Stanford | Institute for Economic Policy Research (SIEPR)

Generalized Random Forests

Susan Athey

Julie Tibshirani

Stefan Wager

July, 2017

Working Paper No. 17-037

Generalized Random Forests

Susan Athey
athey@stanford.edu

Julie Tibshirani
jtibs@cs.stanford.edu

Stefan Wager
swager@stanford.edu

Current version July 2017

Abstract

We propose generalized random forests, a method for non-parametric statistical estimation based on random forests (Breiman, 2001) that can be used to fit any quantity of interest identified as the solution to a set of local moment equations. Following the literature on local maximum likelihood estimation, our method operates at a particular point in covariate space by considering a weighted set of nearby training examples; however, instead of using classical kernel weighting functions that are prone to a strong curse of dimensionality, we use an adaptive weighting function derived from a forest designed to express heterogeneity in the specified quantity of interest. We propose a flexible, computationally efficient algorithm for growing generalized random forests, develop a large sample theory for our method showing that our estimates are consistent and asymptotically Gaussian, and provide an estimator for their asymptotic variance that enables valid confidence intervals. We use our approach to develop new methods for three statistical tasks: non-parametric quantile regression, conditional average partial effect estimation, and heterogeneous treatment effect estimation via instrumental variables. A software implementation, `grf` for `R` and `C++`, is available from `CRAN`.

1 Introduction

Random forests, introduced by Breiman (2001), have become one of the most popular methods in applied statistical learning. From a conceptual perspective, random forests are usually understood as a practical method for non-parametric conditional mean estimation: Given a data-generating distribution for $(X, Y) \in \mathcal{X} \times \mathbb{R}$, forests are used to estimate

$$\mu(x) := \mathbb{E}\left[Y \,\middle|\, X = x\right]. \tag{1}$$

We are grateful for helpful comments from several colleagues; in particular, we are indebted to Jerry Friedman for first recommending we take a closer look at splitting rules for quantile regression forests, to Will Fithian for drawing our attention to connections between our early ideas and gradient boosting, to Guido Imbens for suggesting the local centering scheme discussed in Section 6.1.1, to two anonymous referees, and to seminar participants at the Atlantic Causal Inference Conference, the California Econometrics Conference, Ca'Voscari University of Venice, Columbia, Cornell, the Econometric Society Winter Meetings, EPFL, the European University Institute, INFORMS, Kellogg, the Microsoft Conference on Digital Economics, the MIT Conference on Digital Experimentation, Northwestern, Toulouse, Triangle Computer Science Distinguished Lecture Series, University of Chicago, University of Illinois Urbana–Champaign, University of Lausanne, and the USC Dornsife Conference on Big Data in Economics.

Several theoretical results are available on the asymptotic behavior of variants of such forest-based estimates $\hat{\mu}(x)$, including consistency (Arlot and Genuer, 2014; Biau et al., 2008; Biau, 2012; Denil et al., 2014; Lin and Jeon, 2006; Scornet et al., 2015; Wager and Walther, 2015) and asymptotic normality (Mentch and Hooker, 2016; Wager and Athey, 2017).

The goal of our paper is to generalize Breiman's random forests, and to develop a forest-based method for estimating any quantity $\theta(x)$ identified via local moment conditions. Specifically, given data $(X, O) \in \mathcal{X} \times \mathcal{O}$, we take $\theta(x)$ to be defined via a local estimating equation of the form

$$\mathbb{E}\left[\psi_{\theta(x),\,\nu(x)}\left(O\right) \,\middle|\, X = x\right] = 0, \tag{2}$$

where $\psi(\cdot)$ is some scoring function and $\nu(x)$ is an optional nuisance parameter. This setup encompasses several key statistical problems. For example, if we model the distribution of O conditionally on X as having a density $f_{\theta(x),\,\nu(x)}(\cdot)$ then, under standard regularity conditions, the moment condition (2) with $\psi_{\theta(x),\,\nu(x)}(O) = \nabla \log\left(f_{\theta(x),\,\nu(x)}(O)\right)$ identifies the local maximum likelihood parameters $(\theta(x),\, \nu(x))$. More generally, we can use moment conditions of the form (2) to identify conditional means, conditional quantiles, conditional average partial effects, etc. Our main substantive application of generalized random forests involves heterogeneous treatment effect estimation via instrumental variables.

In developing generalized random forests, our aim is to build a family of non-parametric estimators that inherit the desirable empirical properties of regression forests—such as stability, ease of use, and flexible adaptation to different functional forms as in, e.g., Biau and Scornet (2016) or Varian (2014)—but can be used in the wide range of statistical settings characterized by (2) rather than just in regression problems of the type (1). Our main focus is on addressing the resulting conceptual and methodological challenges, and in establishing formal asymptotic inference results for such forests.

In order to support generic estimating equations as in (2), our method requires several important extensions to the standard regression forests of Breiman (2001). First, while regression forests are typically understood as ensemble methods, i.e., forest predictions $\hat{\mu}(x)$ are written as the average of B noisy tree-based predictors $\hat{\mu}_b(x)$,

$$\hat{\mu}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{\mu}_b(x), \tag{3}$$

this perspective is not appropriate more generally. Noisy solutions to moment equations as in (2) are generally biased, and averaging as in (3) would do nothing to alleviate the bias.[1]

To avoid this issue, we cast forests as a type of adaptive locally weighted estimators that first use a forest to calculate a weighted set of neighbors for each test point x, and then solve a plug-in version of the estimating equation (2) using these neighbors. Section 2.2 gives a detailed treatment of this perspective. This locally weighting view of random forests was previously advocated by Meinshausen (2006) in the context of quantile regression, and also underlies theoretical analyses of regression forests (e.g., Lin and Jeon, 2006). In the context of regression, the averaging and weighting views of forests are equivalent; however, once we move to more general settings, the weighting-based perspective will prove to be substantially more effective, and also brings forests closer to the classical literature on local

[1] In the special case of regression forests, individual trees $\hat{\mu}_b(x)$ have low bias but high variance, and so (3) does meaningfully stabilize predictions; see Scornet et al. (2015) or Wager and Athey (2017) for formal statements.

maximum likelihood estimation (Fan and Gijbels, 1996; Loader, 1999; Newey, 1994b; Stone, 1977; Tibshirani and Hastie, 1987).

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

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

In contrast, we seek a unified, general framework for computationally efficient problem-specific splitting rules, optimized for the primary objective of capturing heterogeneity in a key parameter of interest. In the spirit of gradient boosting (Friedman, 2001), our recursive partitioning method begins by computing a linear, gradient-based approximation to the non-linear estimating equation we are trying to solve, and then uses this approximation to specify the tree-split point. Algorithmically, our procedure reduces to iteratively applying a labeling step where we generate pseudo-outcomes by computing gradients using parameters estimated in the parent node, and a regression step where we pass this labeled data to a standard CART regression routine. Thus, we can make use of pre-existing, optimized tree software to execute the regression step, and obtain high quality neighborhood functions while only using computational resources comparable to those required by standard CART algorithms. In line with this approach, our generalized random forest software package builds on the carefully optimized `ranger` implementation of regression forest splitting rules (Wright and Ziegler, 2017).

Finally, moment conditions of the form (2) typically arise in scientific applications where rigorous statistical inference is required, and so a generalization of random forests to such problems would not be of much use without accompanying theoretical guarantees. The bulk of this paper is devoted to a theoretical analysis of generalized random forests, and to establishing asymptotic consistency and Gaussianity of the resulting estimates $\hat{\theta}(x)$. We also develop methodology for asymptotic confidence intervals. Our technical analysis is motivated by classical theory for local estimating equations, in particular the results of Newey (1994b), paired with machinery from Wager and Athey (2017) to address the adaptivity of the random forest weighting function.

The resulting generalized random forests present a flexible framework for non-parametric statistical estimation and inference with formal asymptotic guarantees. In this paper, we consider applications to quantile regression, conditional average partial effect estimation and heterogeneous treatment effect estimation with instruments; however, there are many other popular statistical models that fit directly into our framework, including panel regression, Huberized robust regression, models of consumer choice, etc. In order to fit any of these models with generalized random forests, the analyst simply needs to provide the problem-specific routines to calculate gradients of the moment conditions evaluated at different observations in the dataset for the "label" step of our algorithm. Moreover, despite all the required formalism, we emphasize that our method is in fact a proper generalization of regression forests: If we apply our framework to build a forest-based method for local least-squares regression, i.e., we use $\psi_{\mu(x)}(Y) = Y - \mu(x)$ in (2), we exactly recover a regression forest.

A high-performance software implementation of generalized random forests, `grf` for `R` and `C++`, is available from `CRAN` and at `https://www.github.com/swager/grf`.

1.1 Related Work

The idea of local maximum likelihood (and closely related, local generalized method of moments) estimation has a long history in statistics, with notable contributions from Fan et al. (1998), Newey (1994b), Staniswalis (1989), Stone (1977), Tibshirani and Hastie (1987), Lewbel (2007) and others. In the economics literature, popular applications of these techniques include local linear regression in regression discontinuity frameworks (see Imbens and Lemieux, 2008, for a review), multinomial choice modeling in a panel data setting (e.g., Honoré and Kyriazidou, 2000), and instrumental variables regression (Su et al., 2013). The basic idea is that when estimating parameters at a particular value of covariates, a kernel weighting function is used to place more weight on nearby observations in the covariate space. A challenge facing this approach is that if the covariate space has more than two or three dimensions, the "curse of dimensionality" implies that plain kernel-based methods may not perform well (e.g., Robins and Ritov, 1997).

Our paper takes the approach of replacing the kernel weighting with forest-based weights, that is, weights derived from the fraction of trees in which an observation appears in the same leaf as the target value of the covariate vector. The original random forest algorithm for non-parametric classification and regression was proposed by Breiman (2001), building on insights from the ensemble learning literature (Amit and Geman, 1997; Breiman, 1996; Dietterich, 2000; Ho, 1998). The perspective we take on random forests as a form of adaptive nearest neighbor estimation, however, most closely builds on the proposal of Meinshausen (2006) for forest-based quantile regression. This adaptive nearest neighbors perspective also underlies several statistical analyses of random forests, including those of Arlot and Genuer (2014), Biau and Devroye (2010), and Lin and Jeon (2006).

Meanwhile, our gradient-based splitting scheme draws heavily from a long tradition in the statistics and econometrics literatures of using gradient-based test statistics to detect change points in likelihood models (Andrews, 1993; Hansen, 1992; Hjort and Koning, 2002; Nyblom, 1989; Ploberger and Krämer, 1992; Zeileis, 2005; Zeileis and Hornik, 2007). In particular, Zeileis et al. (2008) consider the use of such methods for model-based recursive partitioning. Our problem setting differs from the above in that we are not focused on running a hypothesis test, but rather seek an adaptive nearest neighbor weighting that is as sensitive as possible to heterogeneity in our parameter of interest; we then rely on the random forest resampling mechanism to achieve statistical stability (Mentch and Hooker,

2016; Scornet et al., 2015; Wager and Athey, 2017). In this sense, our approach is closely related to the gradient boosting algorithm of Friedman (2001), who uses similar gradient-based approximations to guide a greedy, heuristic, non-parametric regression procedure.

Our asymptotic theory relates to an extensive recent literature on the statistics of random forests, most of which focuses on the regression case (Arlot and Genuer, 2014; Biau, 2012; Biau et al., 2008; Biau and Scornet, 2016; Breiman, 2004; Bühlmann and Yu, 2002; Chipman et al., 2010; Denil et al., 2014; Efron, 2014; Geurts et al., 2006; Ishwaran and Kogalur, 2010; Lin and Jeon, 2006; Meinshausen, 2006; Mentch and Hooker, 2016; Samworth, 2012; Scornet et al., 2015; Sexton and Laake, 2009; Wager and Athey, 2017; Wager and Walther, 2015; Zhu et al., 2015). Our present paper complements this body of work, by showing how methods developed to study regression forests can also be used understand estimated solutions to local moment equations obtained via generalized random forests.

Finally we note that the problem we study, namely estimating how a function $\theta(x)$ varies with covariates, is distinct from the problem of estimating a single, low-dimensional parameter—such as an average treatment effect—while controlling for a non-parametric or high-dimensional set of covariates. Recent contributions to the latter include Athey et al. (2016), Belloni et al. (2013), Chernozhukov et al. (2016), Robins et al. (1995), Robins et al. (2008), and van der Laan and Rubin (2006). In particular, Chernozhukov et al. (2016) and van der Laan and Rubin (2006) discuss how traditional machine learning methods like regression forests can be used as sub-components in efficient inference about such low-dimensional parameters.

2 Generalized Random Forests

2.1 Review of Breiman's Forests

In standard classification or regression forests as proposed by Breiman (2001), the prediction for a particular test point x is determined by averaging predictions across an ensemble of different trees (Amit and Geman, 1997; Breiman, 1996; Dietterich, 2000; Ho, 1998). Individual trees are grown by greedy recursive partitioning, i.e., we recursively add axis-aligned splits to the tree, where each split it chosen to maximize the improvement to model fit (Breiman et al., 1984); see Figure 1 for an example of a tree. The trees are randomized using bootstrap (or subsample) aggregation, whereby each tree is grown on a different random subset of the training data, and random split selection that restricts the variables available at each step of the algorithm.[2]

In generalizing random forests beyond the regression and classification contexts, we preserve several core elements of Breiman's forests—including recursive partitioning, subsampling, and random split selection—but must abandon the idea that our final estimate is obtained by averaging estimates from each member of an ensemble. As discussed below, standard regression forests can equivalently be described as a type of adaptive nearest neighbor estimator, and this alternative perspective is much more amenable to statistical extensions.

2.2 Forest-Based Local Estimation

Suppose now that we have independent and identically data for n samples, indexed $i = 1, ..., n$. For each sample, we have access to an observable quantity O_i that encodes in-

[2]For an introductory overview of random forests, we recommend the chapter of Hastie et al. (2009) dedicated to the method.

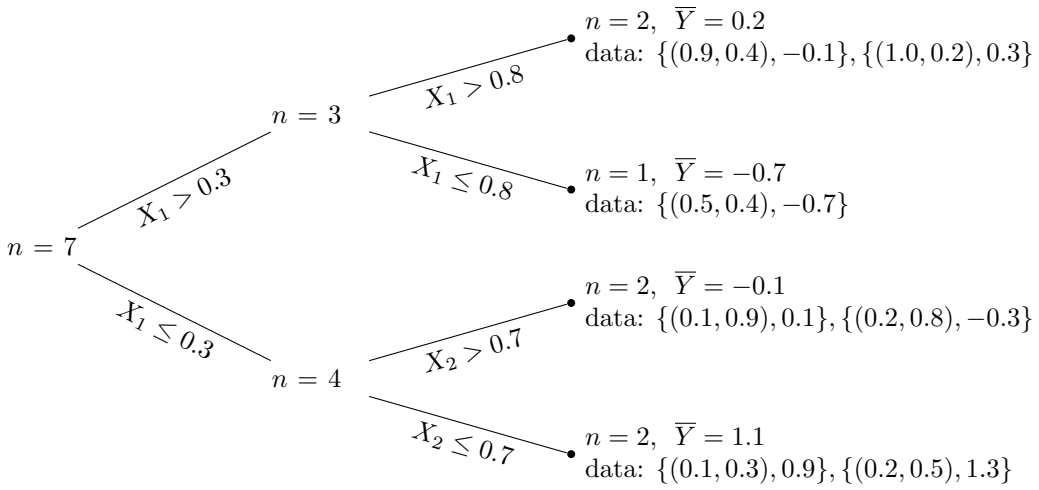


Figure 1: Example of a small regression tree on a sample of size $n = 7$. The examples used to build this tree are of the form $\{X_i, Y_i\} \in \mathbb{R}^2 \times \mathbb{R}$, and axis-aligned splits based on the X_i determine the leaf membership of each training example. In "standard" regression trees as discussed in, e.g., Breiman et al. (1984) or Hastie et al. (2009), the tree predicts by averaging the outcomes Y_i within the relevant leaf; thus, in the example of Figure 1, any test point x with $(x_1 \leq 0.3) \wedge (x_2 \leq 0.7)$ would be assigned a prediction $\hat{\mu}(x) = 1.1$.

formation relevant to estimating $\theta(\cdot)$, along with a set of auxiliary covariates X_i. In the case of non-parametric regression, this observable just consists of an outcome $O_i = \{Y_i\}$ with $Y_i \in \mathbb{R}$; in general, however, it will contain richer information. For example, in the case of treatment effect estimation with exogenous treatment assignment, $O_i = \{Y_i, W_i\}$ also includes the treatment assignment W_i. Given this type of data, our goal is to estimate solutions to local estimation equations of the form

$$\mathbb{E}\left[\psi_{\theta(x),\,\nu(x)}\left(O_i\right) \,\middle|\, X_i = x\right] = 0 \quad \text{for all} \quad x \in \mathcal{X}, \tag{4}$$

where $\theta(x)$ is the parameter we care about and $\nu(x)$ is an optional nuisance parameter. This setting is general, and encompasses many important problems in statistics and econometrics.

A popular approach to estimating such functions $\theta(x)$ is to first define some kind of similarity weights $\alpha_i(x)$ that measure the relevance of the i-th training example to fitting $\theta(\cdot)$ at x, and then fit the target of interest via an empirical version of the estimating equation (Fan et al., 1998; Newey, 1994b; Staniswalis, 1989; Stone, 1977; Tibshirani and Hastie, 1987):

$$\left(\hat{\theta}(x),\, \hat{\nu}(x)\right) \in \text{argmin}_{\theta,\,\nu} \left\{ \left\| \sum_{i=1}^{n} \alpha_i(x)\, \psi_{\theta,\,\nu}\left(O_i\right) \right\|_2 \right\}. \tag{5}$$

When the above expression has a unique root, we can simply say that $\hat{\theta}(x)$, $\hat{\nu}(x)$ solves $\sum_{i=1}^{n} \alpha_i(x)\, \psi_{\hat{\theta}(x),\,\hat{\nu}(x)}\left(O_i\right) = 0$. The weights $\alpha_i(x)$ used to specify the above solution to the heterogeneous estimating equation are traditionally obtained via a deterministic kernel function, perhaps with an adaptively chosen bandwidth parameter (Hastie et al., 2009).

