

## GENERALIZED RANDOM FORESTS

BY SUSAN ATHEY JULIE TIBSHIRANI AND STEFAN WAGER

*Stanford University and Elasticsearch BV*

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 considers 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), are a widely used algorithm for statistical learning. Statisticians usually study random forests as a practical method for non-parametric conditional mean estimation: Given a data-generating distribution for  $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ , forests are used to estimate  $\mu(x) = \mathbb{E}[Y_i | X_i = x]$ . Several theoretical results are available on the asymptotic behavior of such forest-based estimates  $\hat{\mu}(x)$ , including consistency (Arlot and Genuer, 2014; Biau, Devroye and Lugosi, 2008; Biau, 2012; Denil, Matheson and De Freitas, 2014; Lin and Jeon, 2006; Scornet, Biau and Vert, 2015; Wager and Walther, 2015), second-order asymptotics (Mentch and Hooker, 2016), and confidence intervals (Wager and Athey, 2018).

This paper extends Breiman’s random forests into a flexible method for estimating any quantity  $\theta(x)$  identified via local moment conditions. Specifically, given data  $(X_i, O_i) \in \mathcal{X} \times \mathcal{O}$ , we seek forest-based estimates of  $\theta(x)$  defined by a local estimating equation of the form

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

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  $O_i$  conditionally on  $X_i$  as having a density

$f_{\theta(x), \nu(x)}(\cdot)$  then, under standard regularity conditions, the moment condition (1) with  $\psi_{\theta(x), \nu(x)}(O) = \nabla \log(f_{\theta(x), \nu(x)}(O_i))$  identifies the local maximum likelihood parameters  $(\theta(x), \nu(x))$ . More generally, we can use moment conditions of the form (1) to identify conditional means, quantiles, average partial effects, etc., and to develop robust regression procedures via Huberization. Our main substantive application of generalized random forests involves heterogeneous treatment effect estimation with instrumental variables.

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 (1) in addition to standard conditional mean estimation. This paper addresses the resulting conceptual and methodological challenges and establishes formal asymptotic results.

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  $\hat{\mu}_b(x)$ ,  $\hat{\mu}(x) = B^{-1} \sum_{b=1}^B \hat{\mu}_b(x)$ ; and, because individual trees  $\hat{\mu}_b(x)$  have low bias but high variance, such averaging meaningfully stabilizes predictions ([Bühlmann and Yu, 2002](#); [Scornet, Biau and Vert, 2015](#)). However, noisy solutions to moment equations as in (1) are generally biased, and averaging would do nothing to alleviate the bias.

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 (1) using these neighbors. Section 2.1 gives a detailed treatment of this perspective. This locally weighting view of random forests was previously advocated by [Hothorn et al. \(2004\)](#) in the context of survival analysis and by [Meinshausen \(2006\)](#) for quantile regression, and also underlies theoretical analyses of regression forests (e.g., [Lin and Jeon, 2006](#)). For conditional mean estimation, the averaging and weighting views of forests are equivalent; however, once we move to more general settings, the weighting-based perspective proves substantially more effective, and also brings forests closer to the literature on local maximum likelihood estimation ([Fan and Gijbels, 1996](#); [Loader, 1999](#); [Newey, 1994a](#); [Stone, 1977](#); [Tibshirani and Hastie, 1987](#)).

A 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 is looking at how the treatment affects long-term survival, and the other at its effect on the length of hospital stays. It is plausible that the treatment heterogeneity in each setting would be based on disparate 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 on a case by case basis. 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 [Gordon and Olshen \(1985\)](#), [LeBlanc and Crowley \(1992\)](#), [Molinaro, Dudoit and Van der Laan \(2004\)](#) as well as several others for survival analysis. [Zeileis, Hothorn and Hornik \(2008\)](#) propose a method for constructing a single tree for general maximum likelihood problems, where splitting is based on hypothesis tests for improvements in 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](#)).

Moment conditions of the form (1) typically arise in scientific applications where rigorous statistical inference is required. 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 analysis is motivated by classical results for local estimating equations, in particular [Newey \(1994a\)](#), paired with machinery from [Wager and Athey \(2018\)](#) to address the adaptivity of the random forest weighting function.

The resulting framework presents a flexible method for non-parametric statistical estimation and inference with formal asymptotic guarantees. In this paper, we develop applications to quantile regression, conditional average partial effect estimation and heterogeneous treatment effect estimation with instrumental variables; however, there are many other popular statistical models that fit directly into our framework, including panel regression, Huberized robust regression, models of con-

sumer 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, 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, we exactly recover a regression forest. A high-performance software implementation of generalized random forests, `grf` for R and C++, is available from CRAN.

1.1. *Related Work.* The idea of local maximum likelihood (and local generalized method of moments) estimation has a long history, including [Fan, Farmen and Gijbels \(1998\)](#), [Newey \(1994a\)](#), [Staniswalis \(1989\)](#), [Stone \(1977\)](#), [Tibshirani and Hastie \(1987\)](#) and [Lewbel \(2007\)](#). In economics, popular applications of these techniques include multinomial choice modeling in a panel data setting (e.g., [Honoré and Kyriazidou, 2000](#)) and instrumental variables regression (e.g., [Su, Murtazashvili and Ullah, 2013](#)). The core 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, performance can suffer due to the “curse of dimensionality” (e.g., [Robins and Ritov, 1997](#)).

Our paper replaces 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 proposals of [Hothorn et al. \(2004\)](#) and [Meinshausen \(2006\)](#) for forest-based survival analysis and quantile regression. This adaptive nearest neighbors perspective also underlies several statistical analyses of random forests, including [Arlot and Genuer \(2014\)](#), [Biau and Devroye \(2010\)](#), and [Lin and Jeon \(2006\)](#).

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, Hothorn and Hornik \(2008\)](#) 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, Biau and Vert, 2015](#); [Wager and Athey, 2018](#)). In this sense, our approach is related to gradient boosting ([Friedman, 2001](#)), which uses gradient-based approximations to

guide a greedy, 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, Devroye and Lugosi, 2008; Biau and Scornet, 2016; Bühlmann and Yu, 2002; Denil, Matheson and De Freitas, 2014; Geurts, Ernst and Wehenkel, 2006; Ishwaran and Kogalur, 2010; Lin and Jeon, 2006; Meinshausen, 2006; Mentch and Hooker, 2016; Scornet, Biau and Vert, 2015; Sexton and Laake, 2009; Wager and Athey, 2018; Wager and Walther, 2015; Zhu, Zeng and Kosorok, 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.

**2. Generalized Random 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 is chosen to maximize the improvement to model fit (Breiman et al., 1984); see Figure 4 in the Appendix 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. For an introductory overview of random forests, we recommend the chapter of Hastie, Tibshirani and Friedman (2009) dedicated to the method. As discussed below, in generalizing random forests, we preserve several core elements of Breiman’s forests—including recursive partitioning, subsampling, and random split selection—but we abandon the idea that our final estimate is obtained by averaging estimates from each member of an ensemble. Treating forests as a type of adaptive nearest neighbor estimator is much more amenable to statistical extensions.

**2.1. Forest-Based Local Estimation.** Suppose that we have  $n$  independent and identically distributed samples, indexed  $i = 1, \dots, n$ . For each sample, we have access to an observable quantity  $O_i$  that encodes information 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, it may contain richer information. 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  $\mathbb{E}[\psi_{\theta(x), \nu(x)}(O_i) \mid X_i = x] = 0$  for all  $x \in \mathcal{X}$ , where  $\theta(x)$  is the parameter we care about and  $\nu(x)$  is an optional nuisance parameter.

One 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

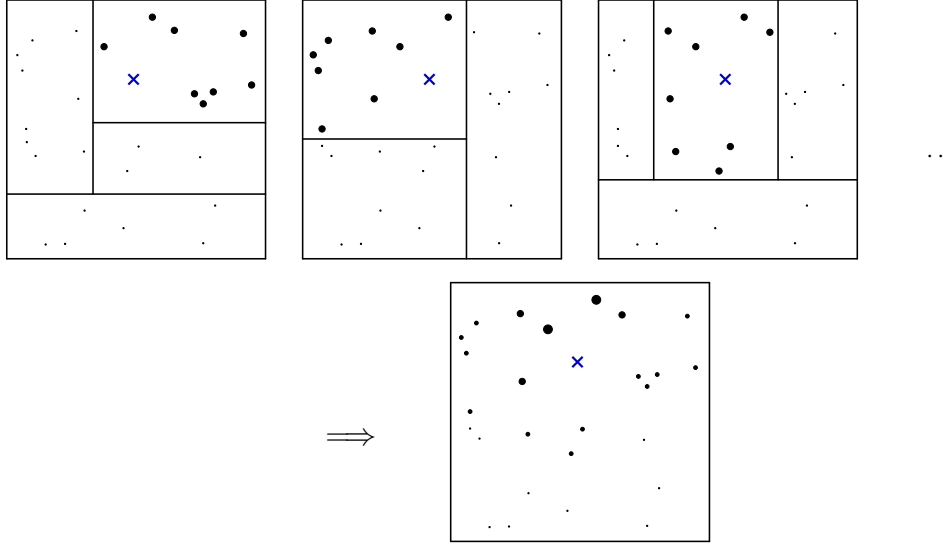


Fig 1: Illustration of the random forest weighting function. The rectangles depicted above correspond to terminal nodes in the dendrogram representation of Figure 4. Each tree starts by giving equal (positive) weight to the training examples in the same leaf as our test point  $x$  of interest, and zero 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  $x$ .

estimating equation (Fan, Farnen and Gijbels, 1998; Newey, 1994a; Staniswalis, 1989; Stone, 1977; Tibshirani and Hastie, 1987):

$$(2) \quad \left( \hat{\theta}(x), \hat{\nu}(x) \right) \in \operatorname{argmin}_{\theta, \nu} \left\{ \left\| \sum_{i=1}^n \alpha_i(x) \psi_{\theta, \nu}(O_i) \right\|_2 \right\}.$$

When the above expression has a unique root, we can simply say that  $(\hat{\theta}(x), \hat{\nu}(x))$  solves  $\sum_{i=1}^n \alpha_i(x) \psi_{\hat{\theta}(x), \hat{\nu}(x)}(O_i) = 0$ . 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, Tibshirani and Friedman, 2009). Although methods of the above kind often work well in low dimensions, they are sensitive to the curse of dimensionality.

Here, we seek to use forest-based algorithms to adaptively learn better, problem-specific, weights  $\alpha_i(x)$  that can be used in conjunction with (2). As in Hothorn et al. (2004) and 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  $b = 1, \dots, B$  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  $x$ :

$$(3) \quad \alpha_{bi}(x) = \frac{\mathbf{1}(\{X_i \in L_b(x)\})}{|L_b(x)|}, \quad \alpha_i(x) = \frac{1}{B} \sum_{b=1}^B \alpha_{bi}(x).$$

These weights sum to 1, and define the forest-based adaptive neighborhood of  $x$ ; see Figure 1 for an illustration of this weighting function.

There are some subtleties in how the sets  $L_b(x)$  are defined—in particular, as discussed in Section 2.4, 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 (2) with weights (3).

Finally, for the special case of regression trees, our weighting-based definition of a random forest is equivalent to the standard “average of trees” perspective taken in Breiman (2001): If we estimate the conditional mean function  $\mu(x) = \mathbb{E}[Y_i | X_i = x]$ , as identified in (1) using  $\psi_{\mu(x)}(Y_i) = Y_i - \mu(x)$ , then we see that  $\sum_{i=1}^n \frac{1}{B} \sum_{b=1}^B \alpha_{bi}(x) (Y_i - \hat{\mu}(x)) = 0$  if and only if  $\hat{\mu}(x) = \frac{1}{B} \sum_{b=1}^B \hat{\mu}_b(x)$ , 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.

