.org/abs/1405.0352.

References

Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000), "Subgroup Analysis and Other (mis) Uses of Baseline Data in Clinical Trials," *The Lancet*, 355, 1064–1069. [1228]

Athey, S., and Imbens, G. (2016), "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, 113, 7353–7360. [1229,1238]

Athey, S., Tibshirani, J., and Wager, S. (2018), "Generalized Random Forests," *Annals of Statistics*, forthcoming. [1230]

Barndorff-Nielsen, O., and Sobel, M. (1966), "On the Distribution of the Number of Admissible Points in a Vector Random Sample," *Theory of Probability & Its Applications*, 11, 249–269. [1235]

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), "Valid Post-Selection Inference," *The Annals of Statistics*, 41, 802–837. [1240]

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., and Li, S. (2013), *FNN: Fast Nearest Neighbor Search Algorithms and Applications*, R package version 1.1. [1238]

Beygelzimer, A., and Langford, J. (2009), "The Offset Tree for Learning with Partial Labels," in *Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 129–138. [1230]

Bhattacharya, D., and Dupas, P. (2012), "Inferring Welfare Maximizing Treatment Assignment Under Budget Constraints," *Journal of Econometrics*, 167, 168–196. [1230]

Biau, G. (2012), "Analysis of a Random Forests Model," *The Journal of Machine Learning Research*, 13, 1063–1095. [1229,1231,1234]

Biau, G., Cérou, F., and Guyader, A. (2010), "On the Rate of Convergence of the Bagged Nearest Neighbor Estimate," *The Journal of Machine Learning Research*, 11, 687–712. [1234]

Biau, G., and Devroye, L. (2010), "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, 2499–2518. [1235]

Biau, G., Devroye, L., and Lugosi, G. (2008), "Consistency of Random Forests and other Averaging Classifiers," *The Journal of Machine Learning Research*, 9, 2015–2033. [1234]

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140. [1234]

——— (2001a), "Random Forests," *Machine Learning*, 45, 5–32. [1228,1232,1233,1236,1239]

——— (2001b), "Statistical Modeling: The Two Cultures" (with comments and a rejoinder by the author), *Statistical Science*, 16, 199–231. [1229]

——— (2004), "Consistency for a Simple Model of Random Forests," *Statistical Department, University of California at Berkeley. Technical Report*. [1234]

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), *Classification and Regression Trees*, Boca Raton, FL: CRC press. [1234,1236]

Bühlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927–961. [1231,1234]

Buja, A., and Stuetzle, W. (2006), "Observations on Bagging," *Statistica Sinica*, 16, 323. [1234]

Chen, S. X., and Hall, P. (2003), "Effects of Bagging and Bias Correction on Estimators Defined by Estimating Equations," *Statistica Sinica*, 13, 97–110. [1234]

Chernozhukov, V., Hansen, C., and Spindler, M. (2015), "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," *Annual Review of Economics*, 7, 649–688. [1240]

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), "BART: Bayesian Additive Regression Trees," *The Annals of Applied Statistics*, 4, 266–298. [1229]

Cook, D. I., Gebski, V. J., and Keech, A. C. (2004), "Subgroup Analysis in Clinical Trials," *Medical Journal of Australia*, 180, 289–292. [1228]

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2008), "Nonparametric Tests for Treatment Effect Heterogeneity," *The Review of Economics and Statistics*, 90, 389–405. [1228]

Dehejia, R. H. (2005), "Program Evaluation as a Decision Problem," *Journal of Econometrics*, 125, 141–173. [1230]

Denil, M., Matheson, D., and De Freitas, N. (2014), "Narrowing the Gap: Random Forests In Theory and In Practice," in *Proceedings of The 31st International Conference on Machine Learning*, pp. 665–673. [1232]

Duan, J. (2011), "Bootstrap-Based Variance Estimators for a Bagging Predictor." Ph.D. thesis, North Carolina State University. [1234]

Dudík, M., Langford, J., and Li, L. (2011), "Doubly Robust Policy Evaluation and Learning," in *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104. [1230]

Efron, B. (2014), "Estimation and Accuracy after Model Selection" (with discussion), *Journal of the American Statistical Association*, 109, 991–1007. [1229,1231,1234,1235,1236]

Efron, B., and Stein, C. (1981), "The Jackknife Estimate of Variance," *The Annals of Statistics*, 9, 586–596. [1229,1235,1236]

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011), "Subgroup Identification from Randomized Clinical Trial Data," *Statistics in Medicine*, 30, 2867–2880. [1229]

Friedman, J. H., and Hall, P. (2007), "On Bagging and Nonlinear Estimation," *Journal of Statistical Planning and Inference*, 137, 669–683. [1234]

Green, D. P., and Kern, H. L. (2012), "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees," *Public Opinion Quarterly*, 76, 491–511. [1229]

Hájek, J. (1968), "Asymptotic Normality of Simple Linear Rank Statistics under Alternatives," *The Annals of Mathematical Statistics*, 39, 325–346. [1229,1231,1235]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning*, New York: Springer. [1230,1239]

Hill, J., and Su, Y.-S. (2013), "Assessing Lack of Common Support in Causal Inference using Bayesian Nonparametrics: Implications for Evaluating the Effect of Breastfeeding on Childrens Cognitive Outcomes," *The Annals of Applied Statistics*, 7, 1386–1420. [1229]

Hill, J. L. (2011), "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240. [1229,1238]

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [1230]

Hirano, K., and Porter, J. R. (2009), "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77, 1683–1701. [1230]

Hoeffding, W. (1948), "A Class of Statistics with Asymptotically Normal Distribution," *The Annals of Mathematical Statistics*, 19, 293–325. [1229,1231,1235]