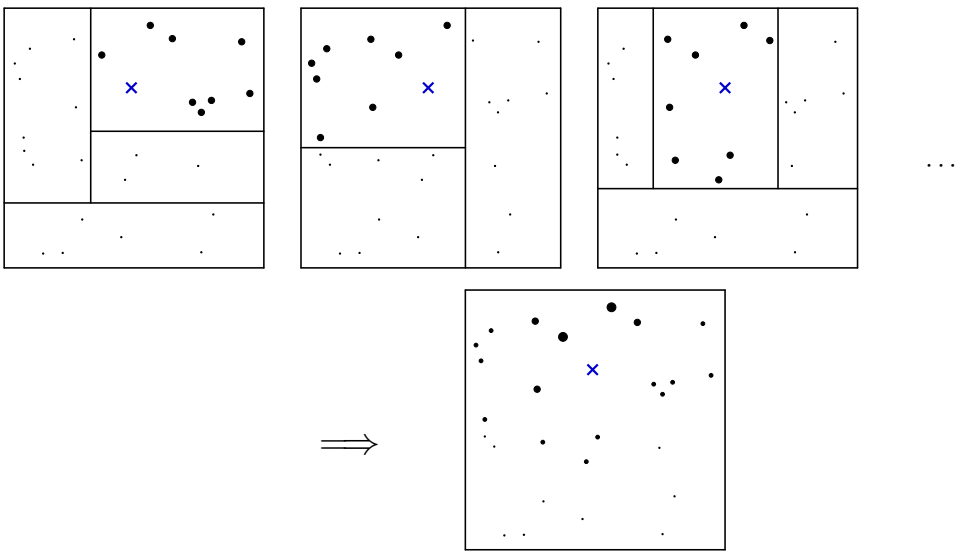


Figure 2: Illustration of the random forest weighting function. The rectangles depticted above correspond to terminal nodes in the dendogram representation of Figure 1. Each tree starts by giving equal (positive) weight to the training examples in the same leaf as our test point x of interest, and zero weight to all the other training examples. Then, the forest averages all these tree-based weightings, and effectively measures how often each training example falls into the same leaf as x.

Although methods of the above kind often work well in low dimensions, they can be very sensitive to the curse of dimensionality.

In this paper, we seek to use forest-based algorithms to adaptively learn better, problem-specific, weights $\alpha_i(x)$ that can be used in conjunction with (5). As in Meinshausen (2006), our generalized random forests obtain such weights by averaging neighborhoods implicitly produced by different trees. First, we grow a set of B trees indexed by $b = 1, ..., B$ and, for each such tree, define $L_b(x)$ as the set of training examples falling in the same "leaf" as x. The weights $\alpha_i(x)$ then capture the frequency with which the i-th training example falls into the same leaf as x:

$$\alpha_{bi}(x) = \frac{\mathbf{1}\left(\{X_i \in L_b(x)\}\right)}{|L_b(x)|}, \quad \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{bi}(x). \tag{6}$$

These weights sum to 1, and define the forest-based adaptive neighborhood of x; see Figure 2 for an illustration of this weighting function. There are of course some subtleties in how the sets $L_b(x)$ are defined—in particular, as discussed in Section 2.5 below, our construction will rely on both subsampling and a specific form of sample-splitting to achieve consistency—but at a high level the estimates $\hat{\theta}(x)$ produced by a generalized random forests are simply obtained by solving (5) with weights (6).

Finally, as claimed above, our present weighting-based definition of a random forest is equivalent to the standard "average of trees" perspective taken in Breiman (2001) for the

special case of regression trees. Specifically, suppose we want to estimate the conditional mean function $\mu(x) = \mathbb{E}\left[Y_i \,\middle|\, X_i = x\right]$, which is identified in (4) using the moment function $\psi_{\mu(x)}(Y_i) = Y_i - \mu(x)$. Then, we can use simple algebra to verify that

$$\sum_{i=1}^{n} \frac{1}{B} \sum_{b=1}^{B} \alpha_{bi}(x) \left(Y_i - \hat{\mu}(x)\right) = 0 \quad \Longleftrightarrow \quad \hat{\mu}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{\mu}_b(x), \tag{7}$$

where $\hat{\mu}_b(x) = \sum_{\{i: X_i \in L_b(x)\}} Y_i \,/\, |L_b(x)|$ is the prediction made by a single CART regression tree, and so we in fact recover the simple averaging definition (3) of a regression forest.

2.3 Splitting to Maximize Heterogeneity

Given this setup, we need to build trees that, when combined into a forest, induce weights $\alpha_i(x)$ that lead to good estimates of $\theta(x)$. The main substantive difference between the random forests of Breiman (2001) relative to other non-parametric regression techniques is their use of recursive partitioning on subsamples to generate these weights $\alpha_i(x)$. Motivated by the empirical success of regression forests across several application areas, our approach here is to mimic the algorithm of Breiman (2001) as closely as possible, while tailoring our splitting scheme to focus on heterogeneity in the target functional $\theta(x)$.

Just like in Breiman's forests, our search for good splits proceeds greedily, i.e., we seek splits that immediately improve the quality of the tree fit as much as possible. Every split starts with a parent node $P \subseteq \mathcal{X}$; given a sample of data $\mathcal{J}$, we define $(\hat{\theta}_P,\, \hat{\nu}_P)(\mathcal{J})$ as the solution to the estimating equation, as follows (where we suppress the dependence on the sample $\mathcal{J}$ in cases where it is unambiguous):

$$\left(\hat{\theta}_P,\, \hat{\nu}_P\right)(\mathcal{J}) \in \text{argmin}_{\theta,\,\nu} \left\{ \left\| \sum_{\{i \in \mathcal{J}: X_i \in P\}} \psi_{\theta,\,\nu}\left(O_i\right) \right\|_2 \right\}. \tag{8}$$

We would like to divide P into two children $C_1,\, C_2 \in \mathcal{X}$ using an axis-aligned cut such as to improve the accuracy of our θ-estimates as much as possible. That is, we wish to evaluate a given split using the expectation of the squared error, where the expectation is taken with respect to the uncertainty that arises when we estimate the parameters using a random sample $\mathcal{J}$ and evaluate the squared error at a random evaluation point X:

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

where the $\hat{\theta}_{C_j}(\mathcal{J})$ are fit over the child nodes C_j in analogy to (8).

Many standard regression tree implementations, such as CART (Breiman et al., 1984), choose their splits by simply minimizing the in-sample prediction error of the node, which corresponds to (9) with plug-in estimators from the training sample. Athey and Imbens (2016) study "honest" trees, where one sample is used to select the splits but a distinct, independent sample $\mathcal{J}$ is used to estimate $(\hat{\theta}_{C_1}, \hat{\theta}_{C_2})(\mathcal{J})$; this motivates an explicit treatment of $\mathcal{J}$ as a random variable. They show that for the case of "causal trees" (estimating the effect of a binary treatment), an unbiased, model-free estimate of (9) is available, one which reduces to adjusting the in-sample squared error in treatment effects by a correction that accounts for the way in which the splits affect the variance of the parameter estimates when re-estimated in new samples $\mathcal{J}$, in the spirit of Mallows (1973).

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

Proposition 1. *Suppose that basic assumptions detailed in Section 4 hold, and that the parent node P has a radius smaller than r for some value $r > 0$. Fix a training sample $\mathcal{J}^{tr}$. We write $n_P = |\{i \in \mathcal{J}^{tr} : X_i \in P\}|$ for the number of observations in the parent and n_{C_j} for the number of observations in each child, and define*

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

where $\hat{\theta}_{C_1}$ and $\hat{\theta}_{C_2}$ are solutions to the estimating equation computed in the children, following (8). Then, treating the children C_1 and C_2 as well as the counts n_{C_1} and n_{C_2} as constant, and assuming that $n_{C_1}, n_{C_2} \gg r^{-2}$, we have

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

where $K(P)$ is a deterministic term that measures the purity of the parent node that does not depend on how the parent is split, and the o-term incorporates terms that depend on sampling variance.

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

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

2.4 The Gradient Tree Algorithm

The above discussion provides some helpful conceptual guidance on how to pick good splits. However, from a computational perspective, actually optimizing the criterion $\Delta(C_1, C_2)$ over all possible axis-aligned splits while explicitly solving for $\hat{\theta}_{C_1}$ and $\hat{\theta}_{C_2}$ in each candidate child using an analogue to (8) may be quite expensive.

To avoid this issue, we instead optimize an approximate criterion $\widetilde{\Delta}(C_1, C_2)$ built using gradient-based approximations for $\hat{\theta}_{C_1}$ and $\hat{\theta}_{C_2}$. For each child C, we use $\tilde{\theta}_C \approx \hat{\theta}_C$ as follows: We first compute A_P as any consistent estimate for the gradient of the expectation of the ψ-function, i.e., $\nabla \mathbb{E}[\psi_{\hat{\theta}_P, \hat{\nu}_P}(O) \,|\, X = x]$, and then set

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

where $\hat{\theta}_P$ and $\hat{\nu}_P$ are obtained by solving (8) once in the parent node, and ξ is a vector that picks out the θ-coordinate from the (θ, ν) vector. When the ψ-function itself is continuously differentiable, we use

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

and, in this case, the quantity $\xi^\top A_P^{-1} \psi_{\hat{\theta}_P, \hat{\nu}_p}(O_i)$ corresponds to the influence function of the i-th observation for computing $\hat{\theta}_P$ in the parent. Cases where ψ is non-differentiable, e.g., with quantile regression, require more care.

Algorithmically, our recursive partitioning scheme now reduces to alternatively applying the following two steps. First, in a **labeling step**, we compute $\hat{\theta}_P$, $\hat{\nu}_P$, and the derivative matrix A_P^{-1} on the parent data as in (8), and use them to get pseudo-outcomes

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

Next, in a **regression step**, we run a standard CART regression split on the pseudo-outcomes ρ_i. Specifically, we split P into two axis-aligned children C_1 and C_2 such as to maximize the criterion

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

Finally, once we have executed the regression step, we relabel the observations in each child by solving the estimating equation, and continue on recursively.

To gain more intuition about this method, it is helpful to examine what it does in the simplest case of least-squares regression, i.e., with $\psi_\theta(x)(Y) = Y - \theta(x)$. Here, the labeling step (14) doesn't change anything—we get $\rho_i = Y_i - \overline{Y}_p$, where $\overline{Y}_p$ is the mean outcome in the parent—while the second step maximizing (15) corresponds to the usual way of making splits as in Breiman (2001). In other words, the special structure of the type of problem we are trying to solve is encoded in (14), while the second scanning step is a universal step shared across all different types of forests.

From a computational perspective, we expect this approach to provide more consistent performance than optimizing (10) at each split directly. When growing a tree, the computation is typically dominated by the split-selection step, and so it is critical for this step to be implemented as efficiently as possible (conversely, the labeling step (14) is only solved once per node, and so is less performance sensitive). From this perspective, using a regression splitting criterion as in (15) is very desirable, as it is possible to evaluate all possible split points along a given feature with only a single pass over the data in the parent node (by representing the criterion in terms of cumulative sums). In contrast, directly optimizing the original criterion (10) may require solving more intricate optimization problems for each possible candidate split.

This type of gradient-based approximation also underlies other popular statistical algorithms, including gradient boosting (Friedman, 2001) and the model-based recursive partitioning algorithm of Zeileis et al. (2008); conceptually, it is closely related to standard techniques for moment-based change-point detection (Andrews, 1993; Hansen, 1992; Hjort and Koning, 2002; Nyblom, 1989; Ploberger and Krämer, 1992; Zeileis, 2005; Zeileis and Hornik, 2007). Finally, in our context, we can verify that the error from using the approximate criterion (15) instead of the exact Δ-criterion (10) is within the tolerance used to

Algorithm 1 Generalized random forest with honesty and subsampling

Note: All tuning parameters, such as the total number of trees B and the sub-sampling s rate used in Subsample, are taken as pre-specified. This function is implemented in the package `grf` for `R` and `C++`.

procedure GeneralizedRandomForest(set of examples $\mathcal{S}$, test point x)
  weight vector $\alpha \leftarrow$ Zeros($|\mathcal{S}|$)
  for $b = 1$ to total number of trees B **do**
    set of examples $\mathcal{I} \leftarrow$ Subsample($\mathcal{S}$, s)
    sets of examples $\mathcal{J}_1$, $\mathcal{J}_2 \leftarrow$ SplitSample($\mathcal{I}$)
    tree $\mathcal{T} \leftarrow$ GradientTree($\mathcal{J}_1$) ▷ Grows a tree by recursive partitioning, alternating the steps (14) and (15).
    $\mathcal{N} \leftarrow$Neighbors(x, $\mathcal{T}$, $\mathcal{J}_2$) ▷ Returns those elements of $\mathcal{J}_2$ that fall into the same leaf as x in the tree $\mathcal{T}$.
    for all example $e \in \mathcal{N}$ **do**
      $\alpha[e]$ += $1/|\mathcal{N}|$
  output $\hat{\theta}(x)$, the solution to (5) with weights α/B

The function call Zeros creates a vector of zeros of length $|\mathcal{S}|$; Subsample draws a sub-sample of size s from $\mathcal{S}$ without replacement; and SplitSample randomly divides a set into two evenly-sized, non-overlapping halves.
The final step (5) can be solved using any numerical estimator. Our implementation `grf` provides an explicit plug-in point where a user can write a solver for (5) that is appropriate for their ψ-function of interest.

motivate the Δ-criterion in Proposition 1, thus suggesting that our use of (12) to guide splitting may not result in too much inefficiency.

Proposition 2. *Under the conditions of Proposition 1, suppose that A_p is consistent, i.e., $A_p \to_p \nabla \mathbb{E}[\psi_{\hat{\theta}_P, \hat{\nu}_P}(O) \,|\, X = x]$. Then,*

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

2.5 Building a Forest with Theoretical Guarantees

Now, given a practical splitting scheme for growing individual trees, we want to grow a forest that allows for consistent estimation of $\theta(x)$ using (5) paired with the forest weights (6). At a high level, we expect each tree to provide small, relevant neighborhoods for x that give us noisy estimates of $\theta(x)$. Then, if every tree has different small, relevant neighborhoods for x, we may hope that forest-based aggregation will provide a single larger but still relevant neighborhood for x that yields stable estimates $\hat{\theta}(x)$.

To ensure that forest-based aggregation succeeds in providing such stable, consistent parameter estimates, we rely on two conceptual ideas that have proven to be successful in the literature on forest-based least-squares regression: Training trees on subsamples of the training data (Mentch and Hooker, 2016; Scornet et al., 2015; Wager and Athey, 2017), and a sub-sample splitting technique that we call honesty (Biau, 2012; Denil et al., 2014; Wager and Athey, 2017). Our final algorithm for forest-based solutions to heterogeneous estimating

equations is given as Algorithm 1; we refer to Section 2.4 of Wager and Athey (2017) for a more in-depth discussion of honesty in the context of forests.

As we will show in the theoretical analysis in Section 4, assuming regularity conditions, the estimates $\hat{\theta}(x)$ obtained using a generalized random forest as described in Algorithm 1 are consistent for $\theta(x)$. Moreover, given appropriate subsampling rates, we establish asymptotic normality of the resulting forest estimates $\hat{\theta}(x)$.

3 Interlude: Quantile Regression Forests

Before developing the asymptotics of generalized random forests below, we take a brief pause to illustrate how the formalisms described above play out in the case of quantile regression. This problem has also been considered in detail by Meinshausen (2006), who proposed a consistent forest-based quantile regression algorithm; his method also fits into the paradigm of solving estimating equations (5) using random forest weights (6). However, unlike us, Meinshausen (2006) does not propose a splitting rule that is tailored to the quantile regression context, and instead builds his forests using plain CART regression splits. Thus, a comparison of our method with that of Meinshausen (2006) provides a perfect opportunity for evaluating the value of our proposed method for constructing forest-based weights $\alpha_i(x)$ that are specifically designed to express heterogeneity in conditional quantiles.

Recall that, in the language of estimating equations, the q-th quantile $\theta_q(x)$ of the distribution of Y conditionally on $X = x$ is identified via (4), using the moment function

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

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

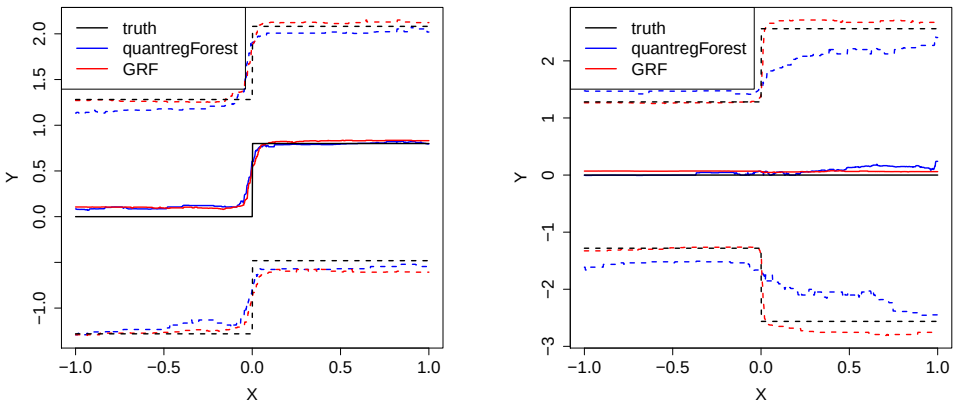
$$\rho_i = \mathbf{1}\left(\left\{Y_i > \hat{\theta}_{q,P}\right\}\right) \text{ where } \hat{\theta}_{q,P} \text{ is the } q\text{-th quantile of the parent } P, \tag{18}$$

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

We compare our method to that of Meinshausen (2006) in Figure 3. In the left panel, we have a mean shift in the distribution of Y_i conditional on X_i at $(X_i)_1 = 0$, and both methods are able to pick it up as expected. However, in the right panel, the mean of Y given X is constant, but there is a scale shift at $(X_i)_1 = 0$. Here, our method still performs well, as our splitting rule targets changes in the quantiles of the Y-distribution. However, the method of Meinshausen (2006) breaks down completely, as it relies on CART regression splits that are only sensitive to changes in the conditional mean of Y given X.

We also note that generalized random forests produce somewhat smoother sample paths than the method of Meinshausen (2006); this is due to our use of honesty as described in Section 2.5. If we run generalized random forests without honesty, then our method still correctly identifies the jumps at $x = 0$, but has sample paths that oscillate locally just as much as the baseline method.

Finally, the purpose of this example is not to claim that our variant of quantile regression forests built using gradient trees is always superior to the method of Meinshausen (2006) that uses regression-based splitting to obtain the weights $\alpha_i(x)$. Rather, we have shown that, as claimed, our splitting rule is specifically sensitive to quantile shifts in a way that regression splits are not—and, moreover, deriving the splitting rule with (18) was fully automatic given



mean shift scale shift

Figure 3: Comparison of quantile regression using generalized random forests and the `quantregForest` package of Meinshausen (2006). In both cases, we have $n = 2,000$ independent and identically distributed examples where X_i is uniformly distributed over $[-1, 1]^p$ with $p = 40$, and Y_i is Gaussian conditionally on $(X_i)_1$: in the left panel, $Y_i \,\big|\, X_i \sim \mathcal{N}\left(0.8 \cdot \mathbf{1}\left(\{(X_i)_1 > 0\}\right), 1\right)$, while in the right panel $Y_i \,\big|\, X_i \sim \mathcal{N}(0, (1 + \mathbf{1}\left(\{(X_i)_1 > 0\}\right))^2)$. The other 39 covariates are noise. We estimate the quantiles at $q = 0.1, 0.5, 0.9$.

our generalized random forest formalism. Of course, there are presumably applications where changes in the conditional mean are good proxies for changes in the conditional quantiles; and, in these cases, we might expect the method of Meinshausen (2006) to work very well. However, if we want to make sure our splitting rule is always sensitive to changes at the quantile of interest, our gradient-based approach may be more robust.