**2.2. Splitting to Maximize Heterogeneity.** We seek trees that, when combined into a forest, induce weights  $\alpha_i(x)$  that lead to good estimates of  $\theta(x)$ . The main difference between random forests 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 mimics 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 (we suppress dependence on  $\mathcal{J}$  when unambiguous):

$$(4) \quad (\hat{\theta}_P, \hat{\nu}_P)(\mathcal{J}) \in \operatorname{argmin}_{\theta, \nu} \left\{ \left\| \sum_{\{i \in \mathcal{J}: X_i \in P\}} \psi_{\theta, \nu}(O_i) \right\|_2 \right\}.$$

We would like to divide  $P$  into two children  $C_1, C_2 \subseteq \mathcal{X}$  using an axis-aligned cut such as to improve the accuracy of our  $\theta$ -estimates as much as possible; formally, we seek to minimize  $\operatorname{err}(C_1, C_2)$  defined as  $\operatorname{err}(C_1, C_2) = \sum_{j=1,2} \mathbb{P}[X \in C_j | X \in P] \mathbb{E}[(\hat{\theta}_{C_j}(\mathcal{J}) - \theta(X))^2 | X \in C_j]$ , where  $\hat{\theta}_{C_j}(\mathcal{J})$  are fit over children  $C_j$  in analogy to (4), and expectations are taken over both the randomness in  $\hat{\theta}_{C_j}(\mathcal{J})$  and a new test point  $X$ .



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  $\text{err}(C_1, C_2)$  with plug-in estimators from the training sample. In the case of estimating the effect of a binary treatment, Athey and Imbens (2016) study sample-splitting trees, and propose an unbiased, model-free estimate of  $\text{err}(C_1, C_2)$  using an overfitting penalty 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  $\text{err}(C_1, C_2)$ . 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 3 hold, and that the parent node  $P$  has a radius smaller than  $r$  for some value  $r > 0$ . We write  $n_P = |\{i \in \mathcal{J} : X_i \in P\}|$  for the number of observations in the parent and  $n_{C_j}$  for the number of observations in each child, and define*

$$(5) \quad \Delta(C_1, C_2) := n_{C_1} n_{C_2} / n_P^2 \left( \hat{\theta}_{C_1}(\mathcal{J}) - \hat{\theta}_{C_2}(\mathcal{J}) \right)^2,$$

where  $\hat{\theta}_{C_1}$  and  $\hat{\theta}_{C_2}$  are solutions to the estimating equation computed in the children, following (4). Then, treating the child nodes  $C_1$  and  $C_2$  as well as the corresponding counts  $n_{C_1}$  and  $n_{C_2}$  as fixed, and assuming that  $n_{C_1}, n_{C_2} \gg r^{-2}$ , we have  $\text{err}(C_1, C_2) = K(P) - \mathbb{E}[\Delta(C_1, C_2)] + o(r^2)$  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  $\Delta$ -criterion (5) 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  $\Delta$ -criterion as favoring splits that increase the heterogeneity of the in-sample  $\theta$ -estimates as fast as possible. The dominant bias term in  $\text{err}(C_1, C_2)$  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.

**2.3. The Gradient Tree Algorithm.** The above discussion provides conceptual guidance on how to pick good splits. But 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 (4) may be quite expensive computationally. To avoid this issue, we instead optimize an approximate criterion  $\tilde{\Delta}(C_1, C_2)$  built using gradient-based approximations for  $\hat{\theta}_{C_1}$  and  $\hat{\theta}_{C_2}$ . For each child  $C$ , we use  $\hat{\theta}_C \approx \theta_C$  as follows: We first compute  $A_P$  as any consistent estimate for the gradient of the



expectation of the  $\psi$ -function, i.e.,  $\nabla \mathbb{E}[\psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) \mid X_i \in P]$ , and then set

$$(6) \quad \tilde{\theta}_C = \hat{\theta}_P - \frac{1}{|\{i : X_i \in C\}|} \sum_{\{i : X_i \in C\}} \xi^\top A_P^{-1} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i),$$

where  $\hat{\theta}_P$  and  $\hat{\nu}_P$  are obtained by solving (4) once in the parent node, and  $\xi$  is a vector that picks out the  $\theta$ -coordinate from the  $(\theta, \nu)$  vector. When the  $\psi$ -function itself is continuously differentiable, we use

$$(7) \quad A_P = \frac{1}{|\{i : X_i \in P\}|} \sum_{\{i : X_i \in P\}} \nabla \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i),$$

and 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  $\psi$  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 (4), and use them to get pseudo-outcomes

$$(8) \quad \rho_i = -\xi^\top A_P^{-1} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) \in \mathbb{R}.$$

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

$$(9) \quad \tilde{\Delta}(C_1, C_2) = \sum_{j=1}^2 \frac{1}{|\{i : X_i \in C_j\}|} \left( \sum_{\{i : X_i \in C_j\}} \rho_i \right)^2.$$

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

For intuition, it is helpful to examine the simplest case of least-squares regression, i.e., with  $\psi_{\theta(x)}(Y) = Y - \theta(x)$ . Here, the labeling step (8) doesn't change anything—we get  $\rho_i = Y_i - \bar{Y}_P$ , where  $\bar{Y}_P$  is the mean outcome in the parent—while the second step maximizing (9) corresponds to the usual way of making splits as in Breiman (2001). Thus, the special structure of the type of problem we are trying to solve is encoded in (8), while the second scanning step is a universal step shared across all different types of forests.

We expect this approach to provide more consistent computational performance than optimizing (5) 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 (8) is only solved once per node, and so is less performance sensitive). From this perspective,

Then, (9) is equal to (9.13) in Hastie et al. (2009)

**Algorithm 1** Generalized random forest with honesty and subsampling

All tuning parameters are pre-specified, including the number of trees  $B$  and the sub-sampling  $s$  rate used in SUBSAMPLE. This function is implemented in the package **grf** for R and C++.

```

1: procedure GENERALIZEDRANDOMFOREST(set of examples  $\mathcal{S}$ , test point  $x$ )
2:   weight vector  $\alpha \leftarrow \text{ZEROS}(|\mathcal{S}|)$ 
3:   for  $b = 1$  to total number of trees  $B$  do
4:     set of examples  $\mathcal{I} \leftarrow \text{SUBSAMPLE}(\mathcal{S}, s)$ 
5:     sets of examples  $\mathcal{J}_1, \mathcal{J}_2 \leftarrow \text{SPLITSAMPLE}(\mathcal{I})$ 
6:     tree  $\mathcal{T} \leftarrow \text{GRADIENTTREE}(\mathcal{J}_1, \mathcal{X})$  ▷ See Algorithm 2.
7:      $\mathcal{N} \leftarrow \text{NEIGHBORS}(x, \mathcal{T}, \mathcal{J}_2)$  ▷ Returns those elements of  $\mathcal{J}_2$  that fall into  
the same leaf as  $x$  in the tree  $\mathcal{T}$ .
8:     for all example  $e \in \mathcal{N}$  do
9:        $\alpha[e] += 1/|\mathcal{N}|$ 
10:  output  $\hat{\theta}(x)$ , the solution to (2) with weights  $\alpha/B$ 

```

The function ZEROS creates a vector of zeros of length  $|\mathcal{S}|$ ; SUBSAMPLE draws a subsample of size  $s$  from  $\mathcal{S}$  without replacement; and SPLITSAMPLE randomly divides a set into two evenly-sized, non-overlapping halves. The step (2) can be solved using any numerical estimator. Our implementation **grf** provides an explicit plug-in point where a user can write a solver for (2) appropriate for their  $\psi$ -function.  $\mathcal{X}$  is the domain of the  $X_i$ . In our analysis, we consider a restricted class of generalized random forests satisfying Specification 1.

using a regression splitting criterion as in (9) 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 (5) may require solving 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, Hothorn and Hornik (2008). Conceptually, tree splitting bears some connection to change-point detection if we imagine tree splits as occurring at detected change-points in  $\theta(x)$ ; and, from this perspective, our approach 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).

In our context, we can verify that the error from using the approximate criterion (9) instead of the exact  $\Delta$ -criterion (5) is within the tolerance used to motivate the  $\Delta$ -criterion in Proposition 1, thus suggesting that our use of (6) to guide splitting may not result in too much inefficiency. Note that consistent estimates of  $A_P$  can in general be derived directly via, e.g., (7), without relying on Proposition 2.

**PROPOSITION 2.** *Under the conditions of Proposition 1, if  $|A_P - \nabla \mathbb{E}[\psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) | X_i \in P]| \rightarrow_P 0$ , i.e.,  $A_P$  is consistent, then  $\Delta(C_1, C_2)$  and  $\tilde{\Delta}(C_1, C_2)$  are approximately equivalent, in that  $\tilde{\Delta}(C_1, C_2) = \Delta(C_1, C_2) + o_P(\max\{r^2, 1/n_{C_1}, 1/n_{C_2}\})$ .*

**Algorithm 2** Gradient tree

Gradient trees are grown as subroutines of a generalized random forest.

```

1: procedure GRADIENTTREE(set of examples  $\mathcal{J}$ , domain  $\mathcal{X}$ )
2:   node  $P_0 \leftarrow \text{CREATENODE}(\mathcal{J}, \mathcal{X})$ 
3:   queue  $\mathcal{Q} \leftarrow \text{INITIALIZEQUEUE}(P_0)$ 
4:   while NOTNULL(node  $P \leftarrow \text{POP}(\mathcal{Q})$ ) do
5:      $(\hat{\theta}_P, \hat{\nu}_P, A_P) \leftarrow \text{SOLVEESTIMATINGEQUATION}(P)$  ▷ Computes (4) and (7).
6:     vector  $R_P \leftarrow \text{GETPSEUDOOUTCOMES}(\hat{\theta}_P, \hat{\nu}_P, A_P)$  ▷ Applies (8) over  $P$ .
7:     split  $\Sigma \leftarrow \text{MAKECARTSPLIT}(P, R_P)$  ▷ Optimizes (9).
8:     if SPLITSUCCCEEDED( $\Sigma$ ) then
9:       SETCHILDREN( $P$ , GETLEFTCHILD( $\Sigma$ ), GETRIGHTCHILD( $\Sigma$ ))
10:      ADDTOQUEUE( $\mathcal{Q}$ , GETLEFTCHILD( $\Sigma$ ))
11:      ADDTOQUEUE( $\mathcal{Q}$ , GETRIGHTCHILD( $\Sigma$ ))
12:   output tree with root node  $P_0$ 

```

The function call INITIALIZEQUEUE initializes a queue with a single element; POP returns and removes the oldest element of a queue  $\mathcal{Q}$ , unless  $\mathcal{Q}$  is empty in which case it returns null. MAKECARTSPLIT runs a CART split on the pseudo-outcomes, and either returns two child nodes or a failure message that no legal split is possible.

*2.4. 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 (2) paired with the forest weights (3). We expect each tree to provide small, relevant neighborhoods for  $x$  that give us noisy estimates of  $\theta(x)$ ; then, 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 good statistical behavior, 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, Biau and Vert, 2015; Wager and Athey, 2018), and a sub-sample splitting technique that we call honesty (Biau, 2012; Denil, Matheson and De Freitas, 2014; Wager and Athey, 2018). 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 (2018) for a more in-depth discussion of honesty in the context of forests. As shown in Section 3, 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. Asymptotic Analysis.** We now turn to a formal characterization of generalized random forests, with the aim of establishing asymptotic Gaussianity of the  $\hat{\theta}(x)$ , and of providing tools for statistical inference about  $\theta(x)$ . We first list assumptions underlying our theoretical results. Throughout, the covariate space and the parameter space are both subsets of Euclidean space; specifically,  $\mathcal{X} = [0, 1]^p$  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  $\infty$ ; 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. All proofs are in the appendix.

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

$$(10) \quad M_{\theta, \nu}(x) := \mathbb{E} [\psi_{\theta, \nu}(O) \mid X = x],$$

varies smoothly in the parameters, even though  $\psi$  itself may be discontinuous. For example, with quantile regression  $\psi_{\theta}(Y) = 1(\{Y > \theta\}) - (1 - q)$  is discontinuous in  $q$ , but  $M_{\theta}(x) = \mathbb{P}[Y > \theta \mid X = x] - (1 - q)$  will be smooth whenever  $Y \mid X = x$  has a smooth density.