Imai, K., and Ratkovic, M. (2013), "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation," *The Annals of Applied Statistics*, 7, 443–470. [1230]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, New York: Cambridge University Press. [1230]

Jaeckel, L. A. (1972), *The Infinitesimal Jackknife*, Technical Report, Bell Laboratories. [1231]

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015), "Prediction Policy Problems," *American Economic Review*, 105, 491–95. [1229]

Lee, M.-J. (2009), "Nonparametric Tests for Distributional Treatment Effect for Randomly Censored Responses," *Journal of the Royal Statistical Society*, Series B, 71, 243–264. [1228]

Liaw, A., and Wiener, M. (2002), "Classification and Regression by Random Forest," *R News*, 2, 18–22. [1234]

Lin, Y., and Jeon, Y. (2006), "Random Forests and Adaptive Nearest Neighbors," *Journal of the American Statistical Association*, 101, 578–590. [1229,1231,1234,1235,1236]

Mallows, C. L. (1973), "Some Comments on Cp," *Technometrics*, 15, 661–675. [1232]

Manski, C. F. (2004), "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 1221–1246. [1230]

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004), "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies," *Psychological Methods*, 9, 403. [1230]

Meinshausen, N. (2006), "Quantile regression forests," *The Journal of Machine Learning Research*, 7, 983–999. [1229,1231,1233,1234,1235]

Mentch, L., and Hooker, G. (2016), "Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests," *Journal of Machine Learning Research*, 17, 1–41. [1229,1233,1234]

Neyman, J. (1923), "Sur les Applications de la Théorie des Probabilités aux Experiences Agricoles: Essai des Principes," *Roczniki Nauk Rolniczych*, 10, 1–51. [1229,1230]

Politis, D. N., Romano, J. P., and Wolf, M. (1999), *Subsampling*, Springer Series in Statistics, New York: Springer. [1234]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [1230,1232]

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688. [1229,1230]

Samworth, R. J. (2012), "Optimal Weighted Nearest Neighbour Classifiers," *The Annals of Statistics*, 40, 2733–2763. [1234]

Schick, A. (1986), "On Asymptotically Efficient Estimation in Semiparametric Models," *The Annals of Statistics*, 1139–1151. [1232]

Scornet, E., Biau, G., and Vert, J.-P. (2015), "Consistency of Random Forests," *The Annals of Statistics*, 43, 1716–1741. [1229,1232,1234]

Sexton, J., and Laake, P. (2009), "Standard Errors for Bagged and Random Forest Estimators," *Computational Statistics & Data Analysis*, 53, 801–811. [1234]

Signorovitch, J. E. (2007), "Identifying Informative Biological Markers in High-Dimensional Genomic Data and Clinical Trials," Ph.D. thesis, Harvard University. [1230]

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007), "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution," *BMC Bioinformatics*, 8, 25. [1234]

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009), "Subgroup Analysis via Recursive Partitioning," *The Journal of Machine Learning Research*, 10, 141–158. [1230]

Taddy, M., Gardner, M., Chen, L., and Draper, D. (2016), "A Nonparametric Bayesian Analysis of Heterogeneous Treatment Effects in Digital Experimentation," *Journal of Business & Economic Statistics*, 34, 661–672. [1230]

Taylor, J., and Tibshirani, R. J. (2015), "Statistical Learning and Selective Inference," *Proceedings of the National Academy of Sciences*, 112, 7629–7634. [1240]

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014), "A Simple Method for Estimating Interactions between a Treatment and a Large Number of Covariates," *Journal of the American Statistical Association*, 109, 1517–1532. [1230]

Van der Vaart, A. W. (2000), *Asymptotic Statistics*, no. 3, New York: Cambridge Univ Press. [1235]

Wager, S., Hastie, T., and Efron, B. (2014), "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife," *The Journal of Machine Learning Research*, 15, 1625–1651. [1229,1231,1233,1234,1235,1236,1238]

Wager, S., and Walther, G. (2015), "Adaptive Concentration of Regression Trees, with Application to Random Forests," arXiv:1503.06388. [1232,1234]

Wang, P., Sun, W., Yin, D., Yang, J., and Chang, Y. (2015), "Robust Tree-Based Causal Inference for Complex ad Effectiveness Analysis," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ACM, pp. 67–76. [1232]

Wasserman, L., and Roeder, K. (2009), "High-Dimensional Variable Selection," *The Annals of Statistics*, 37, 2178–2201. [1232]

Weisberg, H. I., and Pontes, V. P. (2015), "Post hoc Subgroups in Clinical Trials: Anathema or Analytics?," *Clinical Trials*, 12, 357–364. [1230]

Westreich, D., Lessler, J., and Funk, M. J. (2010), "Propensity Score Estimation: Neural Networks, Support Vector Machines, Decision Trees (CART), and Meta-Classifiers as Alternatives to Logistic Regression," *Journal of Clinical Epidemiology*, 63, 826–833. [1230]

Willke, R. J., Zheng, Z., Subedi, P., Althin, R., and Mullins, C. D. (2012), "From Concepts, Theory, and Evidence of Heterogeneity of Treatment Effects to Methodological Approaches: A Primer," *BMC Medical Research Methodology*, 12, 185. [1228]

Zeileis, A., Hothorn, T., and Hornik, K. (2008), "Model-Based Recursive Partitioning," *Journal of Computational and Graphical Statistics*, 17, 492–514. [1230]

Zhu, R., Zeng, D., and Kosorok, M. R. (2015), "Reinforcement Learning Trees," *Journal of the American Statistical Association*, 110, 1770–1784. [1234]