Remark (Estimating many quantiles)**.** In many cases, we want to estimate multiple quantiles at the same time; for example, in Figure 3, we sought to get $q = 0.1, 0.5, 0.9$ at the same time. Estimating different forests for each quantile separately would be undesirable for many reasons: It would be computationally expensive and, moreover, there is a risk that quantile estimates might cross in finite samples due to statistical noise. Thus, we need to build a forest using a splitting scheme that is sensitive to changes at any of our quantiles of interests. Here, we use a simple heuristic inspired by the relabeling transformation (18). Given a set of quantiles of interest $q_1 < ... < q_k$, we first evaluate all these quantiles $\hat{\theta}_{q_1,P} \leq ... \leq \hat{\theta}_{q_k,P}$ in the parent node, and label i-th point by the interval $[\hat{\theta}_{q_{j-1},P}, \hat{\theta}_{q_j,P})$ it falls into. Then, we choose the split point using a multiclass classification rule that classifies each observation into one of the intervals.

4 Asymptotic Analysis

4.1 Theoretical Setup

We now turn to a formal characterization of generalized random forests, with the aim of establishing asymptotic Gaussianity of the estimates $\hat{\theta}(x)$, and of providing tools for statistical inference about $\theta(x)$. We begin by listing the basic assumptions underlying all of our theoretical results. First, we assume that the covariate space and the parameter space are both subsets of Euclidean space; specifically, we assume that $\mathcal{X} = [0, 1]^p$ and $(\theta, \nu) \in \mathcal{B} \subset \mathbb{R}^k$ for some $p, k > 0$, where $\mathcal{B}$ is a compact subset of $\mathbb{R}^k$. Moreover, we assume that the features X have a density that is bounded away from 0 and ∞; as argued in, e.g., Wager and Walther (2015), this is equivalent to imposing a weak dependence condition on the individual features $(X_i)_j$ because trees and forests are invariant to monotone rescaling of the features.

Some practically interesting cases, such as quantile regression, involve discontinuous score functions ψ, which makes the analysis considerably more intricate. Here, we follow standard practice, and assume that the *expected* score function

$$M_{\theta,\,\nu}(x) := \mathbb{E}\left[\psi_{\theta,\,\nu}(O) \,\middle|\, X = x\right] \tag{19}$$

vary smoothly in the parameters, even though ψ itself may be discontinuous. For example, in the case of quantile regression $\psi_\theta(Y) = 1\left(\{Y > \theta\}\right) - (1 - q)$ is discontinuous in q, but $M_\theta(x) = \mathbb{P}\left[Y > \theta \,\middle|\, X = x\right] - (1 - q)$ will be smooth whenever $Y \,\middle|\, X = x$ has a smooth density.

Assumption 1 (Lipschitz x-signal)**.** For fixed values of (θ, ν), we assume that $M_{\theta,\,\nu}(x)$ as defined in (19) is Lipschitz continuous in x.

Assumption 2 (Smooth identification)**.** When x is fixed, we assume that this M-function is twice continuously differentiable in (θ, ν) with a uniformly bounded second derivative. Moreover, writing the derivative of $M_{\theta,\,\nu}(x)$ as

$$V_{\theta,\,\nu}(x) := \frac{\partial}{\partial(\theta,\,\nu)} M_{\theta,\,\nu}(x) \,\big|_{\,\theta(x),\;\nu(x)}, \tag{20}$$

we assume that $V(x) := V_{\theta(x),\;\nu(x)}(x)$ is invertible for all $x \in \mathcal{X}$.

Our next two assumptions control regularity properties of the ψ-function itself. Assumption 3 holds trivially when ψ itself is Lipschitz in (θ, ν) (in fact, having ψ be 0.5-Hölder would be enough); and also holds for quantile regression with the constant $L = \sup_{x,\,y} f_x(y)$ referred to in the results below, where $f_x(y)$ is the density of $Y \,\big|\, X = x$. Assumption 4 is used to show that a certain empirical process is Donsker and clearly holds for all our cases of interest; see the Section 4.3 for details. Finally, Assumption 5 can be taken to hold with $C = 0$ when ψ is continuous and non-degenerate, and with $C = 1/2$ for quantile regression.

Assumption 3 (Lipschitz (θ, ν)-variogram)**.** We assume that the the score functions $\psi_{\theta,\,\nu}(O_i)$ have a continuous covariance structure. Writing γ for the worst-case variogram

$$\gamma\left(\begin{pmatrix}\theta\\ \nu\end{pmatrix}, \begin{pmatrix}\theta'\\ \nu'\end{pmatrix}\right) := \sup_{x \in \mathcal{X}} \left\{ \left\| \mathrm{Var}\left[\psi_{\theta,\,\nu}\left(O_i\right) - \psi_{\theta',\,\nu'}\left(O_i\right)\right] \,\middle|\, X_i = x \right\|_F \right\}, \tag{21}$$

we assume that this variogram is Lipschitz, i.e. for all $(\theta, \nu), (\theta', \nu')$

$$\gamma\left(\begin{pmatrix}\theta\\ \nu\end{pmatrix}, \begin{pmatrix}\theta'\\ \nu'\end{pmatrix}\right) \leq L \left\| \begin{pmatrix}\theta\\ \nu\end{pmatrix} - \begin{pmatrix}\theta'\\ \nu'\end{pmatrix} \right\|_2, \tag{22}$$

for some constant $L > 0$.

Assumption 4 (Regularity of ψ)**.** The ψ-functions can be written as

$$\psi_{\theta,\,\nu}(O) = \lambda\left(\theta,\,\nu;\,O_i\right) + \zeta_{\theta,\,\nu}\left(g(O_i)\right),$$

such that λ is Lipschitz-continuous in $(\theta,\,\nu)$, $g : \{O_i\} \to \mathbb{R}$ is a univariate summary of O_i, $\zeta_{\theta,\,\nu} : \mathbb{R} \to \mathbb{R}$ is any family of monotone and bounded functions.

Assumption 5 (Existence of solutions)**.** We assume that, for any weights α_i with $\sum \alpha_i = 1$, the estimating equation (5) returns a minimizer $(\hat{\theta},\,\hat{\nu})$ that at least approximately solves the estimating equation:

$$\left\| \sum_{i=1}^{n} \alpha_i\, \psi_{\hat{\theta},\,\hat{\nu}}\left(O_i\right) \right\|_2 \leq C \max\left\{\alpha_i\right\}, \text{ for some constant } C \in \mathbb{R}_+. \tag{23}$$

All the previous assumptions only deal with local properties of the estimating equation, and can be used to control the behavior of $(\hat{\theta}(x),\,\hat{\nu}(x))$ in a small neighborhood of the population parameter value $(\theta(x),\,\nu(x))$. Now, to make any use of these assumptions, we first need to verify that $(\hat{\theta}(x),\,\hat{\nu}(x))$ be consistent. Here, we use the following assumption to guarantee consistency; this setup is general enough to cover both instrumental variables regression and quantile regression.

Assumption 6 (Convexity)**.** The score function $\psi_{\theta,\,\nu}(O_i)$ is a sub-gradient of a convex function, and the expected score $M_{\theta,\,\nu}(X_i)$ is the gradient of a strongly convex function.

Finally, our consistency and Gaussianty results require some control on the behavior of the trees comprising the forest. To do so, we follow Wager and Athey (2017), as follows.

Assumption 7 (Regular and honest forest)**.** We assume that our trees are symmetric, in that their output is invariant to permuting the indices of the training examples. We also assume that the tree makes balanced splits, in the sense that every split puts at least a fraction ω of the observations in the parent node into each child, for some $\omega > 0$, take the tree to be randomized in such a way that, at every split, the probability that the tree splits on the j-th feature is bounded from below by some $\pi > 0$.[3] Finally, we assume that our forest is honest and built via subsampling with subsample size s satisfying $s/n \to 0$ and $s \to \infty$, as described in Section 2.5.

In the interest of generality, we set up Assumptions 1–6 in a rather abstract way effort. We end this section by showing that, in the context of our main problems of interest requiring Assumptions 1–6 is not particularly stringent. Further examples that satisfy the above assumptions will be discussed in Sections 6 and 7.

Example 1 (Least squares regression)**.** In the case of least-squares regression, i.e., $\psi_{\theta(x)}(Y_i) = Y_i - \theta(x)$, Assumptions 2–6 hold immediately from the definition of ψ. Meanwhile, Assumption 1 simply means that the conditional mean function $\mathbb{E}\left[Y_i \,\middle|\, X_i = x\right]$ must be Lipschitz in x; this is a standard assumption in the literature on regression forests.

[3] In our implementation, we satisfy this condition by using the device of Denil et al. (2014), whereby the number splitting variables considered at each step of the algorithm is random; specifically, we try $\min\left\{\max\left\{\text{Poisson}(m),\,1\right\},\,p\right\}$ variables at each step, where $m > 0$ is a tuning parameter.

Example 2 (Quantile regression)**.** For quantile regression, Assumption 1 is equivalent to assuming that the conditional exceedance probabilities $\mathbb{P}\left[Y_i > y \,\middle|\, X_i = x\right]$ be Lipschitz-continuous in x for all $y \in \mathbb{R}$. Further, Assumption 2 holds if the conditional density $f_x(y)$ has a continuous uniformly bounded first derivative, and is bounded away from 0 at the quantile of interest $y = F_x^{-1}(q)$, where $F_x(y)$ denotes the cumulative distribution function of Y conditionally on $X = x$; and Assumption 3 holds if $f_x(y)$ is uniformly bounded from above. Assumptions 4–6 hold from the definition of ψ.

4.2 A Central Limit Theorem for Generalized Random Forests

Given these assumptions, we are now ready to provide an asymptotic characterization of generalzed random forests. In doing so, we note that existing asymptotic analyses of regression forests, including Mentch and Hooker (2016), Scornet et al. (2015) and Wager and Athey (2017), were built around the fact that regression forests are averages of regression trees grown over sub-samples, and can thus be analyzed as U-statistics (Hoeffding, 1948). Unlike regression forest predictions, however, the parameter estimates $\hat{\theta}(x)$ provided by generalized random forests are not averages of estimates made by different trees; instead, we obtain $\hat{\theta}(x)$ by solving a single weighted moment equation as in (5). Thus, existing proof strategies do not apply in our setting.

We tackle this problem using the method of influence functions as described by Hampel (1974); in particular, we are motivated by the analysis of Newey (1994b). The core idea of these methods is to first derive a sharp, linearized approximation to the local estimator $\hat{\theta}(x)$, and then to analyze the linear approximation instead.

In our setup, the influence function heuristic motivates a natural approximation $\tilde{\theta}^*(x)$ to $\hat{\theta}(x)$ as follows. Let $\rho_i^*(x)$ denote the influence function of the i-th observation with respect to the true parameter value $\theta(x)$,

$$\rho_i^*(x) := -\xi^\top V(x)^{-1} \psi_{\theta(x),\,\nu(x)}(O_i). \tag{24}$$

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

Then, given any set of forest weights $\alpha_i(x)$ used to define the generalized random forest estimate $\hat{\theta}(x)$ by solving (5), we can also define a pseudo-forest

$$\tilde{\theta}^*(x) := \theta(x) + \sum_{i=1}^{n} \alpha_i \rho_i^*(x), \tag{25}$$

which we will use as an approximation for $\hat{\theta}(x)$. We note that, formally, this pseudo-forest estimate $\tilde{\theta}^*(x)$ is equivalent to the output of an (infeasible) regression forest with weights $\alpha_i(x)$ and outcomes $\theta(x) + \rho_i^*(x)$.

The upshot of this approximation is that, unlike $\hat{\theta}(x)$, the pseudo-forest $\tilde{\theta}^*(x)$ is a U-statistic. More specifically, because $\tilde{\theta}^*(x)$ is a linear function of the pseudo-outcomes $\rho_i^*(x)$, we can write it as an average of pseudo-tree predictions

$$\tilde{\theta}^*(x) = \frac{1}{B} \sum_{b=1}^{B} \tilde{\theta}_b^*(x), \quad \tilde{\theta}_b^*(x) = \sum_{i=1}^{n} \alpha_{ib} \left(\theta(x) + \rho_i^*(x)\right). \tag{26}$$

Then, because each individual pseudo-tree prediction $\tilde{\theta}_b^*(x)$ is trained on a subsample of the training data drawn without replacement (see Section 2.5), the arguments of Mentch and Hooker (2016) or Wager and Athey (2017) can be used to study the averaged estimator $\tilde{\theta}^*(x)$ using classical results about U-statistics (Hoeffding, 1948; Efron and Stein, 1981).

Following this proof strategy, the key difficulty is in showing that our influence-based statistic $\tilde{\theta}^*(x)$ is in fact a good approximation for $\hat{\theta}(x)$. The following result, which is the main technical contribution of this paper, establishes such a coupling provided the estimator $\hat{\theta}(x)$ itself is consistent for $\theta(x)$. Then, in Lemma 4 we guarantee consistency of $\hat{\theta}(x)$ assuming convexity of ψ as in Assumption 6. We note that separating the analysis of moment estimators into a local approximation argument that hinges on consistency and a separate result that establishes consistency is standard; see, e.g., Chapter 5.3 of Van der Vaart (2000).

Lemma 3. *Given Assumptions 1–5, and a forest trained according to 7, suppose that the generalized random forest estimator $\hat{\theta}(x)$ is consistent for $\theta(x)$. Then $\hat{\theta}(x)$ and $\tilde{\theta}^*(x)$ are coupled at the following rate:*

$$\sqrt{\frac{n}{s}}\left(\tilde{\theta}^*(x) - \hat{\theta}(x)\right) = \mathcal{O}_P\left(\max\left\{s^{-\frac{\pi}{2}\frac{\log\left((1-\omega)^{-1}\right)}{\log(\omega^{-1})}}, \left(\frac{s}{n}\right)^{\frac{1}{6}}\right\}\right). \tag{27}$$

Lemma 4. *Given Assumptions 1–6, the estimates $(\hat{\theta}(x), \hat{\nu}(x))$ from a generalized random forest trained according to Assumption 7 converge in probability to $(\theta(x), \nu(x))$ as $n \to \infty$.*

Given this coupling result, it now remains to study the asymptotics of $\tilde{\theta}^*(x)$. In doing so, we re-iterate that $\tilde{\theta}^*(x)$ is *exactly* the output of an infeasible regression forest trained on outcomes $\theta(x) + \rho_i^*(x)$. Thus, the results of Wager and Athey (2017) apply directly to this object, and can be used to establish its Gaussianity. The fact that we cannot actually compute $\tilde{\theta}^*(x)$ in practice does not hinder an application of their results.

Pursuing this approach, we find that whenever trees are grown on subsamples of size s scaling as $s = n^\beta$ for some $\beta_{\min} < \beta < 1$, $\tilde{\theta}^*(x)$—and thus also $\hat{\theta}(x)$—is asymptotically normal.

Theorem 5. *Suppose that Assumptions 1–6 hold, that our forest is trained according to Assumption 7, and moreover that trees are grown on subsamples of size s with*

$$s = n^\beta \quad \text{for some} \quad \beta_{\min} := 1 - \left(1 + \frac{1}{\pi}\frac{\log\left(\omega^{-1}\right)}{\log\left((1-\omega)^{-1}\right)}\right)^{-1} < \beta < 1, \tag{28}$$

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

$$\left(\hat{\theta}_n(x) - \theta(x)\right) / \sigma_n(x) \Rightarrow \mathcal{N}(0, 1), \quad \sigma_n^2(x) = \frac{1}{\mathrm{polylog}(n/s)}\frac{s}{n}, \tag{29}$$

where $\mathrm{polylog}(n/s)$ is a function that is bounded away from 0 and increases at most polynomially with the log-inverse sampling ratio $\log(n/s)$.

4.3 Proof of Main Results

Here, we present arguments leading up to our main result, namely the central limit theorem presented in Theorem 5, starting with some technical lemmas. Due to space considerations,

the proofs of Propositions 1 and 2, Lemma 4, and the technical results stated below are given in the appendix. Throughout our theoretical analysis, we use the following notation: Given our forest weights $\alpha_i(x)$ (6), let

$$\Psi(\theta, \nu) := \sum_{i=1}^{n} \alpha_i(x) \psi_{\theta, \nu}(O_i) \text{ and } \overline{\Psi}(\theta, \nu) := \sum_{i=1}^{n} \alpha_i(x) M_{\theta, \nu}(X_i). \tag{30}$$

We will frequently use the following bounds on the moments of Ψ at the true parameter value $(\theta(x), \nu(x))$.

Lemma 6. *Let $\alpha_i(x)$ be weights from a forest obtained as in Assumption 7, and suppose that the M-function is Lipschitz in x as in Assumption 1. Then, $\Psi(\theta(x), \nu(x))$ satisfies the following moment bounds:*

$$\left\|\mathbb{E}\left[\Psi(\theta(x), \nu(x))\right]\right\|_2 = \mathcal{O}\left(s^{-\frac{\pi}{2}\frac{\log\left((1-\omega)^{-1}\right)}{\log\left(\omega^{-1}\right)}}\right) \tag{31}$$

$$\left\|\text{Var}\left[\Psi(\theta(x), \nu(x))\right]\right\|_F = \mathcal{O}(s/n), \tag{32}$$

where s is the subsampling rate used when building our forest.

Proof. To establish these bounds, we start by expanding Ψ as

$$\Psi(\theta, \nu) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} \alpha_{bi}(x) \psi_{\theta, \nu}(O_i), \tag{33}$$

where the α_{bi} are the individual tree weights used to build the forest weights in (6). Now, $\Psi(\theta, \nu)$ is nothing but the output of a regression forest with response $\psi_{\theta, \nu}(O_i)$. Thus, given our assumptions about the moments of $\psi_{\theta, \nu}(O_i)$ and the fact that our trees are built via honest subsampling, these bounds follow directly from arguments made in Wager and Athey (2017). First, the proof of Theorem 3 of Wager and Athey (2017) shows that the weights $\alpha_i(x)$ are localized:

$$\mathbb{E}\left[\sup\left\{\|X_i - x\|_2 : \alpha_i(x) > 0\right\}\right] = \mathcal{O}\left(s^{-\frac{\pi}{2}\frac{\log\left((1-\omega)^{-1}\right)}{\log\left(\omega^{-1}\right)}}\right), \tag{34}$$

thus directly implying (31) thanks to Assumption 1. Meanwhile, because individual trees are grown on subsamples, we can verify that

$$\frac{n}{s} \text{Var}\left[\Psi(\theta(x), \nu(x))\right] \preceq \text{Var}\left[\sum_{i=1}^{n} \alpha_{bi}(x) \psi_{\theta, \nu}(O_i)\right] = \mathcal{O}(1), \tag{35}$$

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

4.3.1 Local Regularity of Forests

Before proving any of our main results, we need establish a result that gives us some control over the "sample paths" of Ψ. To do so, define the local discrepancy measure

$$\delta_\alpha\left((\theta,\,\nu)\,,\,(\theta',\,\nu')\right) = \Psi\left(\theta,\,\nu\right) - \overline{\Psi}\left(\theta,\,\nu\right) - \left(\Psi\left(\theta',\,\nu'\right) - \overline{\Psi}\left(\theta',\,\nu'\right)\right), \tag{36}$$

which describes how tightly the stochastic fluctuations of $\Psi - \overline{\Psi}$ are coupled for nearby parameter values $(\theta,\,\nu)$ and $(\theta',\,\nu')$. The following lemmas establish uniform local concentration of δ_α: First, in Lemma 7, we control the variogram of the forest, and then Lemma 8 establishes concentration of δ over small balls. Both proofs are given the appendix.