**ASSUMPTION 1** (Lipschitz  $x$ -signal). For fixed values of  $(\theta, \nu)$ , we assume that  $M_{\theta, \nu}(x)$  as defined in (10) is Lipschitz continuous in  $x$ .

**ASSUMPTION 2** (Smooth identification). When  $x$  is fixed, we assume that the  $M$ -function is twice continuously differentiable in  $(\theta, \nu)$  with a uniformly bounded second derivative, and that  $V(x) := V_{\theta(x), \nu(x)}(x)$  is invertible for all  $x \in \mathcal{X}$ , with  $V_{\theta, \nu}(x) := \partial/\partial(\theta, \nu) M_{\theta, \nu}(x) \mid_{\theta(x), \nu(x)}$ .

Our next two assumptions control regularity properties of the  $\psi$ -function itself. Assumption 3 holds trivially when  $\psi$  itself is Lipschitz in  $(\theta, \nu)$  (in fact, having  $\psi$  be 0.5-Hölder would be enough), while Assumption 4 is used to show that a certain empirical process is Donsker. Examples are given at the end of this section.

**ASSUMPTION 3** (Lipschitz  $(\theta, \nu)$ -variogram). The score functions  $\psi_{\theta, \nu}(O_i)$  have a continuous covariance structure. Writing  $\gamma$  for the worst-case variogram and  $\|\cdot\|_F$  for the Frobenius norm, then for some  $L > 0$ ,

$$(11) \quad \begin{aligned} \gamma \left( \begin{pmatrix} \theta \\ \nu \end{pmatrix}, \begin{pmatrix} \theta' \\ \nu' \end{pmatrix} \right) &\leq L \left\| \begin{pmatrix} \theta \\ \nu \end{pmatrix} - \begin{pmatrix} \theta' \\ \nu' \end{pmatrix} \right\|_2 \text{ for all } (\theta, \nu), (\theta', \nu'), \\ \gamma \left( \begin{pmatrix} \theta \\ \nu \end{pmatrix}, \begin{pmatrix} \theta' \\ \nu' \end{pmatrix} \right) &:= \sup_{x \in \mathcal{X}} \left\{ \left\| \text{Var} [\psi_{\theta, \nu}(O_i) - \psi_{\theta', \nu'}(O_i) \mid X_i = x] \right\|_F \right\}. \end{aligned}$$

**ASSUMPTION 4** (Regularity of  $\psi$ ). The  $\psi$ -functions can be written as  $\psi_{\theta, \nu}(O) = \lambda(\theta, \nu; O_i) + \zeta_{\theta, \nu}(g(O_i))$ , such that  $\lambda$  is Lipschitz-continuous in  $(\theta, \nu)$ ,  $g : \{O_i\} \rightarrow \mathbb{R}$  is a univariate summary of  $O_i$ , and  $\zeta_{\theta, \nu} : \mathbb{R} \rightarrow \mathbb{R}$  is any family of monotone and bounded functions.

**ASSUMPTION 5** (Existence of solutions). We assume that, for any weights  $\alpha_i$  with  $\sum \alpha_i = 1$ , the estimating equation (2) returns a minimizer  $(\hat{\theta}, \hat{\nu})$  that at least

approximately solves the estimating equation:  $\|\sum_{i=1}^n \alpha_i \psi_{\hat{\theta}, \hat{\nu}}(O_i)\|_2 \leq C \max \{\alpha_i\}$ , for some constant  $C \geq 0$ .

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 negative sub-gradient of a convex function, and the expected score  $M_{\theta, \nu}(X_i)$  is the negative gradient of a strongly convex function.

Finally, our consistency and Gaussianity results require using some specific settings for the trees from Algorithm 1. In particular, we require that all trees be honest and regular in the sense of [Wager and Athey \(2018\)](#), as follows. In order to satisfy the minimum split probability condition below, our implementation relies on the device of [Denil, Matheson and De Freitas \(2014\)](#), whereby the number splitting variables considered at each step of the algorithm is random; specifically, we try  $\min \{\max \{\text{Poisson}(m), 1\}, p\}$  variables at each step, where  $m > 0$  is a tuning parameter.

**SPECIFICATION 1.** All trees are symmetric, in that their output is invariant to permuting the indices of training examples; make balanced splits, in the sense that every split puts at least a fraction  $\omega$  of the observations in the parent node into each child, for some  $\omega > 0$ ; and are 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  $\pi > 0$ . The forest is honest and built via subsampling with subsample size  $s$  satisfying  $s/n \rightarrow 0$  and  $s \rightarrow \infty$ , as described in Section 2.4.

For generality, we set up Assumptions 1–6 in an abstract way. 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.,  $\psi_{\theta}(Y_i) = Y_i - \theta$ , Assumptions 2–6 hold immediately from the definition of  $\psi$ . In particular,  $V = 1$  in Assumption 2,  $\gamma(\theta, \theta') = 0$  in Assumption 3,  $\psi$  itself is Lipschitz for Assumption 4, and  $\psi_{\theta}(y) = -\frac{d}{d\theta}(y - \theta)^2/2$  for Assumption 6. Meanwhile, Assumption 1 simply means that the conditional mean function  $\mathbb{E}[Y_i | X_i = x]$  must be Lipschitz in  $x$ ; this is a standard assumption in the literature on regression forests.

EXAMPLE 2 (Quantile regression). For quantile regression, we have  $\psi_\theta(Y_i) = q - \mathbf{1}(\{Y_i \leq \theta\})$  and  $M_\theta(x) = q - F_x(\theta)$ , where  $F_x(\cdot)$  denotes the cumulative distribution function of  $Y_i$  given  $X_i = x$ . Assumption 1 is equivalent to assuming that the conditional exceedance probabilities  $\mathbb{P}[Y_i > y | X_i = x]$  be Lipschitz-continuous in  $x$  for all  $y \in \mathbb{R}$ , while 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  $y = F_x^{-1}(q)$ . Assumption 3 holds if  $f_x(y)$  is uniformly bounded from above (specifically,  $\gamma(\theta, \theta') \leq \max_x \{f_x(y)\} |\theta - \theta'|$ ), Assumption 4 holds because  $\psi$  is monotone and  $O_i = Y_i$  is univariate, Assumption 5 is immediate, and Assumption 6 holds because  $-\frac{d}{d\theta} M_\theta(x) = f_x(\theta) > 0$  and  $\psi_\theta(Y_i)$  is the negative sub-gradient of a V-shaped function with elbow at  $Y_i$ .

3.1. *A Central Limit Theorem for Generalized Random Forests.* Given these assumptions, we are now ready to provide an asymptotic characterization of generalized random forests. In doing so, we note that existing asymptotic analyses of regression forests, including Mentch and Hooker (2016), Scornet, Biau and Vert (2015) and Wager and Athey (2018), 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  $\hat{\theta}(x)$  from 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 (2). 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 (1994a). 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  $\theta(x)$ ,  $\rho_i^*(x) := -\xi^\top V(x)^{-1} \psi_{\theta(x), \nu(x)}(O_i)$ . These quantities are closely related to the pseudo-outcomes (8) used in our gradient tree splitting rule; the main difference is that, here, the  $\rho_i^*(x)$  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 (2), we can also define a pseudo-forest

$$(12) \quad \tilde{\theta}^*(x) := \theta(x) + \sum_{i=1}^n \alpha_i(x) \rho_i^*(x),$$

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. 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  $\tilde{\theta}^*(x) = \frac{1}{B} \sum_{b=1}^B \tilde{\theta}_b^*(x)$  with  $\tilde{\theta}_b^*(x) = \sum_{i=1}^n \alpha_{ib}(x) (\theta(x) + \rho_i^*(x))$ . Then, because each individual pseudo-tree prediction  $\tilde{\theta}_b^*(x)$  is trained on a size- $s$  subsample of the training data drawn without replacement (see Section 2.4),  $\tilde{\theta}^*(x)$  is an infinite-order  $U$ -statistic whose order corresponds to the subsample size, and so the arguments of [Mentch and Hooker \(2016\)](#) or [Wager and Athey \(2018\)](#) can be used to study the averaged estimator  $\tilde{\theta}^*(x)$  using 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  $\hat{\theta}(x)$ . To do so, we start by establishing consistency of  $\hat{\theta}(x)$  for  $\theta(x)$  given our assumptions; we note that this is the only point in the paper where we use the fact that  $\psi$  is the negative gradient of a convex loss as in Assumption 6.

**THEOREM 3.** *Given Assumptions 1–6, estimates  $(\hat{\theta}(x), \hat{\nu}(x))$  from a forest satisfying Specification 1 converge in probability to  $(\theta(x), \nu(x))$ .*

Building on this consistency result, we obtain a coupling of the desired type in Lemma 4, the main technical contribution of this paper. 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 \(2000\)](#). The remainder of our analysis assumes that trees are grown on subsamples of size  $s$  scaling as  $s = n^\beta$  for some  $\beta_{\min} < \beta < 1$ , with

$$(13) \quad \beta_{\min} := 1 - \left(1 + \pi^{-1} (\log(\omega^{-1})) / \left(\log((1-\omega)^{-1})\right)\right)^{-1} < \beta < 1,$$

where  $\pi$  and  $\omega$  are as in Specification 1. This scaling guarantees that the errors of forests are variance-dominated.

**LEMMA 4.** *Given Assumptions 1–5, and a forest trained according to Specification 1 with (13), 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, where  $s$ ,  $\pi$  and  $\omega$  are as in Specification 1:*

$$(14) \quad \sqrt{\frac{n}{s}} (\tilde{\theta}^*(x) - \hat{\theta}(x)) = \mathcal{O}_P \left( \max \left\{ s^{-\frac{\pi}{2} \frac{\log((1-\omega)^{-1})}{\log(\omega^{-1})}}, \left(\frac{s}{n}\right)^{\frac{1}{6}} \right\} \right).$$

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 \(2018\)](#) apply directly to this object, and can be used to establish its Gaussianity. That we cannot actually compute  $\tilde{\theta}^*(x)$  does not hinder an application of their results. Pursuing this approach, we find that given (13),  $\tilde{\theta}^*(x)$  and  $\hat{\theta}(x)$  are both



asymptotically normal. By extending the same argument, we could also show that the nuisance parameter estimates  $\hat{\nu}(x)$  are consistent and asymptotically normal; however, we caution that the tree splits are not necessarily targeted to expressing heterogeneity in  $\nu(x)$ , and so the resulting  $\hat{\nu}(x)$  may not be particularly accurate in finite samples.

**THEOREM 5.** *Suppose Assumptions 1–6 and a forest trained according to Specification 1 with trees are grown on subsamples of size  $s = n^\beta$  satisfying (13). Finally, suppose that  $\text{Var}[\rho_i^*(x) | X_i = x] > 0$ . Then, there is a sequence  $\sigma_n(x)$  for which  $(\hat{\theta}_n(x) - \theta(x)) / \sigma_n(x) \Rightarrow \mathcal{N}(0, 1)$  and  $\sigma_n^2(x) = \text{polylog}(n/s)^{-1} s/n$ , where  $\text{polylog}(n/s)$  is a function that is bounded away from 0 and increases at most polynomially with the log-inverse sampling ratio  $\log(n/s)$ .*

**4. Confidence Intervals via the Delta Method.** Theorem 5 can also be used for statistical inference about  $\theta(x)$ . Given any consistent estimator  $\hat{\sigma}_n(x)/\sigma_n(x) \rightarrow_p 1$  of the noise scale of  $\hat{\theta}_n(x)$ , Theorem 5 can be paired with Slutsky's lemma to verify that  $\lim_{n \rightarrow \infty} \mathbb{E}[\theta(x) \in (\hat{\theta}_n(x) \pm \Phi^{-1}(1 - \alpha/2)\hat{\sigma}_n(x))] = \alpha$ . 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  $\text{Var}[\tilde{\theta}^*(x)]/\sigma_n^2(x) \rightarrow_p 1$ , and so it again suffices to study  $\tilde{\theta}^*(x)$ . Moreover, from the definition of  $\tilde{\theta}^*(x)$ , we directly see that

$$(15) \quad \text{Var}[\tilde{\theta}^*(x)] = \xi^\top V(x)^{-1} H_n(x; \theta(x), \nu(x)) (V(x)^{-1})^\top \xi,$$

where  $H_n(x; \theta, \nu) = \text{Var}[\sum_{i=1}^n \alpha_i(x) \psi_{\theta, \nu}(O_i)]$ . Thus, we propose building Gaussian confidence intervals using

$$(16) \quad \hat{\sigma}_n^2(x) := \xi^\top \hat{V}_n(x)^{-1} \hat{H}_n(x) (\hat{V}_n(x)^{-1})^\top \xi,$$

where  $\hat{V}_n(x)$  and  $\hat{H}_n(x)$  are consistent estimators for the quantities in (15).

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

$$(17) \quad V(x) = \begin{pmatrix} \mathbb{E}[Z_i W_i | X_i = x] & \mathbb{E}[Z_i | X_i = x] \\ \mathbb{E}[W_i | X_i = x] & 1 \end{pmatrix},$$

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 entries of (17) 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, Hastie and Efron \(2014\)](#); and, in principle, any of these methods could be adapted to estimate the variance of  $\Psi$ . The only difficulty is that  $\Psi$  depends on the true parameter values  $(\theta(x), \nu(x))$ , and so cannot directly be accessed in practice. Here, we present results 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.

*4.1. 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](#)) (throughout this section, we assume that  $s \leq \lfloor n/2 \rfloor$ ). In our context, the ideal half-sampling estimator would be  $\hat{H}_n^{HS}(x)$  defined as

$$(18) \quad \left( \binom{n}{\lfloor n/2 \rfloor} \right)^{-1} \sum_{\{\mathcal{H}: |\mathcal{H}| = \lfloor \frac{n}{2} \rfloor\}} \left( \Psi_{\mathcal{H}}(\hat{\theta}(x), \hat{\nu}(x)) - \Psi(\hat{\theta}(x), \hat{\nu}(x)) \right)^2,$$

where  $\Psi_{\mathcal{H}}$  denotes a version of  $\Psi$  computed only using all the possible trees that only rely on data from the half sample  $\mathcal{H} \subset \{1, \dots, n\}$  (specifically, in terms of Algorithm 1, we only use trees whose full  $\mathcal{I}$ -subsample is contained in  $\mathcal{H}$ ). If we could evaluate  $\hat{H}_n^{HS}(x)$ , results from [Efron \(1982\)](#) suggest that it would be a good variance estimator for  $\Psi$ , but doing so is effectively impossible computationally as it would require growing very many forests.

Following [Sexton and Laake \(2009\)](#), however, we can efficiently approximate  $\hat{H}_n^{HS}(x)$  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. Then, we grow our forest as follows: First draw  $g = 1, \dots, B/\ell$  random half-samples  $\mathcal{H}_g \subset \{1, \dots, n\}$  of size  $\lfloor n/2 \rfloor$ , and then generate the subsamples  $\mathcal{I}_b$  used to build the forest in Algorithm 1 such that  $\mathcal{I}_b \subseteq \mathcal{H}_{\lfloor b/\ell \rfloor}$  for each  $b = 1, \dots, B$ . In other words, we now generate our forest using little bags of  $\ell$  trees, where all the trees in a given bag only use data from the same half-sample. [Sexton and Laake \(2009\)](#) discuss optimal choices of  $\ell$  for minimizing Monte Carlo error, and show that they depend on the ratio of the sampling variance of a single tree to that of the full forest.

The upshot of this construction is that we can now identify  $\hat{H}_n^{HS}(x)$  using a simple variance decomposition. Writing  $\Psi_b$  for a version of  $\Psi$  computed only using the  $b$ -th tree, we can verify that  $\hat{H}_n^{HS}(x)$  can be expressed in terms of the “between groups”

and “within group” variance terms,

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

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

The result below verifies that, under the conditions of Theorem 5, the optimal half-sampling estimator  $\hat{H}_n^{HS}(x)$  with plug-in values for  $(\hat{\theta}(x), \hat{\nu}(x))$  as in (18) consistently estimates the sampling variance of  $\Psi(\theta(x), \nu(x))$ . We have already seen above that the computationally feasible estimator  $\hat{H}_n^{BLB}(x)$  will match  $\hat{H}_n^{HS}(x)$  whenever  $B$  is large enough and so, given any consistent estimator  $\hat{V}_n(x)$  for  $V(x)$ , we find that the confidence intervals built using (16) will be asymptotically valid.

**THEOREM 6.** *Given the conditions of Theorem 5,  $\hat{H}_n^{HS}(x)$  is consistent,  $\|\hat{H}_n^{HS}(x) - H_n(x; \theta(x), \nu(x))\|_F / \|H_n(x; \theta(x), \nu(x))\|_F \rightarrow_p 0$ . Moreover, given any consistent  $\hat{V}_n(x)$  estimator for  $V(x)$  such that  $\|\hat{V}_n(x) - V(x)\|_F \rightarrow_p 0$ , Gaussian confidence intervals built using (16) will asymptotically have nominal coverage.*

One challenge with the empirical moment estimator based on the above is that, if  $B$  is small, the variance estimates  $\hat{H}_n^{BLB}(x)$  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  $\hat{H}_n^{HS}(x)$  over  $[0, \infty)$ . When  $B$  is large enough, this distinction washes out.

**5. Application: Quantile Regression Forests.** Our first application of generalized random forests is to the classical problem of non-parametric 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 (2) using random forest weights (3). 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 (1), using the moment function  $\psi_{\theta}(Y_i) = q\mathbf{1}(\{Y_i > \theta\}) - (1 - q)\mathbf{1}(\{Y_i \leq \theta\})$ . Plugging this moment function into our splitting scheme (8) gives us pseudo-outcomes  $\rho_i = \mathbf{1}(\{Y_i > \hat{\theta}_{q,P(X_i)}\})$ ,

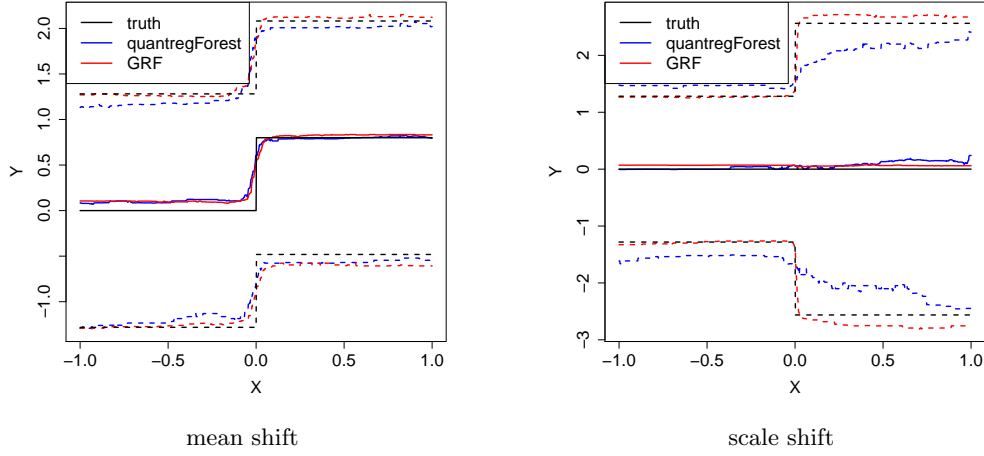


Fig 2: Comparison of quantile regression using generalized random forests and the `quantregForest` package of Meinshausen (2006). In both cases, we have  $n = 2,000$  independent and identically distributed examples where  $X_i$  is uniformly distributed over  $[-1, 1]^p$  with  $p = 40$ , and  $Y_i$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

where  $\hat{\theta}_{q,P(X_i)}$  is the  $q$ -th quantile of the parent node  $P(X_i)$  containing  $X_i$ , 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 try to separate observations that fall above the  $q$ -th quantile of the parent from those below it.

We compare our method to that of Meinshausen (2006) in Figure 2. In the left panel, we have a mean shift in the distribution of  $Y_i$  conditional on  $X_i$  at  $(X_i)_1 = 0$ , and both methods are able to pick it up as expected. However, in the right panel, the mean of  $Y$  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 2.4. 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. 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 found that, our splitting rule is specifically sensitive to quantile shifts in a way that regression splits are not—and, moreover, deriving our splitting rule was fully automatic given the generalized random forest formalism.

In several applications, we want to estimate multiple quantiles at the same time. For example, in Figure 2, we estimate at  $q = 0.1, 0.5, 0.9$ . Estimating different forests for each quantile separately would be undesirable for many reasons: It would be computationally expensive, and 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 our relabeling transformation. Given a set of quantiles of interest  $q_1 < \dots < q_k$ , we first evaluate all these quantiles  $\hat{\theta}_{q_1, P(X_i)} \leq \dots \leq \hat{\theta}_{q_k, P(X_i)}$  in the parent node, and label  $i$ -th point by the interval  $[\hat{\theta}_{q_{j-1}, P(X_i)}, \hat{\theta}_{q_j, P(X_i)})$  it falls into. Then, we choose the split point using a multiclass classification rule that classifies each observation into one of the intervals.

**6. Application: Estimating Conditional Average Partial Effects.** Next, we consider conditional average partial effect estimation 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  $Y_i = W_i \cdot b_i + \varepsilon_i$ ,  $\beta(x) = \mathbb{E}[b_i | X_i = x]$ . Our goal is to estimate  $\theta(x) = \xi \cdot \beta(x)$  for some contrast  $\xi \in \mathbb{R}^p$ . If  $W_i \in \{0, 1\}$  is a treatment assignment, then  $\beta(x)$  corresponds to the conditional average treatment effect.

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, i.e., independent of the unobservables conditionally on  $X_i$ :  $\{b_i, \varepsilon_i\} \perp\!\!\!\perp W_i | X_i$ . If  $W_i$  is a binary treatment, this condition is equivalent to the unconfoundedness assumption used to motivate propensity score methods (Rosenbaum and Rubin, 1983). When exogeneity does not hold, more sophisticated identification strategies are needed (see following section).

**6.1. Growing a Forest.** Our parameter of interest  $\theta(x) = \xi \cdot \beta(x)$  is identified by (1) with  $\psi_{\beta(x), c(x)}(Y_i, W_i) = (Y_i - \beta(x) \cdot W_i - c(x))(1 - W_i)^\top$  where  $c(x)$  is an intercept term; this can also be written more explicitly as  $\theta(x) = \xi^\top \text{Var}[W_i | X_i = x]^{-1} \text{Cov}[W_i, Y_i | X_i = x]$ . Given forest weights  $\alpha_i(x)$  as in (2), the induced estimator  $\hat{\theta}(x)$  for  $\theta(x)$  is

$$(19) \quad \hat{\theta}(x) = \xi^\top \left( \sum_{i=1}^n \alpha_i(x) (W_i - \bar{W}_\alpha)^{\otimes 2} \right)^{-1} \sum_{i=1}^n \alpha_i(x) (W_i - \bar{W}_\alpha) (Y_i - \bar{Y}_\alpha),$$

where  $\bar{W}_\alpha = \sum \alpha_i(x) W_i$  and  $\bar{Y}_\alpha = \sum \alpha_i(x) Y_i$ , and we write  $v^{\otimes 2} = vv^\top$ .

Generalized random forests provide us with a quasi-automatic framework for getting the weights  $\alpha_i(x)$  needed in (19); all that needs to be done is to compute the

pseudo-outcomes  $\rho_i$  from (8) used for recursive partitioning. We use (7) and, for every parent  $P$  and each observation  $i$  with  $X_i \in P$  set

$$(20) \quad \begin{aligned} \rho_i &= \xi^\top A_P^{-1} (W_i - \bar{W}_P) \left( Y_i - \bar{Y}_P - (W_i - \bar{W}_P) \hat{\beta}_P \right), \\ A_P &= \frac{1}{|\{i : X_i \in P\}|} \sum_{\{i : X_i \in P\}} (W_i - \bar{W}_P)^{\otimes 2}, \end{aligned}$$

where now  $\bar{W}_P$  and  $\bar{Y}_P$  stand for averages taken over the parent  $P$ , and  $\hat{\beta}_P$  is the least-squares 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.