Lemma 7. *Let $(\theta,\,\nu)$ and $(\theta',\,\nu')$ be fixed pairs of parameters, and let $\alpha_i(x)$ be forest weights generated according to Assumption 7. Then, provided that Assumptions 1–3 hold,*

$$\mathbb{E}\left[\delta_\alpha\left((\theta,\,\nu)\,,\,(\theta',\,\nu')\right)\right] = 0, \quad \mathbb{E}\left[\left\|\delta_\alpha\left((\theta,\,\nu)\,,\,(\theta',\,\nu')\right)\right\|^2\right] \leq L\frac{s}{n}\left\|\begin{pmatrix}\theta\\ \nu\end{pmatrix} - \begin{pmatrix}\theta'\\ \nu'\end{pmatrix}\right\|_2, \tag{37}$$

where L is the Lipschitz parameter from (22).

Next, to generalize this concentration bound from a single point into a uniform bound, we will need some standard formalism from empirical process theory as presented in, e.g., van der Vaart and Wellner (1996). To do so, we start by defining a bracketing, as follows. For any pair of parameters $(\theta_-,\,\nu_-)$, $(\theta_+,\,\nu_+)$, define the bracket

$$\beta\left(\begin{pmatrix}\theta_-\\ \nu_-\end{pmatrix}, \begin{pmatrix}\theta_+\\ \nu_+\end{pmatrix}\right) := \left\{\begin{pmatrix}\theta\\ \nu\end{pmatrix} \in \mathcal{B} : \Psi\left(\theta_-,\,\nu_-\right) \leq \Psi\left(\theta,\,\nu\right) \leq \Psi\left(\theta_+,\,\nu_+\right)\right\}$$

for all realizations of Ψ, where the inequality is understood coordinate-wise; and define the radius r of the bracket in terms of the L_2-distance of the individual "ψ-trees" that comprise Ψ:

$$r^2\left(\beta\left(\begin{pmatrix}\theta_-\\ \nu_-\end{pmatrix}, \begin{pmatrix}\theta_+\\ \nu_+\end{pmatrix}\right)\right) := \mathbb{E}\left[\left\|\sum_{\{i:i\in\mathcal{J}_1\}} \alpha_i(x;\,\mathcal{J}_2)\left(\psi_{\theta_+,\,\nu_+}\left(O_i\right) - \psi_{\theta_-,\,\nu_-}\left(O_i\right)\right)\right\|_2^2\right], \tag{38}$$

where $\mathcal{J}_1$ and $\mathcal{J}_2$ are two disjoint half-subsamples as in Algorithm 1. For any $\varepsilon > 0$, the ε-bracketing number $N_{[]}(\varepsilon,\,\Psi,\,L_2)$ is the minimum number of brackets of radius at most ε required to cover $\mathcal{B}$.

Given this notation, our concentration bound for δ_α will depend on controlling this covering number. Specifically, we assume that there is a constant κ for which the bracketing entropy $\log N_{[]}$ is bounded by

$$\log\left(N_{[]}(\varepsilon,\,\Psi,\,L_2)\right) \leq \frac{\kappa}{\varepsilon} \quad \text{for all} \quad 0 < \varepsilon \leq 1. \tag{39}$$

We use Assumption 4 to give us bounds of this type; and, in fact, this is the only place we use Assumption 4. Replacing Assumption 4 with (39) would be enough to prove our results, which will only depend on this assumption through Lemma 8 below.

To see how Assumption 4 leads to (39), we first write

$$\Psi\left(\theta,\,\nu\right) = \Psi^\lambda\left(\theta,\,\nu\right) + \Psi^\zeta\left(\theta,\,\nu\right),$$

where Ψ^λ is Lipschitz and Ψ^ζ is a monotone function of a univariate representation of O_i. Writing analogously $N_{[]}(\varepsilon, \Psi^\lambda, L_2)$ and $N_{[]}(\varepsilon, \Psi^\zeta, L_2)$ for the bracketing numbers of these two additive components on their own, we can verify that

$$\log\left(N_{[]}(\varepsilon, \Psi, L_2)\right) \leq \log\left(N_{[]}(\varepsilon/2, \Psi^\lambda, L_2)\right) + \log\left(N_{[]}(\varepsilon/2, \Psi^\zeta, L_2)\right).$$

Because Ψ^ζ is a bounded, monotone, univariate family, Theorem 2.7.5 of van der Vaart and Wellner (1996) implies that $\log N_{[]}(\varepsilon, \Psi^\lambda, L_2) = \mathcal{O}(1/\varepsilon)$. Meanwhile, because Ψ^λ is Lipschitz and our parameter space $\mathcal{B}$ is compact, Lemma 2.7.11 of van der Vaart and Wellner (1996) implies that $\log N_{[]}(\varepsilon, \Psi^\lambda, L_2) = \mathcal{O}\left(\log \varepsilon^{-1}\right)$. Thus, both terms are controlled at the desired order, and so (39) holds.

Lemma 8. *Under the conditions of Lemma 7, suppose moreover that (39) holds. Then,*

$$\begin{aligned}\mathbb{E}\left[\sup_{(\theta', \nu')}\left\{\left\|\delta_\alpha\left((\theta, \nu), (\theta', \nu')\right)\right\|_2 : \left\|\begin{pmatrix}\theta - \theta' \\ \nu - \nu'\end{pmatrix}\right\|_2 \leq \eta\right\}\right] \\ = \mathcal{O}\left(\sqrt{\frac{\kappa L \eta}{\lfloor n/s \rfloor}} + \frac{8\kappa G}{\lfloor n/s \rfloor L\eta}\right),\end{aligned} \tag{40}$$

for any $\eta > 0$ and $1 \leq s \leq n$, where G is an upper bound for $\left\|\psi_{\theta, \nu}(O_i) - \psi_{\theta', \nu'}(O_i)\right\|_\infty \leq G$; note that Assumption 4 guarantees that a finite bound G exists.

4.3.2 Proof of Lemma 3

We now turn to proving one of our key results. We first note that, if $\psi_{\theta, \nu}(O_i)$ were twice differentiable in (θ, ν), then we could verify (27) fairly directly via Taylor expansion of ψ. Now, of course, ψ is not twice differentiable, and so we cannot apply this argument directly. Rather, we need to first apply a Taylor expansion on the expected ψ function, $M_{\theta, \nu}(X_i)$, which *is* twice differentiable; we then use the regularity properties established in Section 4.3.1 to extend this result to ψ.

Now, recall that we have assumed $(\hat{\theta}(x), \hat{\nu}(x))$ to be consistent; thus, there is a sequence $\varepsilon_n \to 0$ such that

$$\left\|\begin{pmatrix}\hat{\theta}(x) - \theta(x) \\ \hat{\nu}(x) - \nu(x)\end{pmatrix}\right\|_2 = \mathcal{O}_P(\varepsilon_n). \tag{41}$$

Using notation established in (30) and (36), we then write

$$\begin{aligned}&\overline{\Psi}\left(\hat{\theta}(x), \hat{\nu}(x)\right) - \overline{\Psi}\left(\theta(x), \nu(x)\right) \\ &\quad = \Psi\left(\hat{\theta}(x), \hat{\nu}(x)\right) - \Psi\left(\theta(x), \nu(x)\right) - \delta_\alpha\left(\left(\theta(x), \nu(x)\right), \left(\hat{\theta}(x), \hat{\nu}(x)\right)\right).\end{aligned} \tag{42}$$

By the assumed smoothness of the moment functions, we know that $\overline{\Psi}$ is twice differentiable in (θ, ν) with a bound on the second derivative that holds uniformly over all realizations of $\alpha_i(x)$ and X_i, and so we can take a Taylor expansion:

$$\overline{\Psi}\left(\hat{\theta}(x), \hat{\nu}(x)\right) - \overline{\Psi}\left(\theta(x), \nu(x)\right) = \left(\sum_{i=1}^n \alpha_i(x) \nabla M_{\theta(x), \nu(x)}(X_i)\right)\begin{pmatrix}\hat{\theta}(x) - \theta(x) \\ \hat{\nu}(x) - \nu(x)\end{pmatrix} + H$$

with $\|H\| \leq c\varepsilon_n^2/2$, where c is the uniform bound on the curvature of M required in Assumption 2. Moreover, because the weights $\alpha_i(x)$ are localized as in (34),

$$\left\| \sum_{i=1}^{n} \alpha_i(x) \nabla M_{\theta(x),\,\nu(x)}(X_i) - V(x) \right\|_F = \mathcal{O}_P\left(s^{-\frac{\pi}{2}\frac{\log\left((1-\omega)^{-1}\right)}{\log\left(\omega^{-1}\right)}} \right), \tag{43}$$

where $s \to \infty$ is the sub-sample size used to grow trees in the forest. This expansion suggests that (42) should be helpful in relating our quantities of interest.

It now remains to bound the extraneous terms. By Assumption 5, we know that

$$\Psi\left(\hat{\theta}(x),\, \hat{\nu}(x)\right) \leq C \max_{1 \leq i \leq n} \{\alpha_i\} \leq C\frac{s}{n}.$$

Next, by the consistency of $(\hat{\theta}(x),\, \hat{\nu}(x))$, we can apply Lemma 8 with "η" set to $\varepsilon_n^{2/3}$ to verify that

$$\left\| \delta_\alpha\left((\theta(x),\, \nu(x)),\, \left(\hat{\theta}(x),\, \hat{\nu}(x)\right)\right) \right\|_2 = \mathcal{O}_P\left(\max\left\{ \varepsilon_n^{1/3}\sqrt{\frac{s}{n}},\, \frac{s}{n\,\varepsilon_n^{2/3}} \right\}\right).$$

Thus, thanks to Assumption 2 which lets us invert $V(x)$, we conclude that

$$\begin{aligned} &\left\| \begin{pmatrix} \hat{\theta}(x) - \theta(x) \\ \hat{\nu}(x) - \nu(x) \end{pmatrix} + V(x)^{-1}\Psi\left(\theta(x),\, \nu(x)\right) \right\|_2 \\ &\qquad = \mathcal{O}_P\left(\max\left\{ s^{-\frac{\pi}{2}\frac{\log\left((1-\omega)^{-1}\right)}{\log\left(\omega^{-1}\right)}}\, \varepsilon_n,\, \varepsilon_n^2,\, \varepsilon_n^{1/3}\sqrt{\frac{s}{n}},\, \frac{s}{n\,\varepsilon_n^{2/3}} \right\}\right). \end{aligned} \tag{44}$$

Finally, recall that $\|\Psi\left(\theta(x),\, \nu(x)\right)\|_2^2 = \mathcal{O}_P\left(s/n\right)$ by Lemma 6. Thus, we can use the bound (44) to get stronger consistency guarantees, and in fact verify that $(\hat{\theta}(x),\, \hat{\nu}(x))$ must have been $\sqrt{s/n}$-consistent; and so, in particular, we can take (41) to hold with $\varepsilon_n = \sqrt{s/n}$. The desired result then follows directly from (44), noting that $\tilde{\theta}^*(x) = \theta(x) + \xi^\top V(x)^{-1}\Psi\left(\theta(x),\, \nu(x)\right)$.

4.3.3 Proof of Theorem 5

As argued in Section 4.2, $\tilde{\theta}^*(x)$ is formally equivalent to the output of a regression forest, and so we can directly apply Theorem 1 of Wager and Athey (2017). Given the assumptions made here, their result shows that

$$\left(\tilde{\theta}^*(x) - \theta(x)\right) \,/\, \sigma_n(x) \Rightarrow \mathcal{N}\left(0,\, 1\right), \quad \sigma_n^2(x) \to_p 0. \tag{45}$$

Moreover, from Theorem 5 and Lemma 7 of Wager and Athey (2017), we see that σ_n^2 scales as in (29). Given this central limit theorem, it only remains to show that the discrepancy between $\hat{\theta}(x)$ and $\tilde{\theta}^*(x)$ established in Lemma 3, decays faster than $\sigma_n(x)$. But, thanks to the consistency result from Lemma 4, the coupling result in Lemma 3 implies that

$$\frac{n}{s}\left(\tilde{\theta}^*(x) - \hat{\theta}(x)\right)^2 = \mathcal{O}_P\left(\max\left\{ s^{-\pi\frac{\log\left((1-\omega)^{-1}\right)}{\log\left(\omega^{-1}\right)}},\, \sqrt[3]{\frac{s}{n}} \right\}\right),$$

and so $(\tilde{\theta}^*(x) - \hat{\theta}(x))/\sigma_n \to_p 0$.

5 Confidence Intervals via the Delta Method

Theorem 5 can also be used for statistical inference about $\theta(x)$. Given any consistent estimator $\hat{\sigma}_n(x)/\sigma_n(x) \to_p 1$ of the noise scale of $\hat{\theta}_n(x)$, Theorem 5 can be paired with Slutsky's lemma to verify that

$$\lim_{n\to\infty} \mathbb{P}\left[\theta(x) \in \left(\hat{\theta}_n(x) \pm \Phi^{-1}\left(1-\alpha/2\right)\hat{\sigma}_n(x)\right)\right] = \alpha. \tag{46}$$

Thus, in order to build asymptotically valid confidence intervals for $\theta(x)$ centered on $\hat{\theta}(x)$, it suffices to derive an estimator for $\sigma_n(x)$.

In order to do so, we again leverage coupling with our approximating pseudo-forest $\tilde{\theta}^*(x)$. In particular, the proof of Theorem 5 implies that $\mathrm{Var}[\tilde{\theta}^*(x)]/\sigma_n^2(x) \to_p 1$, and so it again suffices to study $\tilde{\theta}^*(x)$. Moreover, from the definition of $\tilde{\theta}^*(x)$, we directly see that

$$\begin{aligned} \mathrm{Var}\left[\tilde{\theta}^*(x)\right] &= \xi^\top V(x)^{-1} H_n(x;\, \theta(x),\, \nu(x)) (V(x)^{-1})^\top \xi, \\ H_n(x;\, \theta,\, \nu) &= \mathrm{Var}\left[\sum_{i=1}^n \alpha_i(x)\psi_{\theta,\,\nu}(O_i)\right]. \end{aligned} \tag{47}$$

Thus, we propose building confidence intervals as in (46) using

$$\hat{\sigma}_n^2(x) := \xi^\top \widehat{V}_n(x)^{-1} \widehat{H}_n(x) (\widehat{V}_n(x)^{-1})^\top \xi, \tag{48}$$

where $\widehat{V}_n(x)$ and $\widehat{H}_n(x)$ are any consistent estimators for the quantities defined in (47).

The first quanitity $V(x)$ is a problem specific curvature parameter, and is not directly linked to forest-based methods; in fact, it is the same quantity that is needed to estimate variance of classical local maximum likelihood methods following Newey (1994b). For example, for the instrumental variables problem described in Section 7,

$$V(x) = \begin{pmatrix} \mathbb{E}\left[Z_i W_i \,\middle|\, X_i = x\right] & \mathbb{E}\left[Z_i \,\middle|\, X_i = x\right] \\ \mathbb{E}\left[W_i \,\middle|\, X_i = x\right] & 1 \end{pmatrix}, \tag{49}$$

while for quantile regression, $V(x) = f_x(\theta(x))$. In both cases, several different strategies are available for estimating this term. In the case of instrumental variables forests, we suggest estimating the individual entries of (49) using (honest and regular) regression forests.

The more interesting term is the inner variance term $H_n(x;\, \theta(x),\, \nu(x))$. To study this quantity, we note that the forest score $\Psi(\theta(x),\, \nu(x)) = \sum_{i=1}^n \alpha_i(x)\psi_{\theta(x),\,\nu(x)}(O_i)$ is again formally equivalent to the output of a regression forest with weights $\alpha_i(x)$, this time with effective outcomes $\psi_{\theta(x),\,\nu(x)}(O_i)$. A number of proposals have emerged for estimating the variance of a regression forest, including work by Sexton and Laake (2009), Mentch and Hooker (2016) and Wager et al. (2014); and, in principle, any of these methods could be adapted to estimate the variance of Ψ. The only difficulty is that Ψ depends on the true parameter values $(\theta(x),\, \nu(x))$, and so cannot directly be accessed in practice.

Given this setup, a natural idea is to use the estimates $(\hat{\theta}(x),\, \hat{\nu}(x))$ from the generalized random forest to define a feasible plug-in version of Ψ with $(\theta(x),\, \nu(x))$ replaced by $(\hat{\theta}(x),\, \hat{\nu}(x))$, and then apply standard regression forest inference tools to it. For example, an application of the infinitesimal jackknife for random forests from Wager et al. (2014) would suggest using (see also Efron (2014))

$$\widehat{H}_n^{(IJ)}(x) := \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2 \sum_{i=1}^n \mathrm{Cov}_{ss}\left[\bar{\psi}^{(b)}_{\hat{\theta}(x),\,\hat{\nu}(x)},\, N_{bi}\right]^{\otimes 2}, \tag{50}$$

where N_{bi} is an indicator for whether the i-the observation was used in the b-th subsample used to build the forest, $\bar{\psi}^{(b)}_{\theta,\nu} = \sum_{i=1}^{n} \alpha_{bi}(x)\psi_{\theta,\nu}(O_i)$ is the average of the ψ-function over the leaf containing x in the b-th tree, and Cov_{ss} denotes covariance with respect to the subsampling measure. If we could compute $\widehat{H}_n^{(IJ)}(x)$ with $(\hat{\theta}(x), \hat{\nu}(x))$ replaced by $(\theta(x), \nu(x))$ in (50), then the results of Wager and Athey (2017) would directly imply consistency of $\widehat{H}_n^{(IJ)}(x)$; but now we still need to worry about the plug-in argument (in an earlier draft of this paper we verified that this plug-in strategy is in fact valid for the instrumental variables forests defined in Section 7).

Instead of revisiting technical details from Wager and Athey (2017), however, we here develop our theory based on a variant of the bootstrap of little bags algorithm (or noisy bootstrap) proposed by Sexton and Laake (2009). As a side benefit, we also obtain the first consistency guarantees for this method for any type of forest, including regression forests. One of the main advantages of the infinitesimal jackknife was that, as emphasized by Wager et al. (2014), it operates only on outputs from an already trained forest, and so can be added "on top" of an optimized forest implementation without needing to modify the internal workings of the implementation. But now we already needed to develop our generalized random forest software, `grf`, from the ground up, so adding support for the method of Sexton and Laake (2009) deep within the forest was less of a concern.