Checking the conditions required in Section 3, note that Assumption 1 holds whenever the functions  $\mathbb{E}[Y_i | X_i = x]$ ,  $\mathbb{E}[W_i | X_i = x]$ ,  $\text{Cov}[Y_i, W_i | X_i = x]$  and  $\text{Var}[W_i | X_i = x]$  are all Lipschitz in  $x$ , Assumption 2 holds provided that  $\text{Var}[W_i | X_i = x]$  is invertible, while Assumptions 3–6 hold by construction. Thus, Theorem 5 in fact applies in this setting.

**6.1.1. Local Centering.** The above construction allows for asymptotically valid inference for  $\theta(x)$ , but 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. Writing  $y(x) = \mathbb{E}[Y_i | X = x]$  and  $w(x) = \mathbb{E}[W_i | X = x]$  for the conditional marginal expectations of  $Y_i$  and  $W_i$  respectively, define centered outcomes  $\tilde{Y}_i = Y_i - \hat{y}^{(-i)}(X_i)$  and  $\tilde{W}_i = W_i - \hat{w}^{(-i)}(X_i)$ , where  $\hat{y}^{(-i)}(X_i)$ , etc., are leave-one-out estimates of the marginal expectations, computed without using the  $i$ -th observation. We then run a forest using centered outcomes  $\{\tilde{Y}_i, \tilde{W}_i\}_{i=1}^n$  instead of the original  $\{Y_i, W_i\}_{i=1}^n$ .

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  $x \in \mathcal{S}$ :

$$(21) \quad \begin{aligned} \theta(x) &= \xi^\top \text{Var}[(W_i - \mathbb{E}[W_i | X_i]) | X_i \in \mathcal{S}]^{-1} \\ &\quad \text{Cov}[(W_i - \mathbb{E}[W_i | X_i]), (Y_i - \mathbb{E}[Y_i | X_i]) | X_i \in \mathcal{S}]. \end{aligned}$$

Thus, if we locally center the  $Y_i$  and the  $W_i$  before running our forest, the estimator (19) has the potential to be 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, 1994b; Neyman, 1979); in particular, Robinson (1988) showed that if we have access to a neighborhood  $\mathcal{S}$  over which  $\beta(x) = \beta_{\mathcal{S}}$  is constant, then the moment condition (21) induces a semiparametrically efficient estimator for  $\theta_{\mathcal{S}} = \xi \cdot \beta_{\mathcal{S}}$ .

We note that if we ran a forest with any deterministic centering scheme, i.e., we used  $\tilde{Y}_i = Y_i - \hat{y}(X_i)$  for any Lipschitz function  $\hat{y}(X_i)$  that does not depend on the data, etc., then the theory developed in Section 3 would allow for valid inference about  $\theta(x)$  (in particular, we do not need to assume consistency of  $\hat{y}(X_i)$ ). Moreover, we could also emulate this result by using a form of  $k$ -fold cross-fitting

(Chernozhukov et al., 2016; Schick, 1986). In the context of forests, it is much more practical to carry out residualization via leave-one-out prediction than via  $k$ -fold cross-fitting, because leave-one-out prediction in forests is computationally cheap (Breiman, 2001); however, a practitioner wanting to use results that are precisely covered by theory may prefer to use cross-fitting for centering.

**6.2. Example: Causal Forests.** When  $W_i \in \{0, 1\}$  is a binary treatment assignment, the present setup is equivalent to the standard problem of heterogeneous treatment effect estimation under unconfoundedness. Heterogeneous treatment effect estimation via tree-based methods has received considerable attention in the recent literature: Athey and Imbens (2016) and Su et al. (2009) develop tree-based methods for subgroup analysis, Hill (2011) studies treatment effect estimation via Bayesian additive regression trees (Chipman, George and McCulloch, 2010), and Wager and Athey (2018) propose a causal forest procedure that is very nearly a special case of our generalized random forests. 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 performance of causal forests.

The closest method to ours is Procedure 1 of Wager and Athey (2018), which is almost equivalent to a generalized random forest without centering, the only substantive differences being that they split using the exact loss criterion (5) rather than our gradient-based loss criterion (9), and let each tree compute its own treatment effect estimate rather than using the weighting scheme from Section 2.1 (these methods are exactly equivalent for regression forests, but not for causal forests). Wager and Athey (2018) also consider a second approach, Procedure 2, that obtains its neighborhood function via a classification forest on the treatment assignments  $W_i$ .

A weakness of the methods in Wager and Athey (2018), 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 made it difficult to reconcile their relative advantages. Conversely, given the framing of generalized random forests via estimating equations, it is “obvious” that we can leverage best practices from the literature on estimating equations and orthogonalize our moment conditions by regressing out the main effect of  $X_i$  on  $W_i$  and  $Y_i$  as in Robinson (1988).

To illustrate the value of orthogonalization, we revisit a simulation of Wager and Athey (2018) where  $X_i \sim U([0, 1]^p)$ ,  $W_i | X_i \sim \text{Bernoulli}(e(X_i))$ , and  $Y_i | X_i, W_i \sim \mathcal{N}(m(X_i) + (W_i - 0.5)\tau(X_i), 1)$ . The authors consider two different simulation settings: One with no confounding,  $m(x) = 0$  and  $e(x) = 0.5$ , but with treatment heterogeneity  $\tau(x) = \varsigma(x_1) \varsigma(x_2)$ ,  $\varsigma(u) = 1 + 1 / (1 + e^{-20(u-1/3)})$ , and second with no treatment effect,  $\tau(x) = 0$ , but with 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 estimate heterogeneous treatment effects under unconfoundedness using forests. We compare our generalized random forests with and without local centering (C. GRF and GRF) to Procedures 1 and 2 of [Wager and Athey \(2018\)](#), WA-1 and WA-2. 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 readability.

$e(x) = \frac{1}{4}(1 + \beta_{2,4}(x_3))$ ,  $m(x) = 2x_3 - 1$ , where  $\beta_{a,b}$  is the  $\beta$ -density with shape parameters  $a$  and  $b$ . We also consider a third setting with both heterogeneity and confounding, that combines  $\tau(\cdot)$  from the first setting with  $m(\cdot)$  and  $e(\cdot)$  from the second. For the first setting, [Wager and Athey \(2018\)](#) used their Procedure 1, whereas for the second they used Procedure 2, while noting that it is unfortunate that the practitioner is forced to choose one procedure or the other.

Results presented in Table 1 are reassuring, suggesting that generalized random forests with centering do well under both settings, and can better handle the case with both confounding and treatment heterogeneity than either of the other two procedures. In contrast, Procedure 1 of Wager and Athey does poorly with pure confounding, whereas Procedure 2 of Wager and Athey is good in the pure confounding setting, but does poorly with strong heterogeneity; this is as expected, noting the design of both methods.

**7. Application: Instrumental Variables Regression.** In many applications, we want to measure the causal effect of an intervention on an 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 made above. One 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.

For 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 these effects 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, Venti and Wise, 1996).

**7.1. A Forest for Instrumental Variables Regression.** Classical instrumental variables regression seeks a global understanding of the treatment effect, e.g., on average over the whole U.S. population, does having more children reduce the labor force participation of women? Here, we 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.

We observe  $i = 1, \dots, n$  independent and identically distributed subjects, each of whom has features  $X_i \in \mathcal{X}$ , an outcome  $Y_i \in \mathbb{R}$ , a received treatment  $W_i \in \{0, 1\}$ , and an instrument  $Z_i \in \{0, 1\}$ . We believe that the outcomes  $Y_i$  and received treatment  $W_i$  are related via a structural model  $Y_i = \mu(X_i) + \tau(X_i)W_i + \varepsilon_i$ , where  $\tau(X_i)$  is understood to be the causal effect of  $W_i$  on  $Y_i$ , and  $\varepsilon_i$  is a noise term that may be positively correlated with  $W_i$ . Because  $\varepsilon_i$  is correlated with  $W_i$ , standard regression analyses will not in general be consistent for  $\tau(X_i)$ , and we need to use the instrument  $Z_i$ . If  $Z_i$  is independent of  $\varepsilon_i$  conditionally on  $X_i$  then, provided that  $Z_i$  has an influence on the received treatment  $W_i$ , i.e., that the covariance of  $Z_i$  and  $W_i$  conditionally on  $X_i = x$  is non-zero, the treatment effect  $\tau(x)$  is identified via  $\tau(x) = \text{Cov}[Y_i, Z_i | X_i = x] / \text{Cov}[W_i, Z_i | X_i = x]$ . We can then estimate  $\tau(x)$  by via moment functions  $\mathbb{E}[Z_i(Y_i - W_i\tau(x) - \mu(x)) | X_i = x] = 0$  and  $\mathbb{E}[Y_i - W_i\tau(x) - \mu(x) | X_i = x] = 0$ , where the intercept  $\mu(x)$  is a nuisance parameter. 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 (Abadie, 2003; Imbens and Angrist, 1994).

We then use our formalism to derive a forest that is targeted towards estimating causal effects identified via conditional two-stage least squares. Gradient-based labeling (8) yields pseudo-outcomes for every parent node  $P$  and each observation  $i$  with  $X_i \in P$ ,  $\rho_i = (Z_i - \bar{Z}_P) ((Y_i - \bar{Y}_P) - (W_i - \bar{W}_P) \hat{\tau}_P)$ , where  $\bar{Y}_P$ ,  $\bar{W}_P$ ,  $\bar{Z}_P$  are moments in the parent node, and  $\hat{\tau}_P$  is a solution to the estimating equation in the parent. Given these pseudo-outcomes, the tree executes a CART regression split on the  $\rho_i$  as usual. Finally, we obtain personalized treatment effect estimates  $\hat{\tau}(x)$  by solving (2) with forest weights (3).

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

As in Section 6.1.1, we center our procedure using the transformation of Robinson (1988), and regress out the marginal effects of  $X_i$  first. Writing  $y(x) = \mathbb{E}[Y_i | X = x]$ ,  $w(x) = \mathbb{E}[W_i | X = x]$ , and  $z(x) = \mathbb{E}[Z_i | X = x]$ , we compute conditionally centered outcomes by leave-one-out estimation  $\tilde{Y}_i = Y_i - \hat{y}^{(-i)}(X_i)$ ,  $\tilde{W}_i = W_i - \hat{w}^{(-i)}(X_i)$  and  $\tilde{Z}_i = Z_i - \hat{z}^{(-i)}(X_i)$ , and then run the full instrumental variables forest using centered outcomes  $\{\tilde{Y}_i, \tilde{W}_i, \tilde{Z}_i\}_{i=1}^n$ . 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  $\hat{y}^{(-i)}(X_i)$ ,  $\hat{w}^{(-i)}(X_i)$  and  $\hat{z}^{(-i)}(X_i)$  are first estimated using 3 separate regression forests.

There is a rich literature on non-parametric instrumental variables regression. The above approach generalizes classical approaches based on kernels or series estimation (Abadie, 2003; Su, Murtazashvili and Ullah, 2013; Wooldridge, 2010). In other threads, Darolles et al. (2011) and Newey and Powell (2003) study instrumental variables models that generalize the conditionally linear treatment model and allowing for non-linear effects, and Hartford et al. (2017) develop deep learning tools. Belloni et al. (2012) consider working with high-dimensional instruments  $Z_i$ .

Appendix C has a simulation study for IV forests, comparing them to nearest-neighbor and series regression. We find our method to perform well relative to these baselines, and centering to be helpful. We also evaluate coverage of the bootstrap of little bag confidence intervals.

**7.2. The Effect of Child Rearing on Labor-Force Participation.** We now revisit our motivating example discussed at the beginning of Section 7. We follow Angrist and Evans (1998) in constructing our dataset, and study a sample of  $n = 334,535$  married mothers with at least 2 children (1980 census data), based on the following quantities: The outcome  $Y_i$  is whether the mother did not work in the year preceding the census, the received treatment  $W_i$  is whether the mother had 3 or more chil-

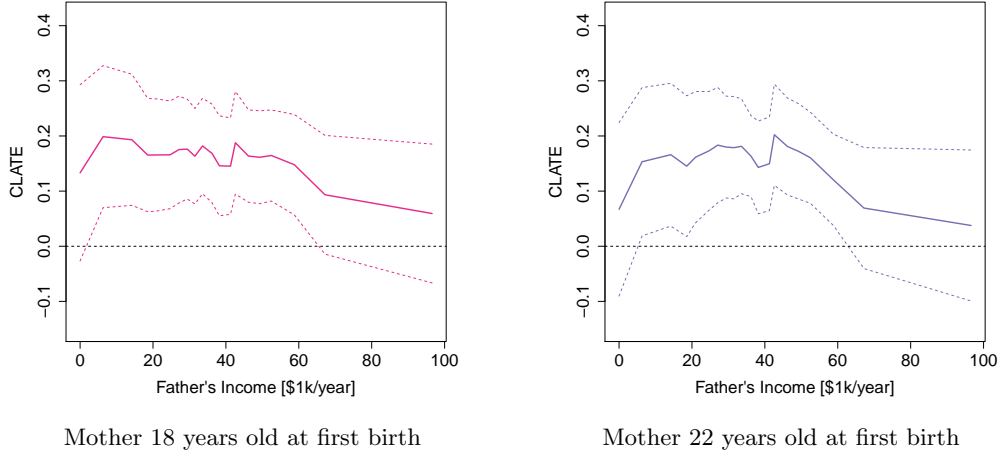


Fig 3: Generalized random forest estimates (along with pointwise 95% confidence intervals) for the causal effect of having a third child on the probability that a mother works for pay, as identified by the same sex instrument of Angrist and Evans (1998); a positive treatment effect means that the treatment *reduces* the probability that the mother works. 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.

dren 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 among mothers with at least two children. In our sample,  $\widehat{\text{Cov}}[W, Z] = 1.6 \cdot 10^{-2}$ , while  $\widehat{\text{Cov}}[Y, Z] = 2.1 \cdot 10^{-3}$ , leading to a 95% confidence interval for the local average treatment effect  $\tau \in (0.14 \pm 0.054)$  using the R function `ivreg` (Kleiberg and Zeileis, 2008). Thus, it appears that having a third child reduces women’s labor force participation on average in the US. Angrist and Evans (1998) conduct extensive sensitivity analysis to corroborate the plausibility of this identification strategy.

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 3. 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 3 are rather wide, attesting to the importance of formal asymptotic theory when using forest-based methods for instrumental variables regression.

**8. Discussion.** We introduced generalized random forests as a versatile method for adaptive, local estimation in a wide variety of statistical models. We discussed our method in the contexts of quantile regression and heterogeneous treatment effect estimation, and our approach also applies 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 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 interest both conceptually and in practice. Moreover, again like standard methods, 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.

**Acknowledgment.** We are grateful 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 in Section 6.1.1, to the associate editor and three anonymous referees for helpful suggestions, and to seminar participants at the Atlantic Causal Inference Conference, the BIRS Workshop on the Interface of Machine Learning and Statistical Inference, 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.

## References.

- ABADIE, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics* **113** 231–263.
- AMIT, Y. and GEMAN, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation* **9** 1545–1588.
- ANDREWS, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica* 821–856.
- ANGRIST, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *AER* 313–336.
- ANGRIST, J. D. and EVANS, W. N. (1998). Children and Their Parents’ Labor Supply: Evidence from Exogenous Variation in Family Size. *AER* 450–477.
- ARLOT, S. and GENUER, R. (2014). Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*.
- ATHEY, S. and IMBENS, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* **113** 7353–7360.
- BELLONI, A., CHEN, D., CHERNOZHUKOV, V. and HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80** 2369–2429.
- BEYGEZIMER, A. and LANGFORD, J. (2009). The offset tree for learning with partial labels. In *Proceedings of KDD* 129–138. ACM.
- BEYGEZIMER, A., KAKADET, S., LANGFORD, J., ARYA, S., MOUNT, D. and LI, S. (2013). FNN: Fast Nearest Neighbor Search Algorithms and Applications.
- BIAU, G. (2012). Analysis of a random forests model. *JMLR* **13** 1063–1095.
- BIAU, G., DEVROYE, L. and LUGOSI, G. (2008). Consistency of random forests and other averaging classifiers. *JMLR* **9** 2015–2033.
- BIAU, G. and DEVROYE, L. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *JMVA* **101** 2499–2518.
- BIAU, G. and SCORNET, E. (2016). A random forest guided tour. *Test* **25** 197–227.
- BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.
- BREIMAN, L. (2001). Random forests. *Machine Learning* **45** 5–32.
- BREIMAN, L., FRIEDMAN, J., STONE, C. J. and OLSHEN, R. A. (1984). *Classification and Regression Trees*. CRC press.
- BÜHLMANN, P. and YU, B. (2002). Analyzing bagging. *Ann. Statist.* **30** 927–961.
- CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C. and NEWEY, W. (2016). Double Machine Learning for Treatment and Causal Parameters. *arXiv preprint arXiv:1608.00060*.
- CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **4** 266–298.
- DAROLLES, S., FAN, Y., FLORENS, J.-P. and RENAULT, E. (2011). Nonparametric instrumental regression. *Econometrica* **79** 1541–1565.
- DENIL, M., MATHESON, D. and DE FREITAS, N. (2014). Narrowing the Gap: Random Forests In Theory and In Practice. In *Proceedings of ICML* 665–673.
- DIETTERICH, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning* **40** 139–157.
- EFRON, B. (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM.
- EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596.
- FAN, J., FARMEN, M. and GIJBELS, I. (1998). Local maximum likelihood estimation and inference. *JRSS-B* **60** 591–608.
- FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications* **66**. CRC Press.
- FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* 1189–1232.
- GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2014). *Bayesian Data Analysis* **2**.

- Chapman & Hall/CRC Boca Raton, FL, USA.
- GEURTS, P., ERNST, D. and WEHENKEL, L. (2006). Extremely randomized trees. *Machine Learning* **63** 3–42.
- GORDON, L. and OLSHEN, R. A. (1985). Tree-structured survival analysis. *Cancer Treatment Reports* **69** 1065–1069.
- HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *JASA* **69** 383–393.
- HANSEN, B. E. (1992). Testing for parameter instability in linear models. *Journal of Policy Modeling* **14** 517–533.
- HARTFORD, J., LEWIS, G., LEYTON-BROWN, K. and TADDY, M. (2017). Deep IV: A Flexible Approach for Counterfactual Prediction. In *Proceedings of ICML* 1414–1423.
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*. New York: Springer.
- HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**.
- HJORT, N. L. and KONING, A. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics* **14** 113–132.
- HO, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence* **20** 832–844.
- HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* **19** 293–325.
- HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *JASA* **58** 13–30.
- HONORÉ, B. E. and KYRIAZIDOU, E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica* **68** 839–874.
- HOTHORN, T., LAUSEN, B., BENNER, A. and RADESPIEL-TRÖGER, M. (2004). Bagging survival trees. *Statistics in Medicine* **23** 77–91.
- IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475.
- ISHWARAN, H. and KOGALUR, U. B. (2010). Consistency of random survival forests. *Statistics & Probability Letters* **80** 1056–1064.
- KALLUS, N. (2016). Learning to Personalize from Observational Data. *arXiv preprint arXiv:1608.08925*.
- KLEIBER, C. and ZEILEIS, A. (2008). *Applied econometrics with R*. Springer Science & Business Media.
- LEBLANC, M. and CROWLEY, J. (1992). Relative risk trees for censored survival data. *Biometrics* **48** 411–425.
- LEWBEL, A. (2007). A local generalized method of moments estimator. *Economics Letters* **94** 124–128.
- LIN, Y. and JEON, Y. (2006). Random forests and adaptive nearest neighbors. *JASA* **101** 578–590.
- LOADER, C. (1999). *Local Regression and Likelihood*. Springer.
- MALLOWS, C. L. (1973). Some comments on Cp. *Technometrics* **15** 661–675.
- MEINSHAUSEN, N. (2006). Quantile regression forests. *JMLR* **7** 983–999.
- MENTCH, L. and HOOKER, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *JMLR* **17** 1–41.
- MOLINARO, A. M., DUDOIT, S. and VAN DER LAAN, M. J. (2004). Tree-based multivariate regression and density estimation with right-censored data. *JMVA* **90** 154–177.
- NEWHEY, W. K. (1994a). Kernel estimation of partial means and a general variance estimator. *Econometric Theory* **10** 1–21.
- NEWHEY, W. K. (1994b). The asymptotic variance of semiparametric estimators. *Econometrica* **62** 1349–1382.
- NEWHEY, W. K. and POWELL, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71** 1565–1578.



- NEYMAN, J. (1979).  $C(\alpha)$  tests and their use. *Sankhyā, Series A* 1–21.
- NYBLÖM, J. (1989). Testing for the constancy of parameters over time. *JASA* **84** 223–230.
- PLOBERGER, W. and KRÄMER, W. (1992). The CUSUM test with OLS residuals. *Econometrica* 271–285.
- POTERBA, J. M., VENTI, S. F. and WISE, D. A. (1996). How retirement saving programs increase saving. *JEP* **10** 91–112.
- ROBINS, J. M. and RITOV, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16**.
- ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* 931–954.
- ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.
- SCHICK, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* 1139–1151.
- SCORNET, E., BIAU, G. and VERT, J.-P. (2015). Consistency of random forests. *Ann. Statist.* **43** 1716–1741.
- SEXTON, J. and LAAKE, P. (2009). Standard errors for bagged and random forest estimators. *Computational Statistics & Data Analysis* **53** 801–811.
- STANISWALIS, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *JASA* **84** 276–283.
- STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* 595–620.
- SU, L., MURTAZASHVILI, I. and ULLAH, A. (2013). 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** 184–207.
- SU, X., TSAI, C.-L., WANG, H., NICKERSON, D. M. and LI, B. (2009). Subgroup analysis via recursive partitioning. *JMLR* **10** 141–158.
- TIBSHIRANI, R. and HASTIE, T. (1987). Local likelihood estimation. *JASA* **82** 559–567.
- VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.
- VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.
- VARIAN, H. R. (2014). Big data: New tricks for econometrics. *JEP* **28** 3–27.
- WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *JASA* **just-accepted**.
- WAGER, S., HASTIE, T. and EFRON, B. (2014). Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife. *JMLR* **15**.
- WAGER, S. and WALTHER, G. (2015). Adaptive Concentration of Regression Trees, with Application to Random Forests. *arXiv preprint arXiv:1503.06388*.
- WOOLDRIDGE, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.
- WRIGHT, M. N. and ZIEGLER, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* **77** 1–17.
- ZEILEIS, A. (2005). A unified approach to structural change tests based on ML scores, F statistics, and OLS residuals. *Econometric Reviews* **24** 445–466.
- ZEILEIS, A. and HORNIK, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica* **61** 488–508.
- ZEILEIS, A., HOTHORN, T. and HORNIK, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics* **17** 492–514.
- ZHU, R., ZENG, D. and KOSOROK, M. R. (2015). Reinforcement learning trees. *JASA* **110** 1770–1784.