5.1 Consistency of the Bootstrap of Little Bags

To motivate the bootstrap of little bags, we first note that building confidence intervals via half-sampling—whereby we evaluate an estimator on random halves of the training data to estimate its sampling error—is closely related to the bootstrap (Efron, 1982). In our context, the ideal half-sampling estimator would be

$$\widehat{H}_n^{BLB*}(x) := \binom{n}{\lfloor n/2 \rfloor}^{-1} \sum_{\left\{\mathcal{H} \subset \{1,\ldots,n\}: |\mathcal{H}| = \lfloor \frac{n}{2} \rfloor\right\}} \left(\Psi_{\mathcal{H}}\left(\hat{\theta}(x), \hat{\nu}(x)\right) - \Psi\left(\hat{\theta}(x), \hat{\nu}(x)\right)\right)^2, \quad (51)$$

where $\Psi_{\mathcal{H}}$ denotes a version of Ψ computed only using all the possible trees that only rely on data from the half sample $\mathcal{H}$ (specifically, in terms of Algorithm 1, we only use trees whose full $\mathcal{I}$-subsample is contained in $\mathcal{H}$). If we could evaluate $\widehat{H}_n^{BLB*}(x)$, results from Efron (1982) strongly suggest that it would be a good variance estimator for Ψ; however, from a computational point of view, doing so is effectively impossible as it would require evaluating all possible trees on n samples.

Following Sexton and Laake (2009), however, we can efficiently approximate $\widehat{H}_n^{BLB*}(x)$ at almost no computational cost if we are willing to slightly modify our subsampling scheme. To do so, let $\ell \geq 2$ denote a little bag size and assume, for simplicity, that B is an integer multiple of it.[4] Then, we grow our forest as follows: First draw $g = 1, \ldots, B/\ell$ random half-samples $\mathcal{H}_g \subset \{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$, and then generate the subsamples $\mathcal{I}_b$ used to build the forest in Algorithm 1 such that $\mathcal{I}_b \subseteq \mathcal{H}_{\lceil b/\ell \rceil}$ for each $b = 1, \ldots, B$. In other words, we now generate our forest using little bags of ℓ trees, where all the trees in a given bag only use data from the same half-sample.[5]

The upshot of this construction is that we can now identify $\widehat{H}_n^{BLB*}(x)$ using a simple variance decomposition. Writing Ψ_b for a version of Ψ computed only using the b-th tree, we

[4] This discussion is valid for any fixed value of $\ell \geq 2$; however, different values of ℓ will affect the number of trees B required for the method to stabilize. Sexton and Laake (2009) discuss optimal choices for ℓ, and show that they depend on the ratio of the sampling variance of a single tree to that of the full forest.

[5] Of course, this procedure only works if the subsample size s is less than or equal to $\lfloor n/2 \rfloor$.

can verify that $\widehat{H}_n^{BLB*}(x)$ can be expressed in terms of the "between groups" and "within group" variance terms,

$$\mathbb{E}_{ss}\left[\left(\frac{1}{\ell}\sum_{b=1}^{\ell}\Psi_b - \Psi\right)^2\right] = \widehat{H}_n^{BLB*}(x) + \frac{1}{\ell-1}\mathbb{E}_{ss}\left[\frac{1}{\ell}\sum_{b=1}^{\ell}\left(\Psi_b - \frac{1}{\ell}\sum_{b=1}^{\ell}\Psi_b\right)^2\right], \tag{52}$$

where $\mathbb{E}_{ss}$ denotes expectations over the subsampling mechanism while holding the data fixed. We define our feasible boostrap of little bags variance estimator $\widehat{H}_n^{BLB}(x)$ via the version of (52) that uses empirical moments and note that, given a large enough number of trees B, this converges to the ideal half-sampling estimator.[6]

The result below (proved in the Appendix) verifies that, under the conditions of Theorem 5, the optimal bootstrap of little bags $\widehat{H}_n^{BLB*}(x)$, with plug-in values for $(\hat{\theta}(x), \hat{\nu}(x))$ as in (51), in fact consistently estimates the sampling variance of $\Psi(\theta(x), \nu(x))$; and we have already seen above that the computationally feasible estimator $\widehat{H}_n^{BLB}(x)$ will match $\widehat{H}_n^{BLB*}(x)$ whenever B is large enough. Moreover, given any consistent estimator $\widehat{V}_n(x)$ for $V(x)$, we find that the confidence intervals built using (48) will also be asymptotically valid.

Theorem 9. *Given the conditions of Therorem 5, the bootstrap of little bags is consistent for the sampling variance of Ψ, i.e.,*

$$\left\|\widehat{H}_n^{BLB*}(x) - H_n(x;\, \theta(x),\, \nu(x))\right\| \Big/ \left\|H_n(x;\, \theta(x),\, \nu(x))\right\| \to_p 0. \tag{53}$$

Moreover, given any consistent $\widehat{V}_n(x)$ estimator for $V(x)$ such that $\|\widehat{V}(x) - V(x)\| \to_p 0$, confidence intervals as in (46) built using (48) will asymptotically have nominal coverage.

6 Estimating Conditional Average Partial Effects

Our first application of generalized random forests is to estimating conditional average partial effects under exogeneity; procedurally, the statistical task is equivalent to solving linear regression problems conditionally on features. Suppose that we observe samples $(X_i,\, Y_i,\, W_i) \in \mathcal{X} \times \mathbb{R} \times \mathbb{R}^q$, and posit a random effects model

$$Y_i = W_i \cdot b_i + \varepsilon_i, \quad \beta(x) = \mathbb{E}\left[b_i \,\middle|\, X_i = x\right]. \tag{54}$$

Our goal will be to estimate $\theta(x) = \xi \cdot \beta(x)$ for some contrast $\xi \in \mathbb{R}^p$. Problems of this type arise across several different application areas (e.g., Wooldridge, 2010). If $W_i \in \{0,\, 1\}$ is a treatment assignment, then $\beta(x)$ corresponds to the conditional average treatment effect (e.g., Imbens and Rubin, 2015).

In order for the average effect $\beta(x)$ to be identified, we need to make certain distributional assumptions. Here, we assume that the W_i are exogenous, in the sense that they are independent of the unobservables conditionally on X_i:

$$\{b_i,\, \varepsilon_i\} \perp\!\!\!\perp W_i \,\big|\, X_i. \tag{55}$$

[6] One practical difficulty with the empirical moment estimator based on (52) is that, if B is too small, the resulting variance estimates $\widehat{H}_n^{BLB}(x)$ may be negative. In our software, we avoid this problem by using a Bayesian analysis of variance following, e.g., Gelman et al. (2014), with an improper uniform prior for $\widehat{H}_n^{BLB*}(x)$ over $[0,\, \infty)$. When B is large enough, this distinction washes out.

If W_i is a binary treatment, this condition is equivalent to the standard unconfoundedness assumption used to motivate propensity score methods (Rosenbaum and Rubin, 1983). When exogeneity of this type does not hold, more sophisticated identification strategies are needed; in the following section, we consider one possible solution to a failure of (55), and study the problem of treatment effect estimation via instrumental variables.

6.1 Growing a Forest

Given (55), our parameter of interest $\theta(x) = \xi \cdot \beta(x)$ is identified by the moment condition (4) where $\psi_{\beta(x),\, c(x)}(Y_i,\, W_i) = (Y_i - \beta(x) \cdot W_i - c(x))(1 \;\; W_i^\top)^\top$ and $c(x)$ is an intercept term; this can also be written more explicitly as

$$\theta(x) = \xi^\top \operatorname{Var}\left[W_i \,\middle|\, X_i = x\right]^{-1} \operatorname{Cov}\left[W_i,\, Y_i \,\middle|\, X_i = x\right]. \tag{56}$$

Given forest weights $\alpha_i(x)$ as in (5), the induced estimator $\hat{\theta}(x)$ for $\theta(x)$ is then

$$\hat{\theta}(x) = \xi^\top \left(\sum_{i=1}^n \alpha_i(x) \left(W_i - \overline{W}_\alpha\right)^{\otimes 2}\right)^{-1} \sum_{i=1}^n \alpha_i(x) \left(W_i - \overline{W}_\alpha\right) \left(Y_i - \overline{Y}_\alpha\right), \tag{57}$$

where $\overline{W}_\alpha = \sum \alpha_i(x) W_i$ and $\overline{Y}_\alpha = \sum \alpha_i(x) Y_i$, and $v^{\otimes 2} = vv^\top$ denotes an outer product of a vector with itself.

Given this setup, generalized random forests provide us with a quasi-automatic framework for getting the weights $\alpha_i(x)$ needed in (57); all that needs to be done is to compute the pseudo-outcomes ρ_i from (14) used for recursive partitioning. Here, it is natural to use

$$\begin{aligned} \rho_i &= \xi^\top A_P^{-1} \left(W_i - \overline{W}_P\right) \left(Y_i - \overline{Y}_P - \left(W_i - \overline{W}_P\right) \hat{\beta}_P\right), \\ A_P &= \frac{1}{|\{i : X_i \in P\}|} \sum_{\{i : X_i \in P\}} \left(W - \overline{W}_P\right)^{\otimes 2}, \end{aligned} \tag{58}$$

where now $\overline{W}_P$ and $\overline{Y}_P$ stand for averages taken over the parent P, and $\hat{\beta}_P$ is the least-sqaures regression solution of Y_i on W_i in the parent. Note that the matrix inverse A_P^{-1} only needs to be evaluated once per parent node, and not for every candidate split.

It is straight-forward to check that the conditons required in Section 4 hold here: Assumption 1 holds whenever the functions $\mathbb{E}\left[Y_i \,\middle|\, X_i = x\right]$, $\mathbb{E}\left[W_i \,\middle|\, X_i = x\right]$, $\operatorname{Cov}\left[Y_i,\, W_i \,\middle|\, X_i = x\right]$ and $\operatorname{Var}\left[W_i \,\middle|\, X_i = x\right]$ are all Lipschitz continuous in x, Assumption 2 holds provided that $\operatorname{Var}\left[W_i \,\middle|\, X_i = x\right]$ in invertible, while Assumptions 3–6 hold by construction. Thus, Theorem 5 in fact applies in this setting.

6.1.1 Local Centering

Although the above construction allows for asymptotically valid inference for $\theta(x)$, the performance of the forests can in practice be improved by first regressing out the effect of the features X_i on all the outcomes separately. Specifically, writing

$$y(x) = \mathbb{E}\left[Y_i \,\middle|\, X = x\right] \;\text{ and }\; w(x) = \mathbb{E}\left[W_i \,\middle|\, X = x\right] \tag{59}$$

for the conditional marginal expectations of Y_i and W_i respectively with respect to the sampling distribution used to collect the data, define conditionally centered outcomes

$$\widetilde{Y}_i = Y_i - \hat{y}^{(-i)}\left(X_i\right) \;\text{ and }\; \widetilde{W}_i = W_i - \hat{w}^{(-i)}\left(X_i\right), \tag{60}$$

where $\hat{y}^{(-1)}(X_i)$, etc., are leave-one-out estimates of the marginal expectation defined in (59), computed without using the i-th observation. We then run the full forest using centered outcomes $\{\widetilde{Y}_i, \widetilde{W}_i\}_{i=1}^n$ instead of the original outcomes $\{Y_i, W_i\}_{i=1}^n$.

In order to justify this transformation, we note if there is any set $\mathcal{S} \subseteq \mathcal{X}$ over which $\beta(x)$ is constant (and so $\theta(x)$ is also constant), the following expression also identifies $\theta(x)$ for any $x \in \mathcal{S}$:

$$\begin{aligned}\theta(x) = \xi^\top \operatorname{Var}\left[\left(W_i - \mathbb{E}\left[W_i \,\middle|\, X_i\right]\right) \,\middle|\, X_i \in \mathcal{S}\right]^{-1} \\ \operatorname{Cov}\left[\left(W_i - \mathbb{E}\left[W_i \,\middle|\, X_i\right]\right), \left(Y_i - \mathbb{E}\left[Y_i \,\middle|\, X_i\right]\right) \,\middle|\, X_i \in \mathcal{S}\right].\end{aligned} \tag{61}$$

Thus, if we locally center the Y_i and the W_i before running our forest, the estimator (57) has the potential to be more robust to confounding effects even when the weights $\alpha_i(x)$ are not sharply concentrated around x. Similar orthogonalization ideas have proven to be useful in many statistical contexts (e.g., Chernozhukov et al., 2016; Newey, 1994a; Neyman, 1979).[7]

Finally, in terms of theoretical guarantees, we note that if we ran a forest with any deterministic centering scheme, i.e., we used $\widetilde{Y}_i = Y_i - \hat{y}(X_i)$ for any Lipschitz function $\hat{y}(X_i)$ that does not depend on the data, etc., then the theory developed in Section 4 would allow for valid inference about $\theta(x)$ (note, in particular, that we do not need to assume consistency of $\hat{y}(X_i)$). Moreover, using ideas from, e.g., Chernozhukov et al. (2016) or Schick (1986), we could also emulate this result by using a form of k-fold data splitting or cross-fitting. In the context of forests, it is much more practical to carry out the residualization in (60) via leave-one-out prediction than via k-fold cross-fitting, because leave-one-out prediction in forests is effectively computationally free (Breiman, 2001); however, a practitioner wanting to use results that are precisely covered by theory may prefer to use cross-fitting for the centering step (60).

6.2 Application: Causal Forests

As discussed above, when $W_i \in \{0, 1\}$ is a binary treatment assignment, the setup considered here is equivalent to the standard problem of heterogeneous treatment effect estimation under unconfoundedness.[8] The problem of heterogeneous treatment effect estimation via tree-based methods has received considerable attention in the recent literature. In particular, Athey and Imbens (2016) and Su et al. (2009) develop tree-based methods for subgroup analysis, Green and Kern (2012) and Hill (2011) study treatment effect estimation via Bayesian additive regression trees (BART) proposed by Chipman et al. (2010), and Wager and Athey (2017) propose a causal forest procedure that is very nearly a special case of our generalized random forests. Of course, the main interest of our method is in how it can handle situations for which no comparable methods exist, such as instrumental variables regression as discussed below. Here, however, we briefly discuss how some concepts developed as a part of our more general approach directly improve the practical performance of causal forests.

[7] In some applications, we may believe that $\beta(x)$ varies with a smaller set of covariates x, but that unconfoundedness (55) only holds if we control for a larger set of covariates (x, x'). In this case, motivated by (61), we may want to perform residualization (60) with the richer model for $\mathbb{E}\left[Y \,\middle|\, (X, X') = (x, x')\right]$ and $\mathbb{E}\left[W \,\middle|\, (X, X') = (x, x')\right]$, and then train the generalized random forest using the restricted covariates x.

[8] To draw a tight connection between our setting and the classical potential outcomes framework (Neyman, 1923; Rubin, 1974), note that if we write $Y_i(0)$ and $Y_i(1)$ for the untreated and treated potential outcomes for subject i, then we simply have $\varepsilon_i = Y_i(0)$ and $b_i = Y_i(1) - Y_i(0)$ in (54). Moreover, the condition (55) can equivalently be written as $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \,|\, X_i$, which is in fact the standard unconfoundedness assumption (Rosenbaum and Rubin, 1983).

The most similar method to our forest developed in Section 6 is Procedure 1 of Wager and Athey (2017), which is almost equivalent to our method without centering, the only potentially substantive difference being that they split using the exact loss criterion (10) rather than our gradient-based loss criterion (15), and let each tree compute its own treatment effect estimate rather than using the weighting scheme from Section 2.2 (these methods are exactly equivalent for regression forests, but not for causal forests). Wager and Athey (2017) also consider a second approach, Procedure 2, that obtains its neighborhood function by building a classification forest on the treatment assignments W_i.

A weakness of the methods in Wager and Athey (2017), as they note in their discussion, is that these two procedures have different strengths—Procedure 1 is more sensitive to changes in the treatment effect function, while Procedure 2 is more robust to confounding—but the hard coded nature of these methods makes it difficult to reconcile their relative advantages. But here, since we framed our generalized random forests in terms of more abstract estimating equations, it is "obvious" that we can leverage best practices from the literature on estimating equations and seek to orthogonalize our moment conditions by regressing our the main effect of X_i on W_i and Y_i (e.g., Chernozhukov et al., 2016; Newey, 1994a; Neyman, 1979).

To illustrate the value of this orthogonalization, we revisit the simulation setup from Wager and Athey (2017) where $X_i \sim U([0,\,1]^p)$, $W_i \,\big|\, X_i \sim \text{Bernoulli}(e(X_i))$, and $Y_i \,\big|\, X_i,\, Y_i \sim \mathcal{N}\left(m(X_i) + (W_i - 0.5)\tau(X_i),\, 1\right)$. The authors consider two different simulation settings: One where there is no confounding, $m(x) = 0$ and $e(x) = 0.5$, but there is treatment heterogeneity

$$\tau\left(x\right) = \varsigma\left(x_1\right)\varsigma\left(x_2\right), \quad \varsigma\left(u\right) = 1 + \frac{1}{1 + e^{-20(u-1/3)}}, \tag{62}$$

and second where there is no treatment effect, $\tau(x) = 0$, but confounding is a problem,

$$e(x) = \frac{1}{4}\left(1 + \beta_{2,\,4}(x_3)\right), \quad m(x) = 2x_3 - 1, \tag{63}$$

where $\beta_{a,\,b}$ is the β-density with shape parameters a and b. For the first setting, Wager and Athey (2017) used their Procedure 1, whereas for the second they used Procedure 2, while noting that it is unfortunate that the practitioner must correctly choose which procedure to use based on the problem hand.

Results presented in Table 1 are reassuring, suggesting that that generalized random forests with centering do well under both settings and, moreover, can better handle the case with both confounding and treatment heterogeneity than either of the two previously proposed procedures. Procedure 2 of Wager and Athey is good in the pure confounding setting, but does poorly with strong treatment heterogeneity; this is as expected, as the method only seeks to correct for uneven sampling probabilities, but does not make any splits that directly target treatment heterogeneity. What surprised us is that we had expected Procedure 1 of Wager and Athey and our un-centered generalized random forests to be roughly equivalent—and with either pure confounding and pure heterogeneity they are.[9] However, with both confounding and heterogeneity, our generalized random forests do substantially better, even without centering.

[9] Although the mean-squared error numbers are somewhat different, this appears to be mostly due to implementation differences; recall that both methods rely on fairly complex, non-overlapping software packages. In the case of pure confounding or pure heterogeneity, plotting the mean-squared errors of WA-1 and GRF against each other across different simulation realizations reveals a fairly tight monotone relationship; this type of monotone relationship does not appear in our third setting with both heterogeneity and confounding.

conf.	heterog.	p	n	WA-1	WA-2	GRF	C. GRF
no	yes	10	800	1.37	6.48	0.85	0.87
no	yes	10	1600	0.63	6.23	0.58	0.59
no	yes	20	800	2.05	8.02	0.92	0.93
no	yes	20	1600	0.71	7.61	0.52	0.52
yes	no	10	800	0.81	0.16	1.12	0.27
yes	no	10	1600	0.68	0.10	0.80	0.20
yes	no	20	800	0.90	0.13	1.17	0.17
yes	no	20	1600	0.77	0.09	0.95	0.11
yes	yes	10	800	4.51	7.67	1.92	0.91
yes	yes	10	1600	2.45	7.94	1.51	0.62
yes	yes	20	800	5.93	8.68	1.92	0.93
yes	yes	20	1600	3.54	8.61	1.55	0.57

Table 1: Mean squared error of various "causal forest" methods, that seek to estimate heterogeneous treatment effects under unconfoundedness using forests. We compare our generalized random forests with and without local centering (C. GRF and GRF) with Procedures 1 and 2 of Wager and Athey (2017), WA-1 and WA-2. The first of their procedures uses a direct extension of the CART splitting rule for regression targeted to treatment effect estimation, while the second splits only on the treatment assignments. The simulation settings toggle the presence of confounding (conf.) and treatment heterogeneity (heterog.), as well as the number of features p and the sample size n. All forests have $B = 2,000$ trees, and results are aggregated over 60 simulation replications with 1,000 test points each. The mean-squared errors numbers are multiplied by 10 for readbility.