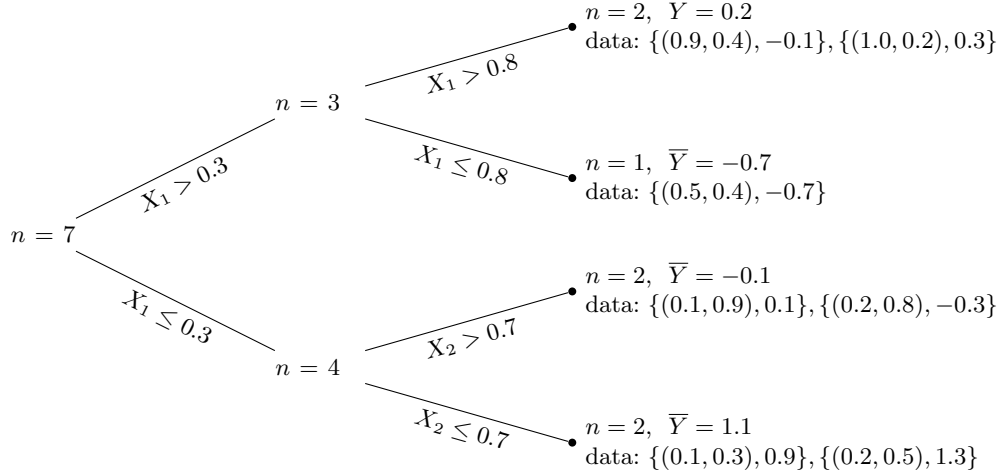


Fig 4: Example of a small regression tree on a sample of size  $n = 7$ . The examples used to build this tree are of the form  $\{X_i, Y_i\} \in \mathbb{R}^2 \times \mathbb{R}$ , and axis-aligned splits based on the  $X_i$  determine the leaf membership of each training example. In “standard” regression trees as discussed in, e.g., [Breiman et al. \(1984\)](#) or [Hastie, Tibshirani and Friedman \(2009\)](#), 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  $\hat{\mu}(x) = 1.1$ . In our method, we do not consider tree predictions directly, but instead use trees to construct a neighborhood weighting as in Figure 1. Our approach also relies on a form of subsample splitting where different subsets of the data are used to grow the tree and make within-leaf predictions; see Section 2.4 for details.

## APPENDIX A: 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. The proofs of Propositions 1 and 2, as well as the technical results stated below are given in Appendix B. Throughout our theoretical analysis, we use the following notation: Given our forest weights  $\alpha_i(x)$  (3), let

$$(22) \quad \Psi(\theta, \nu) := \sum_{i=1}^n \alpha_i(x) \psi_{\theta, \nu}(O_i) \text{ and } \bar{\Psi}(\theta, \nu) := \sum_{i=1}^n \alpha_i(x) M_{\theta, \nu}(X_i).$$

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

**LEMMA 7.** *Let  $\alpha_i(x)$  be weights from a forest obtained as in Specification 1, 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:

$$(23) \quad \|\mathbb{E}[\Psi(\theta(x), \nu(x))]\|_2 = \mathcal{O}\left(s^{-\frac{\pi}{2} \frac{\log((1-\omega)^{-1})}{\log(\omega^{-1})}}\right)$$

$$(24) \quad \|\text{Var}[\Psi(\theta(x), \nu(x))]\|_F = \mathcal{O}(s/n),$$

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

PROOF. To establish these bounds, we start by expanding  $\Psi$  as

$$(25) \quad \Psi(\theta, \nu) = \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^n \alpha_{bi}(x) \psi_{\theta, \nu}(O_i),$$

where the  $\alpha_{bi}$  are the individual tree weights used to build the forest weights in (3). 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 \(2018\)](#). First, the proof of Theorem 3 of [Wager and Athey \(2018\)](#) shows that the weights  $\alpha_i(x)$  are localized:

$$(26) \quad \mathbb{E}[\sup\{\|X_i - x\|_2 : \alpha_i(x) > 0\}] = \mathcal{O}\left(s^{-\frac{\pi}{2} \frac{\log((1-\omega)^{-1})}{\log(\omega^{-1})}}\right),$$

thus directly implying (23) thanks to Assumption 1. Meanwhile, because individual trees are grown on subsamples, we can verify that

$$(27) \quad \frac{n}{s} \text{Var}[\Psi(\theta(x), \nu(x))] \preceq \text{Var}\left[\sum_{i=1}^n \alpha_{bi}(x) \psi_{\theta, \nu}(O_i)\right] = \mathcal{O}(1).$$

The first inequality results from classical results about  $U$ -statistics going back to [Hoeffding \(1948\)](#), and simply states that the variance of the forest score is at most  $s/n$  times the variance of a tree score; see Appendix C3 of [Wager and Athey \(2018\)](#) for a discussion in the context of regression forests. The second inequality follows from second-moment bounds on  $\psi$  along with the fact that our trees are grown on honest subsamples.  $\square$

**A.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  $\Psi$ . To do so, define the local discrepancy measure

$$(28) \quad \delta((\theta, \nu), (\theta', \nu')) = \Psi(\theta, \nu) - \bar{\Psi}(\theta, \nu) - (\Psi(\theta', \nu') - \bar{\Psi}(\theta', \nu')),$$

which describes how tightly the stochastic fluctuations of  $\Psi - \bar{\Psi}$  are coupled for nearby parameter values  $(\theta, \nu)$  and  $(\theta', \nu')$ . The following lemmas establish uniform local concentration of  $\delta$ : First, in Lemma 8, we control the variogram of the forest, and then Lemma 9 establishes concentration of  $\delta$  over small balls. Both proofs are given in Appendix B.

LEMMA 8. *Let  $(\theta, \nu)$  and  $(\theta', \nu')$  be fixed pairs of parameters, and let  $\alpha_i(x)$  be forest weights generated according to Specification 1. Then, provided that Assumptions 1–3 hold,*

$$(29) \quad \begin{aligned} \mathbb{E} [\delta((\theta, \nu), (\theta', \nu'))] &= 0, \\ \mathbb{E} [\|\delta((\theta, \nu), (\theta', \nu'))\|_2^2] &\leq L \frac{s}{n} \left\| \begin{pmatrix} \theta \\ \nu \end{pmatrix} - \begin{pmatrix} \theta' \\ \nu' \end{pmatrix} \right\|_2, \end{aligned}$$

where  $L$  is the Lipschitz parameter from (11).

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

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

for all realizations of  $\Psi$ , 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 “ $\psi$ -trees” that comprise  $\Psi$ :

$$(30) \quad \begin{aligned} r^2 \left( \beta \left( \begin{pmatrix} \theta_- \\ \nu_- \end{pmatrix}, \begin{pmatrix} \theta_+ \\ \nu_+ \end{pmatrix} \right) \right) \\ := \mathbb{E} \left[ \left\| \sum_{\{i:i \in \mathcal{J}_1\}} \alpha_i(x; \mathcal{J}_2) (\psi_{\theta_+, \nu_+}(O_i) - \psi_{\theta_-, \nu_-}(O_i)) \right\|_2^2 \right], \end{aligned}$$

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

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

$$(31) \quad \log(N_{[]}(\varepsilon, \Psi, L_2)) \leq \frac{\kappa}{\varepsilon} \quad \text{for all } 0 < \varepsilon \leq 1.$$

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 (31) would be enough to prove our results, which will only depend on this assumption through Lemma 9 below.

To see how Assumption 4 leads to (31), we first write

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

where  $\Psi^\lambda$  is Lipschitz and  $\Psi^\zeta$  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

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

Because  $\Psi^\zeta$  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  $\Psi^\lambda$  is Lipschitz and our parameter space  $\mathcal{B}$  is compact, Lemma 2.7.11 of [van der Vaart and Wellner \(1996\)](#) implies that  $\log N_{[]}(\varepsilon, \Psi^\lambda, L_2) = \mathcal{O}(\log \varepsilon^{-1})$ . Thus, both terms are controlled at the desired order, and so (31) holds.

**LEMMA 9.** *Under the conditions of Lemma 8, suppose moreover that (31) holds. Then,*

$$(32) \quad \mathbb{E} \left[ \sup_{(\theta', \nu')} \left\{ \|\delta((\theta, \nu), (\theta', \nu'))\|_2 : \left\| \begin{pmatrix} \theta - \theta' \\ \nu - \nu' \end{pmatrix} \right\|_2 \leq \eta \right\} \right] \\ = \mathcal{O} \left( \sqrt{\frac{\kappa L \eta}{n/s}} + \frac{8\kappa G}{(n/s)L\eta} \right),$$

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

**Proof of Theorem 3.** First, thanks to Lemma 7, we know that

$$(33) \quad \|\Psi(\theta(x), \nu(x))\|_2 \rightarrow_p 0.$$

Thus, thanks to Assumption 5, we know there must exist a sequence  $\varepsilon_n > 0$  with  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$  such that

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

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

LEMMA 10. *Suppose that Assumptions 1–6 hold, and that the forest is trained according to Specification 1. Then, all approximate solutions to (2) are close to each other, in the following sense: for any sequence  $\varepsilon_n > 0$  with  $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ ,*

$$(34) \quad \sup \left\{ \left\| \begin{pmatrix} \theta - \theta' \\ \nu - \nu' \end{pmatrix} \right\|_2 : \|\Psi(\theta, \nu)\|_2, \|\Psi(\theta', \nu')\|_2 < \varepsilon_n \right\} \rightarrow_p 0.$$

PROOF. Starting with some notation, let

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

where  $F$  and  $\bar{F}$  are the respectively convex and  $\sigma^2$ -strongly convex functions implicitly defined in the hypothesis statement. Recall that  $(\hat{\theta}, \hat{\nu})$  is assumed to satisfy Assumption 5, and let  $\eta_n > 0$  be any sequence with  $\lim_{n \rightarrow \infty} \eta_n = 0$ ,  $\eta_n > \max\{4\varepsilon_n/\sigma^2, \sqrt[4]{s/n}\}$  for all  $n = 1, 2, \dots$ , and  $\eta_n^{-1} \|\Psi(\hat{\theta}, \hat{\nu})\|_2 \rightarrow_p 0$ .

Now, thanks to Assumptions 1–4, we can apply Lemma 9. Because  $\eta_n \geq \sqrt[4]{s/n}$ , we can pair (32) with the fundamental theorem of calculus for line integrals to check that

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

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

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

again for  $(\theta, \nu)$  within  $\eta_n$  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  $\eta_n^{-1} \|\Psi(\hat{\theta}, \hat{\nu})\|_2 \rightarrow_p 0$ . 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 (34) must hold.  $\square$

**Proof of Lemma 4.** If  $\psi_{\theta, \nu}(O_i)$  were twice differentiable in  $(\theta, \nu)$ , then we could verify (14) fairly directly via Taylor expansion of  $\psi$ . Now, of course,  $\psi$  is not twice differentiable, and so we cannot apply this argument directly. Rather, we need to first apply a Taylor expansion on the expected  $\psi$  function,  $M_{\theta, \nu}(X_i)$ , which *is* twice differentiable; we then use the regularity properties established in Section A.1 to extend this result to  $\psi$ .

Given consistency of  $(\hat{\theta}(x), \hat{\nu}(x))$ , there is a sequence  $\varepsilon_n \rightarrow 0$  such that

$$(35) \quad \left\| \begin{pmatrix} \hat{\theta}(x) - \theta(x) \\ \hat{\nu}(x) - \nu(x) \end{pmatrix} \right\|_2 = \mathcal{O}_P(\varepsilon_n).$$

Using notation established in (22) and (28), we then write

$$(36) \quad \begin{aligned} \bar{\Psi}(\hat{\theta}(x), \hat{\nu}(x)) - \bar{\Psi}(\theta(x), \nu(x)) &= \Psi(\hat{\theta}(x), \hat{\nu}(x)) - \Psi(\theta(x), \nu(x)) \\ &\quad + \delta((\theta(x), \nu(x)), (\hat{\theta}(x), \hat{\nu}(x))). \end{aligned}$$

By the assumed smoothness of the moment functions, we know that  $\bar{\Psi}$  is twice differentiable in  $(\theta, \nu)$  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:

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

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 (26),

$$(37) \quad \left\| \sum_{i=1}^n \alpha_i(x) \nabla M_{\theta(x), \nu(x)}(X_i) - V(x) \right\|_F = \mathcal{O}_P \left( s^{-\frac{\pi}{2} \frac{\log((1-\omega)^{-1})}{\log(\omega^{-1})}} \right),$$

where  $s \rightarrow \infty$  is the sub-sample size used to grow trees in the forest. This expansion suggests that (36) should be helpful in relating our quantities of interest.