7 Heterogeneous Treatment Effect Estimation via Instrumental Variables

In many economics applications, we want to measure the causal effect of an intervention on some outcome, all while recognizing that the intervention and the outcome may also be tied together through non-causal pathways, thus ruling out the exogeneity assumption (55) made above. A popular approach in this situation is to rely on instrumental variables (IV) regression, where we find an auxiliary source of randomness that can be used to identify causal effects.

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

To study this question, Angrist and Evans (1998) found a source of auxiliary randomness that can be used to distinguish causal versus correlational effects: They found that, in the United States, parents who already have two children of mixed sexes, i.e., one boy and one girl, will have fewer kids in the future than parents whose first two children were of the same sex. Assuming that the sexes of the first two children in a family are effectively random,

this observed preference for having children of both sexes provides an exogenous source of variation in family size that can be used to identify causal effects: If the mixed sex indicator is unrelated to the mother's propensity to work for a fixed number of children, then the effect of the mixed sex indicator on the observed propensity to work can be attributed to its effect on family size. The instrumental variable estimator normalizes this effect by the effect of mixed sex on family size, so that the normalized estimate is a consistent estimate of the treatment effect of family size on work. Other classical uses of instrumental variables regression include measuring the impact of military service on lifetime income by using the Vietnam draft lottery as an instrument (Angrist, 1990), and measuring the extent to which 401(k) savings programs crowd out other savings, using eligibility for 401(k) savings programs as an instrument (Abadie, 2003; Poterba et al., 1996).

There is a rich literature on non-parametric instrumental variables regression. Classical approaches based on kernels and series estimation have been studied by several authors including Darolles et al. (2011), Hall and Horowitz (2005), Newey and Powell (2003), and Su et al. (2013); see Newey (2013) for a review. More recently, Belloni et al. (2012) consider estimating average treatment effects using the lasso to control for high-dimensional covariates, and Hartford et al. (2016) develop deep learning tools to estimate heterogeneous treatment effects using instrumental variables regression. In our experiments, we will examine how the adaptivity of generalized random forests enables us to improve over traditional kernel- or series-based methods.

7.1 A Forest for Instrumental Variables Regression

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

Suppose that we observe $i = 1, ..., n$ independent and identically distributed subjects, each of whom has features $X_i \in \mathcal{X}$, an outcome $Y_i \in \mathbb{R}$, a treatment assignment $W_i \in \{0, 1\}$, and an instrument $Z_i \in \{0, 1\}$. We believe that the outcomes Y_i and treatment assignment W_i are related via a structural model[10]

$$Y_i = \mu\left(X_i\right) + \tau\left(X_i\right) W_i + \varepsilon_i, \tag{64}$$

where $\tau(X_i)$ is understood to be the causal effect of W_i on Y_i, and ε_i is a noise term that may be positively correlated with W_i. Because ε_i is correlated with W_i, standard regression analyses will not in general be consistent for $\tau(X_i)$. This is where we need to use the instrument Z_i. Suppose we know that Z_i is independent of ε_i conditionally on X_i. Then, provided that Z_i has an influence on the treatment assignment W_i, i.e., that the covariance of Z_i and W_i conditionally on $X_i = x$ is non-zero, we can verify that the treatment effect $\tau(x)$ is identified via

$$\tau(x) = \text{Cov}\left[Y_i,\, Z_i \,\middle|\, X_i = x\right] \,/\, \text{Cov}\left[W_i,\, Z_i \,\middle|\, X_i = x\right]. \tag{65}$$

[10] If we are not willing to assume that every individual i with features $X_i = x$ has the same treatment effect $\tau(x)$, then heterogeneous instrumental variables regression allows us to estimate a (conditional) local average treatment effect (Imbens and Angrist, 1994); see, e.g., Abadie (2003). Here, however, we use the additive structure (64) for simplicity of exposition.

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

$$\psi_{\tau(x),\,\mu(x)} = \begin{pmatrix} Z_i \left(Y_i - W_i\,\tau(x) - \mu(x)\right) \\ Y_i - W_i\,\tau(x) - \mu(x) \end{pmatrix}, \tag{66}$$

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

We can again use our gradient-based formalism to derive a forest that is targeted towards estimating causal effects identified via (65). When growing the forest, the gradient-based labeling (14) gives us pseudo-outcomes

$$\rho_i = \left(Z_i - \overline{Z}_P\right)\left(\left(Y_i - \overline{Y}_P\right) - \left(W_i - \overline{W}_P\right)\hat{\tau}_P\right), \tag{67}$$

where $\overline{Y}_P$, $\overline{W}_P$, $\overline{Z}_P$ are moments in the parent node, and $\hat{\tau}_P$ is a solution to the estimating equation with moments (66) in the parent. Then, given these pseudo-outcomes, the tree executes a CART regression split on the ρ_i as usual. Finally, we obtain personalized treatment effect estimates $\hat{\tau}(x)$ by solving the estimation equation (5) with forest weights (6).

To verify that Theorem 5 holds in this setting, we note that Assumption 1 holds whenever the conditional moment functions $\mathbb{E}\left[W_i \,\middle|\, X_i = x\right]$, $\mathbb{E}\left[Y_i \,\middle|\, X_i = x\right]$, $\mathbb{E}\left[Z_i \,\middle|\, X_i = x\right]$, $\mathbb{E}\left[W_i Z_i \,\middle|\, X_i = x\right]$ and $\mathbb{E}\left[Y_i Z_i \,\middle|\, X_i = x\right]$ are all Lipschitz continuous in x, while Assumption 2 holds whenever the instrument is correlated with treatment assignment (i.e., the instrument is valid). Assumptions 3–6 hold thanks to the definition of ψ.

Finally, as a practical measure, we again center our procedure by regressing out the effect of X_i on all the outcomes separately. Writing

$$y(x) = \mathbb{E}\left[Y_i \,\middle|\, X = x\right],\ w(x) = \mathbb{E}\left[W_i \,\middle|\, X = x\right] \text{ and } z(x) = \mathbb{E}\left[Z_i \,\middle|\, X = x\right], \tag{68}$$

we compute conditionally centered outcomes by leave-one-out estimation

$$\widetilde{Y}_i = Y_i - \hat{y}^{(-i)}\left(X_i\right),\ \widetilde{W}_i = W_i - \hat{w}^{(-i)}\left(X_i\right) \text{ and } \widetilde{Z}_i = Z_i - \hat{z}^{(-i)}\left(X_i\right), \tag{69}$$

and then run the full instrumental variables forest using centered outcomes $\{\widetilde{Y}_i, \widetilde{W}_i, \widetilde{Z}_i\}_{i=1}^n$. We recommend working with centered outcomes by default, and we do so in our simulations. Our package `grf` provides the option of making this transformation automatically, where $\hat{y}^{(-i)}\left(X_i\right)$, $\hat{w}^{(-i)}\left(X_i\right)$ and $\hat{z}^{(-i)}\left(X_i\right)$ are first estimated using 3 separate regression forests.

7.2 Evaluating the Instrumental Variables Splitting Rule

We illustrate the behavior of IV forests in Figure 4 using two simple simulation designs. In both examples, X is uniformly spread over a cube, $X_i \sim [-1,\,1]^p$, but the causal effect $\tau(X_i)$ only depends on the first coordinate $(X_i)_1$. In both panels of Figure 4, we show estimates of $\tau(x)$ produced by different methods, where we vary x_1 and set all other coordinates to 0.

In the first panel, we illustrate the importance of using an IV forests when the treatment assignment may be endogenous. We consider a case where the true causal effect of has a single jump, $\tau(X_i) = 2 \times \mathbf{1}\left(\{(X_i)_1 > -1/3\}\right)$. Meanwhile, at $(X_i)_1 = +1/3$, there is a change in the correlation structure between W_i and ε_i that leads to a spurious (i.e., non-causal) jump in the correlation between W_i and Y_i. As expected, our IV forest correctly picks out the first jump while ignoring the second one. Conversely, a plain causal forest as in Section 6.2 that assumes that the treatment assignment W_i is exogenous will mistakenly also pick out the second spurious jump in the correlation structure of W_i and Y_i.

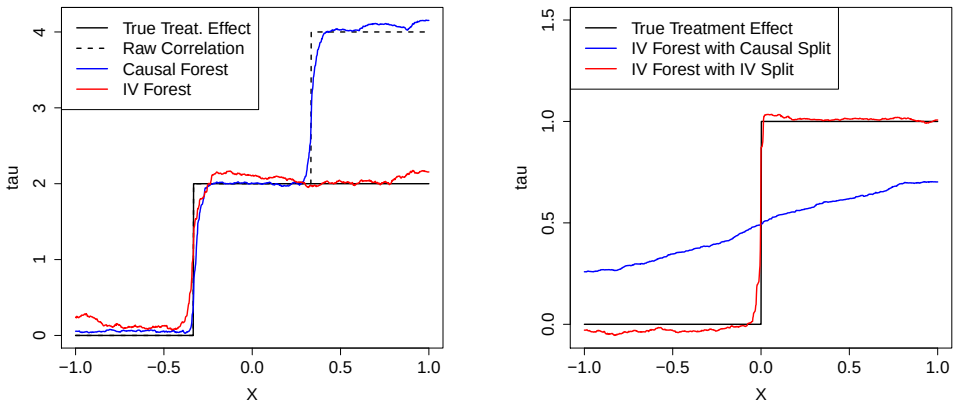


Figure 4: In both panels, we generate data as $X_i \sim [-1, 1]^p$, with $n = 10,000$ and $p = 20$.

Meanwhile, in the second panel, we test our splitting rule. We have a simulation design where there is a jump in the true causal effect, $\tau(X_i) = \mathbf{1}(\{(X_i)_1 > 0\})$. However this causal effect is masked by a change in the correlation of W_i and ε_i, such that the joint distribution of W_i and Y_i does not depend on X_i. Here, the IV forest described above again performs well. However, if we try to use the simpler causal tree splitting from Section 6 that was not designed for IV regression instead of the present splitting rule with pseudo-outcomes (67), then the forest fails to detect any signal.

7.3 Simulation Study

We present some brief simulation results that showcase the value of adaptivity in local instrumental variables regression using generalized random forests. We compare the following four methods: **nearest neighbors** instrumental variables regression, which sets $\alpha_i(x) = 1/k$ in (5) for the k nearest neighbors of x in Euclidean distance, **series** instrumental variables regression, plain **generalized random forests** as described above, and finally **centered generalized random forests**, using residualization as in (69).

Due to computational constraints, we used a fairly limited amount of tuning for each method. For the nearest neighbors method, we tried $k = 10$, 30, 100, 300, and report results for the best resulting choice of k in each setting. For series estimation, we expanded out each feature into a natural spline basis with 3 degrees of freedom, using the `R` function `ns`. We also considered adding interactions of these spline terms to the series basis; however, this led to poor estimates in all of our experiments and so we do not detail these results. Thus, our series method effectively amounts to additive modeling. Finally, we made no effort to tune generalized random forests, and simply ran them with the default tuning parameters in our `grf` software, including a subsample size $s = n/2$. We implemented the nearest neighbors method with the `R` package `FNN` (Beygelzimer et al., 2013), and used the function `ivreg` from the `R` package `AER` (Kleiber and Zeileis, 2008) for series regression.

In all of our simulations, we drew our data from the following generative model, motivated

by an intention to treat design:[11]

$$X_i \sim \mathcal{N}\left(0,\, I_{p\times p}\right),\ \varepsilon_i \sim \mathcal{N}\left(0,\, 1\right),\ Z_i \sim \text{Binom}\left(1/3\right),\ Q_i \sim \text{Binom}\left(1/\left(1+e^{-\omega\varepsilon_i}\right)\right),$$
$$W_i = Z_i \wedge Q_i,\ Y_i = \mu\left(X_i\right) + \frac{2W_i - 1}{2}\tau\left(X_i\right) + \varepsilon_i. \tag{70}$$

In other words, we exogenously draw features X_i, a noise term ε_i and a binary instrument Z_i. Then, the treatment W_i itself depends on both Z_i and Q_i, where Q_i is a random noise term that is correlated with the noise ε_i when $\omega > 0$.

In the context of the above simulation design, we varied the following problem parameters. **Confounding:** We toggled the confounding parameter ω in (70) between $\omega = 0$ (no confounding) and $\omega = 1$ (confounding). **Sparsity of signal:** The signal $\tau(x)$ depended on κ_τ features; we used $\kappa_\tau \in \{2, 4\}$. **Additivity of signal:** When true, we set $\tau(x) = \sum_{j=1}^{\kappa_\tau} \max\{0, x_j\}$; when false, we set $\tau(x) = \max\{0, \sum_{j=1}^{\kappa_\tau} x_j\}$. **Presence of nuisance terms:** When true, we set $\mu(x) = 3\max\{0, x_5\} + 3\max\{0, x_6\}$ or $\mu(x) = 3\max\{0, x_5 + x_6\}$ depending on the additive signal condition; when false we set $\mu(x) = 0$. We also varied the **ambient dimension** p and **sample size** n.

Results from the simulation study are presented in Table 2. We see that the forest-based methods achieve consistently good performance across a wide variety of simulation designs, and do not appear to be too sensitive to non-additive signals or the presence of fairly strong confounding in the treatment assignment. Moreover, we see that the centering behaves as we might have hoped. When there is no nuisance from $\mu(\cdot)$, the centered and uncentered forests perform comparably, while when we add in the nuisance term, the centering substantially improves the performance of generalized random forests.

It is also interesting to examine the few situations where the series method substantially improves over generalized random forests. This only happens in situations where the true signal is additive (as expected), and, moreover, the ambient dimension is small ($p = 10$) while the signal dimension is relatively high ($\kappa_\tau = 4$). In other words, these are the simulation designs where the potential upside from adaptively learning a sparse neighborhood function are the smallest. These results corroborate the intuition that forests provide a form of variable selection for nearest-neighbor estimation.[12]

7.4 Evaluating Confidence Intervals

We also examine the quality of the delta method confidence intervals discussed in Section 5, built using the bootstrap of little bags (Sexton and Laake, 2009). In Table 3, we report coverage results in a subset of the simulation settings from the previous section. We always have confounding ($\omega = 1$) and nuisance terms ($\mu(x) = \max\{0, x_5\} + \max\{0, x_6\}$ or $\mu(x) = \max\{0, x_5 + x_6\}$); we also only consider centered forests. As discussed in Wager et al. (2014), forests typically require more trees to provide accurate confidence intervals; thus, we use $B = 4,000$ trees per forest, rather than the default $B = 2,000$ used in Table 2.

[11] Intuitively, we could think of Z_i as a random intention to treat and of Q_i as a compliance variable; then, if $\omega > 0$, subjects with better outcomes ε_i are more likely to comply, and we need to use the instrument Z_i to deal with this non-compliance effect.

[12] In this simulation design, sparse series regression using the lasso, as in Belloni et al. (2012), might be expected to perform well. Here, however, we only examine the ability of generalized random forests to improve over non-adaptive baselines; a thorough comparison of when lasso- versus forest-based methods perform better is a question that falls beyond the scope of this paper, and hinges on the experience of practitioners in different application areas. In the traditional regression context, both lasso- and forest-based methods have been found to work best in different application areas, and can be considered complementary methods in an applied statistician's toolbox.

					No nuisance from $\mu(\cdot)$				Presence of main effect $\mu(\cdot)$			
add.	conf.	κ_τ	p	n	kNN	series	GRF	C. GRF	kNN	series	GRF	C. GRF
yes	no	2	10	1000	0.50	0.87	**0.33**	0.33	0.77	1.08	0.74	**0.40**
yes	no	2	10	2000	0.42	0.36	**0.23**	0.23	0.64	0.43	0.56	**0.27**
yes	no	2	20	1000	0.56	2.18	0.41	**0.40**	0.82	2.67	0.76	**0.48**
yes	no	2	20	2000	0.51	0.75	0.31	**0.31**	0.78	0.89	0.64	**0.34**
yes	no	4	10	1000	0.87	0.86	0.65	**0.64**	1.23	1.01	1.11	**0.71**
yes	no	4	10	2000	0.79	**0.37**	0.49	0.48	1.03	**0.43**	0.86	0.51
yes	no	4	20	1000	1.09	2.06	0.85	**0.83**	1.35	2.52	1.33	**0.94**
yes	no	4	20	2000	0.96	0.80	0.64	**0.62**	1.23	0.94	1.07	**0.70**
yes	yes	2	10	1000	0.51	0.89	**0.35**	0.36	0.72	1.01	0.69	**0.38**
yes	yes	2	10	2000	0.43	0.37	**0.23**	0.24	0.66	0.42	0.57	**0.26**
yes	yes	2	20	1000	0.57	2.25	0.40	**0.39**	0.86	2.47	0.79	**0.47**
yes	yes	2	20	2000	0.51	0.79	0.28	**0.28**	0.79	0.94	0.65	**0.34**
yes	yes	4	10	1000	0.87	0.88	0.63	**0.62**	1.21	0.99	1.12	**0.69**
yes	yes	4	10	2000	0.78	**0.37**	0.47	0.46	1.02	**0.44**	0.87	0.51
yes	yes	4	20	1000	1.05	2.41	0.80	**0.78**	1.33	2.52	1.28	**0.91**
yes	yes	4	20	2000	0.97	0.78	0.64	**0.62**	1.22	0.93	1.07	**0.67**
no	no	2	10	1000	0.49	0.94	**0.38**	0.39	0.76	1.86	0.85	**0.47**
no	no	2	10	2000	0.41	0.44	0.29	**0.29**	0.61	0.77	0.63	**0.32**
no	no	2	20	1000	0.57	2.34	0.47	**0.47**	0.88	4.47	0.89	**0.57**
no	no	2	20	2000	0.50	0.89	0.35	**0.35**	0.80	1.59	0.71	**0.43**
no	no	4	10	1000	0.83	1.17	0.77	**0.74**	1.18	2.09	1.31	**0.87**
no	no	4	10	2000	0.74	0.66	0.64	**0.61**	1.00	1.02	1.05	**0.66**
no	no	4	20	1000	1.04	2.43	0.98	**0.95**	1.32	4.57	1.35	**1.04**
no	no	4	20	2000	0.93	1.10	0.80	**0.77**	1.18	1.85	1.18	**0.87**
no	yes	2	10	1000	0.49	0.96	**0.37**	0.37	0.73	1.86	0.88	**0.48**
no	yes	2	10	2000	0.41	0.44	**0.28**	0.28	0.62	0.85	0.65	**0.34**
no	yes	2	20	1000	0.55	2.42	0.44	**0.43**	0.85	4.16	0.89	**0.57**
no	yes	2	20	2000	0.49	0.88	0.34	**0.33**	0.75	1.59	0.70	**0.41**
no	yes	4	10	1000	0.83	1.15	0.77	**0.74**	1.19	2.00	1.23	**0.84**
no	yes	4	10	2000	0.73	0.64	0.62	**0.60**	1.01	1.04	1.05	**0.66**
no	yes	4	20	1000	1.04	2.70	0.96	**0.94**	1.36	4.67	1.37	**1.05**
no	yes	4	20	2000	0.94	1.08	0.81	**0.78**	1.22	1.86	1.17	**0.84**

Table 2: Results from simulation study described in Section 7.3, in terms of mean-squared error for the treatment effect on a test set, i.e., $\mathbb{E}[(\hat{\tau}(X) - \tau(X))^2]$, where X is a test example. The methods under comparison are centered generalized random forests (C. GRF), plain generalized random forests (GRF), series instrumental variables regression, and the nearest neighbors method (kNN). The simulation specification varies by whether or not the function $\mu(\cdot)$ in (70) is 0, whether all signals are additive (add.), whether the treatment assignment W is confounded (conf.), the signal dimension (κ_τ), the ambient dimension (p), and the sample size (n). All errors are aggregated over 100 runs of the simulation with 1,000 test points each, and all forests have $B = 2,000$ trees.

		additive	yes	no
κ_τ	p	n	coverage	
2	6	2000	0.91	0.89
2	6	4000	0.92	0.92
2	6	8000	0.93	0.93
2	12	2000	0.85	0.85
2	12	4000	0.91	0.89
2	12	8000	0.94	0.93
2	18	2000	0.85	0.82
2	18	4000	0.91	0.87
2	18	8000	0.94	0.92
4	6	2000	0.85	0.83
4	6	4000	0.87	0.82
4	6	8000	0.90	0.86
4	12	2000	0.75	0.70
4	12	4000	0.80	0.75
4	12	8000	0.86	0.78
4	18	2000	0.66	0.69
4	18	4000	0.76	0.73
4	18	8000	0.81	0.76

Table 3: Emprical coverage of 95% confidence intervals for instrumental variables forests, averaged over 20 replications with 1,000 test points each.

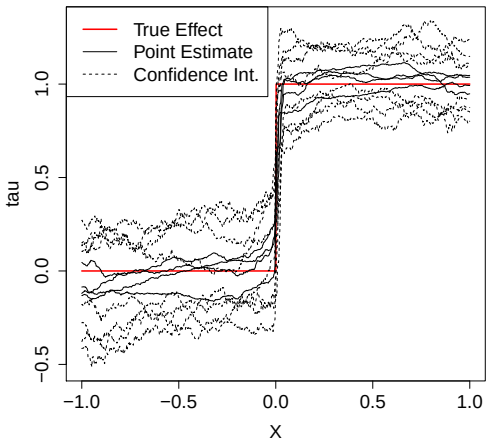


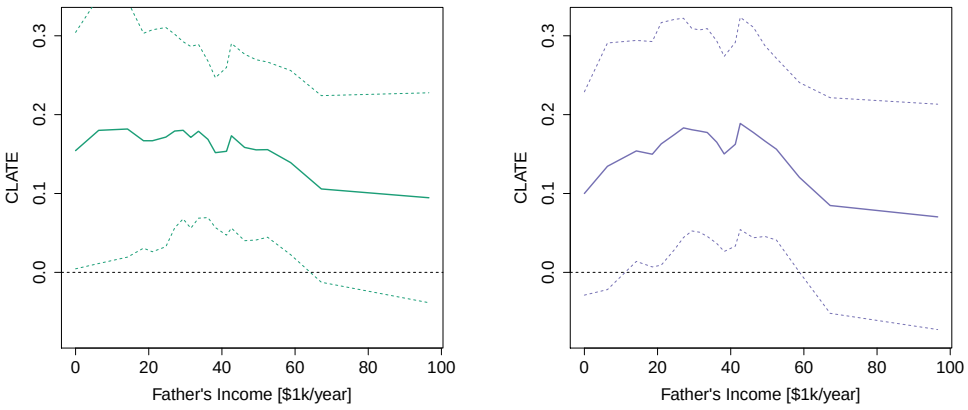
Figure 5: Illustration of 95% confidence intervals for instrumental variables forests across 4 simulation replications. We use the same simulation setting as in the right panel of Figure 4, except now with $n = 4,000$, $p = 20$, and $B = 10,000$ trees.

As expected, coverage results are better when n is larger, the ambient dimension p is smaller, the true signal is sparser, and the true signal is additive. Of these effects, the most important one in Table 3 is the sparsity of τ. When $\kappa_\tau = 2$, i.e., the true signal can be expressed as a bivariate function, our confidence intervals achieve nearly nominal coverage; however, when $\kappa_\tau = 4$, performance declines considerably at the sample sizes n under investigation. Figure 5 gives an illustration of our confidence intervals by superimposing the output from 4 different simulation runs from a single data-generating distribution.

7.5 The Effect of Child Rearing on Labor-Force Participation

Finally, we revisit the motivating example from the beginning of this section. As in Angrist and Evans (1998), we use the sibling-sex composition of a mother's first two children to measure the effect of having a third child on the mother's labor force participation. As discussed earlier, US mothers with two children are more likely to have a third child if their first two children are of the same sex than if they are of mixed sexes. Assuming that the sibling-sex composition of a mother's first two children is effectively randomized, the mixed-sibling-sex indicator can be used as an instrument to identify the causal effect of interest. Angrist and Evans (1998) conduct extensive sensitivity analysis to corroborate the plausibility of this identification strategy. In our analysis, we follow Angrist and Evans (1998) in constructing our dataset.

We study a sample of $n = 334,535$ married mothers with at least 2 children (1980 census



Mother 18 years old at first birth | Mother 22 years old at first birth

Figure 6: Generalized random forest estimates (along with pointwise 95% confidence intervals) for the causal effect of having a third child on the probability that a mother works for pay, as identified by the same sex instrument of Angrist and Evans (1998); a positive treatment effect means that the treatment *reduces* the probability that the mother works. The analysis is based on a sample of $n = 334,535$ married mothers. We vary the mother's age at first birth and the father's income; other covariates are set to their median values in the above plots. The forest was grown with a sub-sample fraction $s/n = 0.05$, a minimum leaf size $k = 800$, and consists of $B = 100,000$ trees.

data), based on the following quantities: The outcome Y_i is whether the mother did not work in the year preceding the census, the treatment W_i is whether the mother had 3 or more children at census time, and the instrument Z_i measures whether or not the mother's first two children were of different sexes.

Based on this data, Angrist and Evans (1998) estimated the local average treatment effect of having a third child. In the data we use for our analysis, $\widehat{\text{Cov}}[W, Z] = 1.6 \cdot 10^{-2}$, while $\widehat{\text{Cov}}[Y, Z] = 2.1 \cdot 10^{-3}$, leading to a 95% confidence interval for the local average treatment effect $\tau \in (0.14 \pm 0.054)$, obtained using the `R` function `ivreg` (Kleiber and Zeileis, 2008). Thus, it appears that having a third child reduces women's labor force participation on average in the US.

Here, we seek to extend this analysis by fitting heterogeneity on several covariates, including the mother's age at the birth of her first child, her age at census time, her years of education and her race (black, hispanic, other), as well as the father's income. Formally, our analysis identifies a conditional local average treatment effect $\tau(x)$ (Abadie, 2003; Imbens and Angrist, 1994).

Results from a generalized random forest analysis are presented in Figure 6. These results suggest that the observed treatment effect is driven by mothers whose husbands have a lower income. Such an effect would be intuitively easy to justify: it seems plausible that mothers with wealthier husbands can afford to hire help in raising their children, and so can choose whether or not to work based on other considerations. That being said, we caution that the father's income was measured in the census, so there is potentially an endogeneity problem:

perhaps a mother's choice not to work after having a third child enables the husband to earn more. Ideally, we would have wanted to measure the husband's income at the time of the second child's birth, but we do not have access to this measurement in the present data. Moreover, the confidence intervals in Figure 6 are rather wide, attesting to the importance of formal asymptotic theory when using forest-based methods for instrumental variables regression.

8 Discussion

In this paper, we introduced generalized random forests as a versatile method for adaptive, local estimation in a wide variety of statistical models. Here, we discussed generalized random forests in the contexts of quantile regression and heterogeneous treatment effect estimation; however, our approach also applies directly to a wide variety of other settings, such as demand estimation or panel data analysis. Our software, `grf`, is implemented in a modular way that should enable users to easily implement splitting rules motivated by new statistical questions.

Many of the remaining challenges with generalized random forests are closely related to those with standard nonparametric methods for local likelihood estimation. In particular, as discussed above, our confidence interval construction relies on undersmoothing to get valid asymptotic coverage (without undersmoothing, the confidence intervals account for sampling variability of the forest, but do not capture bias). Developing a principled way to bias-correct our confidence intervals, and thus avoid the need for undersmoothing, would be of considerable interst both conceptually and in practice. Morevoer, again like standard metods, forests can exhibit edge effects whereby the slope of our estimates $\hat{\theta}(x)$ may taper off as we approach the edge of $\mathcal{X}$-space, even when the true function $\theta(x)$ keeps changing. Finding an elegant way to deal with such edge effects could improve the quality of the confidence intervals provided by generalized random forests.

References

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

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

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

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

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

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

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

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

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program evaluation with high-dimensional data. *arXiv preprint arXiv:1311.2645*, 2013.

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

A. Beygelzimer, S. Kakadet, J. Langford, S. Arya, D. Mount, and S. Li. *FNN: Fast Nearest Neighbor Search Algorithms and Applications*, 2013. URL http://CRAN.R-project.org/package=FNN. R package version 1.1.

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

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

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

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

I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistics Surveys*, 5:44–71, 2011.

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

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

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

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. CRC press, 1984.

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

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*, 2016.

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

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

S. Darolles, Y. Fan, J.-P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

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

T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000.

B. Efron. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

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

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

J. Fan and I. Gijbels. *Local Polynomial Modelling and its Applications*, volume 66. CRC Press, 1996.

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

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

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63 (1):3–42, 2006.

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

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

P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929, 2005.

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

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

J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Counterfactual prediction with deep instrumental variables networks. *arXiv preprint arXiv:1612.09596*, 2016.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. New York: Springer, 2009.

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

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

T. K. Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

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

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

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

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

G. W. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635, 2008.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

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

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

C. Kleiber and A. Zeileis. *Applied econometrics with R*. Springer Science & Business Media, 2008.

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

A. Lewbel. A local generalized method of moments estimator. *Economics Letters*, 94(1): 124–128, 2007.

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

C. Loader. *Local Regression and Likelihood*. Springer, 1999.

C. L. Mallows. Some comments on Cp. *Technometrics*, 15(4):661–675, 1973.

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

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

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

W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, 62(6):1349–1382, 1994a.

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

W. K. Newey. Nonparametric instrumental variables estimation. *The American Economic Review*, 103(3):550–556, 2013.

W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

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

J. Neyman. C(α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 1–21, 1979.

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

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

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

J. Robins, L. Li, E. Tchetgen, and A. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

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

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

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

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

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

A. Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151, 1986.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

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

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

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

L. Su, I. Murtazashvili, and A. Ullah. Local linear gmm estimation of functional coefficient iv models with an application to estimating the rate of return to schooling. *Journal of Business & Economic Statistics*, 31(2):184–207, 2013.

X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li. Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research*, 10:141–158, 2009.

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

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

A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

H. R. Varian. Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2):3–27, 2014.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

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

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

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.

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

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

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

R. Zhu, D. Zeng, and M. R. Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784, 2015.

9 Appendix

The proofs presented here depend on arguments and notation established in Section 4.3. The proof of Proposition 1 builds on that of Proposition 2, so we present the latter first.

Proof of Proposition 2

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

$$\tilde{\theta}^*_{C_j}(x) = \theta(x) + \frac{1}{|C_j|} \sum_{i \in C_j} \rho_i^*(x), \tag{71}$$

where x is some anchor point $x \in P$.

Because the leaf C_j is considered fixed, we can use second-moment bounds on ψ to verify that $\text{Var}[\tilde{\theta}^*_{C_j}(x)] = \mathcal{O}\left(1/n_{C_j}\right)$; meanwhile, by Lipschitz-continuity of the M-function (19), we see that $\mathbb{E}[\tilde{\theta}^*_{C_j}(x) - \theta(x)] = \mathcal{O}(r)$, where r is the radius of the leaf. Finally, given assumptions made so far about the estimating equation, it is straight-forward to show that $\hat{\theta}_{C_j}$ is consistent for $\theta(x)$ in a limit where $r \to 0$ and $n_{C_j} \to \infty$. Thus, a direct analogue to our result, Lemma 3, implies that

$$\tilde{\theta}^*_{C_j}(x) - \hat{\theta}_{C_j} = o_P\left(r, 1/\sqrt{n_{C_j}}\right). \tag{72}$$

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

$$\begin{aligned} \tilde{\theta}_{C_j} - \tilde{\theta}^*_{C_j}(x) = \hat{\theta}_P - \theta(x) - \frac{1}{n_{C_j}} \xi^\top V(x)^{-1} \sum_{i \in C_j} \left(\psi_{\hat{\theta}_P,\, \hat{\nu}_P}(O_i) - \psi_{\theta(x),\, \nu(x)}(O_i)\right) \\ - \frac{1}{n_{C_j}} \xi^\top \left(A_P^{-1} - V(x)^{-1}\right) \sum_{i \in C_j} \psi_{\hat{\theta}_P,\, \hat{\nu}_P}(O_i); \end{aligned} \tag{73}$$

our goal is then to bound the terms on the first and second lines at the desired rate. The first line term is bounded by $o_P(r)$ by smoothness of the M-function as we change θ and ν, as well as an analogue to Lemma 8; while the second line term can be bounded by recalling that $\|A_P^{-1} - V(x)^{-1}\| = o_P(1)$, and verifying that $\sum_{i \in C_j} \psi_{\hat{\theta}_P,\, \hat{\nu}_P}(O_i) = \mathcal{O}_P\left(1/\sqrt{n_{C_j}}, r\right)$. Everything we have showed so far implies that

$$\tilde{\theta}_{C_j} - \hat{\theta}_{C_j} = o_P\left(r, 1/\sqrt{n_{C_j}}\right), \text{ for } j = 1, 2. \tag{74}$$

Finally, it is straight-forward to check that

$$\tilde{\theta}_{C_2} - \tilde{\theta}_{C_1} = O_P\left(r, 1/\sqrt{n_{C_1}}, 1/\sqrt{n_{C_2}}\right), \tag{75}$$

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

Proof of Proposition 1

First, we show that we can replace $\hat{\theta}_{C_j}(\mathcal{J})$ with the influence-based approximation $\tilde{\theta}^*_{C_j}(x;\, \mathcal{J})$ (where we make explicit the dependence of $\tilde{\theta}^*_{C_j}$ on the sample $\mathcal{J}$ for clarity) when computing the error function $\text{err}(C_j)$. To simplify notation without changing the essence of the argument, we restrict attention to samples $\mathcal{J}$ where the number of observations in C_1 and C_2 are held fixed at n_{C_1} and n_{C_2}, respectively (and recall from the main text that P, C_1, and C_2, subsets of $\mathcal{X}$, are also held fixed). To start, let $x \in P$ be an anchor point in the parent leaf, and observe that

$$\begin{aligned}
\text{err}(C_j) &= \mathbb{E}_{X \in C_j}\left[\left(\hat{\theta}_{C_j}(\mathcal{J}) - \theta(X)\right)^2\right] \\
&= \mathbb{E}_{X \in C_j}\left[\left(\tilde{\theta}^*_{C_j}(x;\, \mathcal{J}) - \theta(X)\right)^2\right] + \underbrace{\mathbb{E}\left[\left(\hat{\theta}_{C_j}(\mathcal{J}) - \tilde{\theta}^*_{C_j}(x;\, \mathcal{J})\right)^2\right]}_{o\left(r^2,\, 1/n_{C_j}\right)} \\
&\quad + 2\underbrace{\mathbb{E}\left[\hat{\theta}_{C_j}(\mathcal{J}) - \tilde{\theta}^*_{C_j}(x;\, \mathcal{J})\right]}_{o\left(r,\, 1/\sqrt{n_{C_j}}\right)}\underbrace{\left(\mathbb{E}\left[\tilde{\theta}^*_{C_j}(x;\, \mathcal{J})\right] - \mathbb{E}_{X \in C_j}\left[\theta(X)\right]\right)}_{\mathcal{O}(r^2)},
\end{aligned}$$

where the first two bounds given in underbraces follow from the proof of Proposition 2, while the last one is a direct consequence of Assumption 2, by noting that

$$\mathbb{E}_{X \in C_j}\left[\theta(X)\right] - \mathbb{E}\left[\tilde{\theta}^*_{C_j}(x;\, \mathcal{J})\right] = \mathbb{E}_{X \in C_j}\left[\theta(X) - \theta(x) - \xi^\top \left(\nabla M_{\theta(x),\, \nu(x)}(x)\right)^{-1} M_{\theta(x),\, \nu(x)}(X)\right],$$

and so this term is just the average error from Taylor expanding $M_{\theta(x),\, \nu(x)}(\cdot)$ over C_j.

Now, using the above expansion, we find that

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

Following arguments of Athey and Imbens (2016), we see that

$$\begin{aligned}
\mathbb{E}_{X \in C_j}\left[\left(\tilde{\theta}^*_{C_j}(x;\, \mathcal{J}) - \theta(X)\right)^2\right] &= \text{Var}_{X \in C_j}\left[\theta(X)\right] + \text{Var}\left[\tilde{\theta}^*_{C_j}(x;\, \mathcal{J})\right] \\
&\quad + \left(\mathbb{E}\left[\tilde{\theta}^*_{C_j}(x;\, \mathcal{J})\right] - \mathbb{E}_{X \in C_j}\left[\theta(X)\right]\right)^2,
\end{aligned}$$

and the last term is bounded by $\mathcal{O}\left(r^4\right)$ as argued above. Thus,

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

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

$$\mathbb{E}\left[\Delta\left(C_1,\, C_2\right)\right] = \frac{n_{C_1} n_{C_2}}{n_P^2}\, \mathbb{E}\left[\left(\tilde{\theta}^*_{C_2}(x;\, \mathcal{J}) - \tilde{\theta}^*_{C_1}(x;\, \mathcal{J})\right)^2\right] + o\left(r^2,\, \frac{1}{n_{C_1}},\, \frac{1}{n_{C_2}}\right). \tag{76}$$

Thus, writing $K(P) := \text{Var}_{X \in P}\left[\theta(X)\right]$ as the split-independent error term, all that remains is the sampling variance of $\Delta\left(C_1,\, C_2\right)$ due to noise in the training sample $\mathcal{J}^{tr}$ (which becomes negligible as n gets large), and a term

$$\mathcal{E} := \frac{1}{n_P} \sum_{j=1}^{2} n_{C_j} \left(2 - \frac{n_{C_j}}{n_P}\right) \text{Var}\left[\tilde{\theta}^*_{C_j}(x;\, \mathcal{J})\right] \tag{77}$$

that captures the effect of overfitting to random noise when estimating $\tilde{\theta}^*_{C_j}(x)$. This last term scales as $\mathcal{E} = \mathcal{O}_P(1/n_{C_1},\, 1/n_{C_2})$, and so can be ignored since we assume that $n_P \gg r^{-2}$. Note that if we attempt to correct for a plug-in version of $\mathcal{E}$, we recover exactly the variance correction of Athey and Imbens (2016), up to an additive term that is the same for all splits and so doesn't affect split selection.

9.1 Proof of Lemma 4

First, thanks to Lemma 6, we know that

$$\left\|\Psi\left(\theta(x),\, \nu(x)\right)\right\|_2 \to_p 0. \tag{78}$$

Thus, thanks to (23), we know there must exist a sequence $\varepsilon_n > 0$ with $\lim_{n \to \infty} \varepsilon_n = 0$ such that

$$\left\|\Psi\left(\theta(x),\, \nu(x)\right)\right\|_2,\, \left\|\Psi\left(\hat{\theta}(x),\, \hat{\nu}(x)\right)\right\|_2 \leq \varepsilon_n$$

with probability tending to 1; and so Lemma 10 below implies the desired result.

Lemma 10. *Suppose that Assumptions 1–7 hold. Then, all approximate solutions to (5) are close to each other, in the following sense: for any sequence $\varepsilon_n > 0$ with $\lim_{n \to \infty} \varepsilon_n = 0$,*

$$\sup\left\{\left\|\begin{pmatrix}\theta - \theta' \\ \nu - \nu'\end{pmatrix}\right\|_2 : \left\|\Psi\left(\theta,\, \nu\right)\right\|_2,\, \left\|\Psi\left(\theta',\, \nu'\right)\right\|_2 \leq \varepsilon_n\right\} \to_p 0. \tag{79}$$

Proof. Starting with some notation, let

$$\Psi(\theta,\,\nu) \in \partial F(\theta,\,\nu), \quad \overline{\Psi}(\theta,\,\nu) = \nabla \overline{F}(\theta,\,\nu),$$

where F and $\overline{F}$ are the respectively convex and σ^2-strongly convex functions implicitly defined in the hypothesis statement. Recall that $(\hat{\theta},\,\hat{\nu})$ is assumed to satisfy (23), and let $\eta_n > 0$ be any sequence with $\lim_{n\to\infty} \eta_n = 0$, $\eta_n > \max\{4\varepsilon_n/\sigma^2,\,\sqrt{s/n}\}$ for all $n = 1,\,2,\,...$, and $\eta_n^{-1}\|\Psi(\hat{\theta},\,\hat{\nu})\|_2 \to_p 0$.

Now, thanks to Assumptions 1–4, we can apply Lemma 8. Because $\eta_n \geq \sqrt{s/n}$, we can pair (40) with the fundamental theorem of calculus for line integrals to check that

$$\begin{aligned} &F(\theta,\,\nu) - F(\hat{\theta},\,\hat{\nu}) - \Psi(\hat{\theta},\,\hat{\nu}) \cdot \begin{pmatrix} \theta - \hat{\theta} \\ \nu - \hat{\nu} \end{pmatrix} \\ &\quad = \overline{F}(\theta,\,\nu) - \overline{F}(\hat{\theta},\,\hat{\nu}) - \overline{\Psi}(\hat{\theta},\,\hat{\nu}) \cdot \begin{pmatrix} \theta - \hat{\theta} \\ \nu - \hat{\nu} \end{pmatrix} + o_P\left(\eta_n^2\right), \end{aligned}$$

for points $(\theta,\,\nu)$ within L_2-distance η_n of $(\hat{\theta},\,\hat{\nu})$. By strong convexity of $\overline{F}$, this implies that

$$F(\theta,\,\nu) \geq F(\hat{\theta},\,\hat{\nu}) + \Psi(\hat{\theta},\,\hat{\nu}) \cdot \begin{pmatrix} \theta - \hat{\theta} \\ \nu - \hat{\nu} \end{pmatrix} + \frac{\sigma^2}{2} \left\| \begin{pmatrix} \theta - \hat{\theta} \\ \nu - \hat{\nu} \end{pmatrix} \right\|_2^2 + o_P\left(\eta_n^2\right),$$

again for $(\theta,\,\nu)$ within η_n of $(\hat{\theta},\,\hat{\nu})$. Thus, with probability tending to 1,

$$\inf\left\{ F(\theta,\,\nu) - F(\hat{\theta},\,\hat{\nu}) : \left\| \begin{pmatrix} \theta - \hat{\theta} \\ \nu - \hat{\nu} \end{pmatrix} \right\|_2 = \eta_n \right\} \geq \frac{\sigma^2}{4}\,\eta_n^2;$$

note that, here, we also used the fact that $\eta_n^{-1}\|\Psi(\hat{\theta},\,\hat{\nu})\|_2 \to_p 0$. Finally, by convexity of F, this last fact implies that, with probability tending to 1,

$$\|\Psi\,(\theta,\,\nu)\|_2 \geq \frac{\sigma^2}{4}\,\eta_n \quad \text{for all} \quad \left\| \begin{pmatrix} \theta - \hat{\theta} \\ \nu - \hat{\nu} \end{pmatrix} \right\|_2 \geq \eta_n.$$

Recall that, by construction, $\varepsilon_n < \sigma^2\eta_n/4$, and so (79) must hold. □

9.2 Proof of Theorem 9

Following our discussion in Section 5.1, we here only consider the $B \to \infty$ limiting bootstrap of little bags estimator. We start by considering its expectation,

$$\mathbb{E}\left[\widehat{H}_n^{BLB*}(x)\right] = \mathbb{E}\left[\left(\Psi_{\mathcal{H}}\left(\hat{\theta}(x),\,\hat{\nu}(x)\right) - \Psi\left(\hat{\theta}(x),\,\hat{\nu}(x)\right)\right)^{\otimes 2}\right],$$

for $\mathcal{H} = \{1,\,...,\,\lfloor n/2 \rfloor\}$. By the proof of Theorem 5, we know that $\|(\hat{\theta}(x),\,\hat{\nu}(x)) - (\theta(x),\,\nu(x))\|_2^2 = \mathcal{O}_P(s/n)$, and so we can use Lemma 8 with $\eta = (s/n)^{1/3}$ to verify that

$$\begin{aligned} &\Psi_{\mathcal{H}}\left(\hat{\theta}(x),\,\hat{\nu}(x)\right) - \Psi\left(\hat{\theta}(x),\,\hat{\nu}(x)\right) = Q_{\mathcal{H}} + R_{\mathcal{H}} + \mathcal{O}_P\left(\left(\frac{s}{n}\right)^{2/3}\right), \\ &Q_{\mathcal{H}} := \Psi_{\mathcal{H}}\left(\theta(x),\,\nu(x)\right) - \Psi\left(\theta(x),\,\nu(x)\right), \\ &R_{\mathcal{H}} := \overline{\Psi}_{\mathcal{H}}\left(\hat{\theta}(x),\,\hat{\nu}(x)\right) - \overline{\Psi}_{\mathcal{H}}\left(\theta(x),\,\nu(x)\right) - \left(\overline{\Psi}\left(\hat{\theta}(x),\,\hat{\nu}(x)\right) - \overline{\Psi}\left(\theta(x),\,\nu(x)\right)\right), \end{aligned}$$

where $\overline{\Psi}_{\mathcal{H}}$ is defined analogously to $\overline{\Psi}$ in (30).

The first term above, $Q_{\mathcal{H}}$, is the type of term used by an oracle half-sampling estimator that gets to use the true parameter values $(\theta(x), \nu(x))$ rather than their plug-in analogues. Given our assumptions and because $(\theta(x), \nu(x))$ is non-random, we can use results from Wager and Athey (2017) to directly verify that (see their Lemma 7 and Theorem 8)

$$\begin{aligned}
&\Psi\left(\theta(x), \nu(x)\right) - \mathbb{E}\left[\Psi\left(\theta(x), \nu(x)\right)\right] \\
&\quad = \left(1 + o_P(1)\right) \sum_{i=1}^{n} \left(\mathbb{E}\left[\Psi\left(\theta(x), \nu(x)\right) \mid \left(X_i, O_i\right)\right] - \mathbb{E}\left[\Psi\left(\theta(x), \nu(x)\right)\right]\right), \\
&\Psi_{\mathcal{H}}\left(\theta(x), \nu(x)\right) - \mathbb{E}\left[\Psi\left(\theta(x), \nu(x)\right)\right] \\
&\quad = \left(1 + o_P(1)\right) \frac{n}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \left(\mathbb{E}\left[\Psi\left(\theta(x), \nu(x)\right) \mid \left(X_i, O_i\right)\right] - \mathbb{E}\left[\Psi\left(\theta(x), \nu(x)\right)\right]\right).
\end{aligned} \tag{80}$$

The reason this result holds is that, as discussed in Wager and Athey (2017), forests have regularity properties by which the scaled first-order effects $n(\mathbb{E}\left[\Psi\left(\theta(x), \nu(x)\right) \mid \left(X_i, O_i\right)\right] - \mathbb{E}\left[\Psi\left(\theta(x), \nu(x)\right)\right])$ depend only on the type of tree being grown; and here of course Ψ and $\Psi_{\mathcal{H}}$ are built using exactly the same type of trees ($\Psi_{\mathcal{H}}$ just averages over fewer of them). Given tail bounds to control moments, it follows immediately that

$$\begin{aligned}
\mathbb{E}\left[Q_{\mathcal{H}}^{\otimes 2}\right] &= n\left(1 + o(1)\right) \mathbb{E}\left[\left(\mathbb{E}\left[\Psi\left(\theta(x), \nu(x)\right) \mid \left(X_1, O_1\right)\right] - \mathbb{E}\left[\Psi\left(\theta(x), \nu(x)\right)\right]\right)^{\otimes 2}\right] \\
&= \left(1 + o(1)\right) H_n(x; \theta(x), \nu(x)),
\end{aligned}$$

where the latter is again immediate by the proof of Theorem 8 in Wager and Athey (2017). Thus, taking second moments term $Q_{\mathcal{H}}$ gives us the limiting expectation we want.

It remains to show that the residual term $R_{\mathcal{H}}$, used to account for the plug-in effects, is negligible. Recall that $\overline{\Psi}$ is twice differentiable with a uniform second derivative, so we can take a Taylor expansion as in the proof of Lemma 3:

$$R_{\mathcal{H}} = \left(\nabla\Psi_{\mathcal{H}}\left(\theta(x), \nu(x)\right) - \nabla\Psi\left(\theta(x), \nu(x)\right)\right) \left(\begin{pmatrix}\hat{\theta}(x) \\ \hat{\nu}(x)\end{pmatrix} - \begin{pmatrix}\theta(x) \\ \nu(x)\end{pmatrix}\right) + \mathcal{O}_P\left(\frac{s}{n}\right),$$

where the s/n error term is a bound on the squared error of $(\hat{\theta}(x), \hat{\nu}(x))$. Now, by the same argument as in (43), we see that $\|\nabla\Psi_{\mathcal{H}}\left(\theta(x), \nu(x)\right) - \nabla\Psi\left(\theta(x), \nu(x)\right)\| \to_P 0$, whereas the squared distance between $(\hat{\theta}(x), \hat{\nu}(x))$ and $(\theta(x), \nu(x))$ is of the same order as $H_n(x; \theta(x), \nu(x))$; and so in fact

$$\left\|\mathbb{E}\left[R_{\mathcal{H}}^{\otimes 2}\right]\right\| = o_P\left(\|H_n(x; \theta(x), \nu(x))\|\right),$$

implying that

$$\left\|\mathbb{E}\left[\widehat{H}_n^{BLB*}(x)\right] - H_n(x; \theta(x), \nu(x))\right\| = o_P\left(\|H_n(x; \theta(x), \nu(x))\|\right).$$

To establish (53), it remains to verify concentration of $\widehat{H}_n^{BLB*}(x)$; which, given that the contribution of $R_{\mathcal{H}}$ is negligible, also follows immediately from (80). Finally, given (53) and Theorem 5, the validity of the delta method confidence intervals is immediate by Slutsky's theorem whenever $\|\widehat{V}(x) - V(x)\| \to_p 0$; in particular, recall that $V(x)$ is invertible by Assumption 2.

Proof of Lemma 7

We first note that, because we grew our forest honestly (Assumption 7) and so α_i is independent of O_i conditionally on X_i, we can use the chain rule to verify that

$$\mathbb{E}\left[\Psi\left(\theta,\,\nu\right)-\overline{\Psi}\left(\theta,\,\nu\right)\right]=\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}\left[\alpha_i(x)\,\middle|\,X_i\right]\left(\mathbb{E}\left[\psi_{\theta,\,\nu}(O_i)\,\middle|\,X_i\right]-M_{\theta,\,\nu}(X_i)\right)\right]=0,$$

and so δ_α must be mean-zero.

Next, to establish bounds on the second moments, we start by considering individual trees. To do so, define

$$\mathcal{E}_{\theta,\,\nu}(O_i,\,X_i)=\psi_{\theta,\,\nu}(O_i)-M_{\theta,\,\nu}(X_i).$$

Because $\mathbb{E}\left[\mathcal{E}_{\theta,\,\nu}(O_i,\,X_i)\,\middle|\,M_{\theta,\,\nu}(X_i)\right]=0$ and $M_{\theta,\,\nu}(X_i)$ is locally $(\theta,\,\nu)$-Lipschitz continuous by Assumption 2, we can verify that the worst-case variogram of the $\mathcal{E}_{\theta,\,\nu}(O_i,\,X_i)$ must also satisfy (22). Now, as in our Algorithm 1 let $\mathcal{J}_1$, $\mathcal{J}_2$ be any non-overlapping subset of points of size $\lfloor s/2\rfloor$ and $\lceil s/2\rceil$ respectively. Let $\alpha_i\geq 0$ be weights summing to 1 such that $\{\alpha_i : i\in\mathcal{J}\}$ depends only on $\mathcal{J}_2$ and on $\{X_i : i\in\mathcal{J}_1\}$, and write

$$T_{\theta,\,\nu}(\mathcal{J}_1,\,\mathcal{J}_2)=\sum_{\{i\in\mathcal{J}_1\}}\alpha_i\mathcal{E}_{\theta,\,\nu}(O_i,\,X_i).$$

By the previous argument, we already know that $\mathbb{E}\left[T_{\theta,\,\nu}(\mathcal{J}_1,\,\mathcal{J}_2)\right]=0$; meanwhile, thanks to the variogram bound, for any pair of points $(\theta,\,\nu)$ and $(\theta',\,\nu')$,

$$\begin{aligned}&\mathbb{E}\left[\left\|T_{\theta,\,\nu}(\mathcal{J}_1,\,\mathcal{J}_2)-T_{\theta',\,\nu'}(\mathcal{J}_1,\,\mathcal{J}_2)\right\|_2^2\right]\\&\quad\leq\mathbb{E}\left[\sum_{\{i\in\mathcal{J}_1\}}\alpha_i^2\mathbb{E}\left[\left\|\mathcal{E}_{\theta,\,\nu}(O_i,\,X_i)-\mathcal{E}_{\theta',\,\nu'}(O_i,\,X_i)\right\|_2^2\,\middle|\,X_i\right]\right]\leq L\left\|\begin{pmatrix}\theta\\\nu\end{pmatrix}-\begin{pmatrix}\theta'\\\nu'\end{pmatrix}\right\|_2.\end{aligned}\tag{81}$$

As in arguments used by Wager and Athey (2017), we see that our quantity of interest U-statistic over T, and in particular

$$\delta_\alpha\left(\left(\theta,\,\nu\right),\left(\theta',\,\nu'\right)\right)=\binom{n}{\lfloor s/2\rfloor,\,\lceil s/2\rceil}^{-1}\sum_{\{\mathcal{S}_1,\,\mathcal{S}_2\in\{1,\,\ldots,\,n\}\}}T_{\theta,\,\nu}(\mathcal{J}_1,\,\mathcal{J}_2)-T_{\theta',\,\nu'}(\mathcal{J}_1,\,\mathcal{J}_2).$$

Thus, combing our above variogram bound for T with results on U-statistics going back to Hoeffding (1948), we see that (37) holds.

Proof of Lemma 8

We start by establishing a concentration bound for δ_α at a single point. Given Assumption 4, we know that $\|\delta_\alpha\|_\infty$ is bounded by $2G$, where G is as defined in the problem statement. Thus, recalling that δ_α is a U-statistic and using (81) to bound the variance of a single tree, we can use the Bernstein bound for U-statistics established by Hoeffding (1963) to verify that, for any $\eta>0$,

$$\mathbb{P}\left[\left\|\delta_\alpha\left(\left(\theta,\,\nu\right),\left(\theta',\,\nu'\right)\right)\right\|_\infty>\eta\right]\leq 2k\exp\left(-\lfloor n/s\rfloor\eta^2\,/\,\left(2L\left\|\begin{pmatrix}\theta-\theta'\\\nu-\nu'\end{pmatrix}\right\|_2+\frac{4G}{3}\eta\right)\right).\tag{82}$$

In other words, as expected, the forest concentrates at a rate $\sqrt{s/n}$.

Now, the kernel of δ_α, i.e., the function evaluated on subsamples, is a sum of 4 components that can all be bracketed into a number of brackets bounded as in (39), using the radius (38). Thus, the kernel of δ_α can be bracketed with respect to L_2-measure with a bracketing entropy of at most $16\kappa/\eta$. Given these preliminaries, we proceed by replicating the argument from Lemma 3.4.2 of van der Vaart and Wellner (1996) and, in particular, replacing all applications of Bernstein's inequality with Bernstein's inequality for U-statistics as in (82), we find that for any set $\mathcal{S}$ with $\|T_{\theta,\,\nu}(\mathcal{J}_1,\,\mathcal{J}_2) - T_{\theta',\,\nu'}(\mathcal{J}_1,\,\mathcal{J}_2)\|_2^2 \leq r^2$ for all $((\theta,\,\nu),\,(\theta',\,\nu')) \in \mathcal{S}$, we have

$$\mathbb{E}\left[\sup\left\{\delta_\alpha\left((\theta,\,\nu)\,,\,(\theta',\,\nu')\right) : ((\theta,\,\nu),\,(\theta',\,\nu')) \in \mathcal{S}\right\}\right] = \mathcal{O}\left(\frac{J_{[]}(r,\,\delta_\alpha,\,L_2)}{\sqrt{\lfloor n/s \rfloor}} + \frac{J_{[]}^2(r,\,\delta_\alpha,\,L_2)}{r^2\,\lfloor n/s \rfloor} 2G\right),$$

where $J_{[]}$ is the bracketing entropy integral

$$J_{[]}(r,\,\delta_\alpha,\,L_2) := \int_0^r \sqrt{1 + \log\left(N_{[]}(\eta,\,\delta_\alpha,\,L_2)\right)}\; d\eta.$$

From our bounds on the bracketing number we get $J_{[]}(r,\,\delta_\alpha,\,L_2) \leq 4\sqrt{\kappa r} + o(\sqrt{r})$. Thus, thanks to Lemma 7, we can apply the above result with $r = L\eta$ to obtain the desired conclusion.