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

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

Next, by the consistency of  $(\hat{\theta}(x), \hat{\nu}(x))$ , we can apply Lemma 9 with “ $\eta$ ” set to  $\varepsilon_n^{2/3}$  to verify that

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



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

$$(38) \quad \left\| \begin{pmatrix} \hat{\theta}(x) - \theta(x) \\ \hat{\nu}(x) - \nu(x) \end{pmatrix} + V(x)^{-1} \Psi(\theta(x), \nu(x)) \right\|_2 \\ = \mathcal{O}_P \left( \max \left\{ s^{-\frac{\pi}{2} \frac{\log((1-\omega)^{-1})}{\log(\omega^{-1})}} \varepsilon_n, \varepsilon_n^2, \varepsilon_n^{1/3} \sqrt{\frac{s}{n}}, \frac{s}{n \varepsilon_n^{2/3}} \right\} \right).$$

Finally, recall that  $\|\Psi(\theta(x), \nu(x))\|_2^2 = \mathcal{O}_P(s/n)$  by Lemma 7 and (13). Thus, we can use the bound (38) 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 (35) to hold with  $\varepsilon_n = \sqrt{s/n}$ . The desired result then follows directly from (38), noting that  $\tilde{\theta}^*(x) = \theta(x) + \xi^\top V(x)^{-1} \Psi(\theta(x), \nu(x))$ .

**Proof of Theorem 5.** As argued in Section 3.1,  $\tilde{\theta}^*(x)$  is formally equivalent to the output of a regression forest, and so we can directly apply Theorem 1 of Wager and Athey (2018). Given the assumptions made here, their result shows that

$$(39) \quad \left( \tilde{\theta}^*(x) - \theta(x) \right) / \sigma_n(x) \Rightarrow \mathcal{N}(0, 1), \quad \sigma_n^2(x) \rightarrow_p 0.$$

Moreover, from Theorem 5 and Lemma 7 of Wager and Athey (2018), we see that  $\sigma_n^2$  scales as discussed in the hypothesis statement. Given this central limit theorem, it only remains to show that the discrepancy between  $\hat{\theta}(x)$  and  $\tilde{\theta}^*(x)$  established in Lemma 4, decays faster than  $\sigma_n(x)$ . But, thanks to the consistency result from Theorem 3, the coupling result in Lemma 4 implies that

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

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

**Proof of Theorem 6.** Following our discussion in Section 4.1, we here only consider the ideal “ $B \rightarrow \infty$ ” half-sampling estimator. We start by considering its expectation,

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

for  $\mathcal{H} = \{1, \dots, \lfloor n/2 \rfloor\}$ . 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 9 with  $\eta =$

$(s/n)^{1/3}$  to verify that

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

where  $\bar{\Psi}_{\mathcal{H}}$  is defined analogously to  $\bar{\Psi}$  in (22).

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 \(2018\)](#) to directly verify that (see their Lemma 7 and Theorem 8)

$$\begin{aligned} (1 + o_P(1)) (\Psi(\theta(x), \nu(x)) - \mathbb{E}[\Psi(\theta(x), \nu(x))]) \\ = \sum_{i=1}^n (\mathbb{E}[\Psi(\theta(x), \nu(x)) \mid (X_i, O_i)] - \mathbb{E}[\Psi(\theta(x), \nu(x))]) , \\ (40) \quad (1 + o_P(1)) (\Psi_{\mathcal{H}}(\theta(x), \nu(x)) - \mathbb{E}[\Psi(\theta(x), \nu(x))]) \\ = \frac{n}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} (\mathbb{E}[\Psi(\theta(x), \nu(x)) \mid (X_i, O_i)] - \mathbb{E}[\Psi(\theta(x), \nu(x))]) . \end{aligned}$$

This holds because, as discussed in [Wager and Athey \(2018\)](#), forests have regularity properties by which the scaled first-order effects  $n(\mathbb{E}[\Psi(\theta(x), \nu(x)) \mid (X_i, O_i)] - \mathbb{E}[\Psi(\theta(x), \nu(x))])$  depend only on the type of tree being grown; and here of course  $\Psi$  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

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

where the latter is again immediate by the proof of Theorem 8 in [Wager and Athey \(2018\)](#). 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  $\bar{\Psi}$  is twice differentiable with a uniform second derivative, so we can take a Taylor expansion as in the proof of Lemma 4:

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

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 (37), we see that  $\|\nabla \Psi_{\mathcal{H}}(\theta(x), \nu(x)) - \nabla \Psi(\theta(x), \nu(x))\| \rightarrow_P 0$ , 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

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

implying that

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

To establish consistency, it remains to verify concentration of  $\hat{H}_n^{HS}(x)$ ; which, given that the contribution of  $R_{\mathcal{H}}$  is negligible, also follows immediately from (40). Finally, given consistency of  $\hat{H}_n^{HS}(x)$  and Theorem 5, the validity of the delta method confidence intervals is immediate by Slutsky's theorem whenever  $\|\hat{V}(x) - V(x)\| \rightarrow_P 0$ ; in particular, recall that  $V(x)$  is invertible by Assumption 2.

## APPENDIX B: TECHNICAL RESULTS

The proofs presented here depend on arguments and notation established in Appendix A. From now on, we also use shorthand

$$(41) \quad \mathcal{O}(a, b, c) = \mathcal{O}(\max\{a, b, c\}),$$

etc. 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  $\rho_i^*(x_P)$  averaged over  $C_j$ , namely

$$(42) \quad \tilde{\theta}_{C_j}^*(x_P) = \theta(x_P) + \frac{1}{|C_j|} \sum_{i \in C_j} \rho_i^*(x_P),$$

where  $x_P$  is the center of mass of the parent node  $P$ .

Because the leaf  $C_j$  is considered fixed, we can use second-moment bounds on  $\psi$  to verify that  $\text{Var}[\tilde{\theta}_{C_j}^*(x_P)] = \mathcal{O}(1/n_{C_j})$ ; meanwhile, by Lipschitz-continuity of the  $M$ -function (10), we see that  $\mathbb{E}[\tilde{\theta}_{C_j}^*(x_P) - \theta(x_P)] = \mathcal{O}(r)$ , 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  $\theta(x_P)$  in a limit where  $r \rightarrow 0$  and  $n_{C_j} \rightarrow \infty$ . Thus, a direct analogue to our result, Lemma 4, implies that

$$(43) \quad \tilde{\theta}_{C_j}^*(x_P) - \hat{\theta}_{C_j} = o_P(r, 1/\sqrt{n_{C_j}}).$$

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

$$\begin{aligned}
 \tilde{\theta}_{C_j} - \tilde{\theta}_{C_j}^*(x_P) &= \hat{\theta}_P - \theta(x_P) \\
 &\quad - \frac{1}{n_{C_j}} \xi^\top V(x_P)^{-1} \sum_{i \in C_j} \left( \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) - \psi_{\theta(x_P), \nu(x_P)}(O_i) \right) \\
 &\quad - \frac{1}{n_{C_j}} \xi^\top (A_P^{-1} - V(x_P)^{-1}) \sum_{i \in C_j} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i);
 \end{aligned}
 \tag{44}$$

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  $\theta$  and  $\nu$ , as well as an analogue to Lemma 9; while the second line term can be bounded by recalling that  $\|A_P^{-1} - V(x_P)^{-1}\| = o_P(1)$ , and verifying that  $\sum_{i \in C_j} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) = \mathcal{O}_P(1/\sqrt{n_{C_j}}, r)$ . Everything we have showed so far implies that

$$\tilde{\theta}_{C_j} - \hat{\theta}_{C_j} = o_P(r, 1/\sqrt{n_{C_j}}), \text{ for } j = 1, 2.
 \tag{45}$$

Finally, it is straight-forward to check that

$$\tilde{\theta}_{C_2} - \tilde{\theta}_{C_1} = O_P(r, 1/\sqrt{n_{C_1}}, 1/\sqrt{n_{C_2}}),
 \tag{46}$$

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

**Proof of Proposition 1.** First, we show that we can replace  $\hat{\theta}_{C_j}(\mathcal{J})$  with the influence-based approximation  $\tilde{\theta}_{C_j}^*(x_P; \mathcal{J})$  (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  $x_P \in P$  be the center of mass of the parent leaf, and observe that

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

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

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

and so this term is just the average error from Taylor expanding  $M_{\theta(x_P), \nu(x_P)}(\cdot)$  over  $C_j$ . Now, using the above expansion, we find that

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

Following arguments of [Athey and Imbens \(2016\)](#), we see that

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

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

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

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

$$\begin{aligned} \mathbb{E} [\Delta (C_1, C_2)] &= \frac{n_{C_1} n_{C_2}}{n_P^2} \mathbb{E} \left[ \left( \tilde{\theta}_{C_2}^*(x_P; \mathcal{J}) - \tilde{\theta}_{C_1}^*(x_P; \mathcal{J}) \right)^2 \right] \\ &\quad + o \left( r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}} \right). \end{aligned}$$

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

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

that captures the effect of overfitting to random noise when estimating  $\tilde{\theta}_{C_j}^*(x_P)$ . 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.

**Proof of Lemma 8.** We first note that, because we grew our forest honestly (Specification 1) and so  $\alpha_i$  is independent of  $O_i$  conditionally on  $X_i$ , we can use the chain rule to verify that

$$\begin{aligned} &\mathbb{E} [\Psi(\theta, \nu) - \bar{\Psi}(\theta, \nu)] \\ &= \sum_{i=1}^n \mathbb{E} [\mathbb{E} [\alpha_i(x) \mid X_i] (\mathbb{E} [\psi_{\theta, \nu}(O_i) \mid X_i] - M_{\theta, \nu}(X_i))] = 0, \end{aligned}$$

and so  $\delta$  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  $\mathbb{E} [\mathcal{E}_{\theta, \nu}(O_i, X_i) \mid M_{\theta, \nu}(X_i)] = 0$  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 (11). 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  $\mathbb{E}[T_{\theta, \nu}(\mathcal{J}_1, \mathcal{J}_2)] = 0$ ; meanwhile, thanks to the variogram bound, for any pair of points  $(\theta, \nu)$  and  $(\theta', \nu')$ ,

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

As in arguments used by [Wager and Athey \(2018\)](#), we see that our quantity of interest  $U$ -statistic over  $T$ , and in particular

$$\begin{aligned}
 & \delta((\theta, \nu), (\theta', \nu')) \\
 & = \left( \binom{n}{\lfloor s/2 \rfloor, \lceil s/2 \rceil} \right)^{-1} \sum_{\{S_1, S_2 \in \{1, \dots, n\}\}} T_{\theta, \nu}(\mathcal{J}_1, \mathcal{J}_2) - T_{\theta', \nu'}(\mathcal{J}_1, \mathcal{J}_2).
 \end{aligned}$$

Thus, combining our above variogram bound for  $T$  with results on  $U$ -statistics going back to [Hoeffding \(1948\)](#), we see that (29) holds.

**Proof of Lemma 9.** We start by establishing a concentration bound for  $\delta$  at a single point. Given Assumption 4, we know that  $\|\delta\|_\infty$  is bounded by  $2G$ , where  $G$  is as defined in the problem statement. Thus, recalling that  $\delta$  is a  $U$ -statistic and using (47) 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  $\eta > 0$ ,

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

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

Now, the kernel of  $\delta$ , 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 (31), using the radius (30). Thus, the kernel of  $\delta$  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 (48), 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

$$\begin{aligned}
 & \mathbb{E} \left[ \sup \{ \delta((\theta, \nu), (\theta', \nu')) : ((\theta, \nu), (\theta', \nu')) \in \mathcal{S} \} \right] \\
 & = \mathcal{O} \left( \frac{J_{[]} (r, \delta, L_2)}{\sqrt{n/s}} + \frac{J_{[]}^2 (r, \delta, L_2)}{r^2 n/s} 2G \right),
 \end{aligned}$$



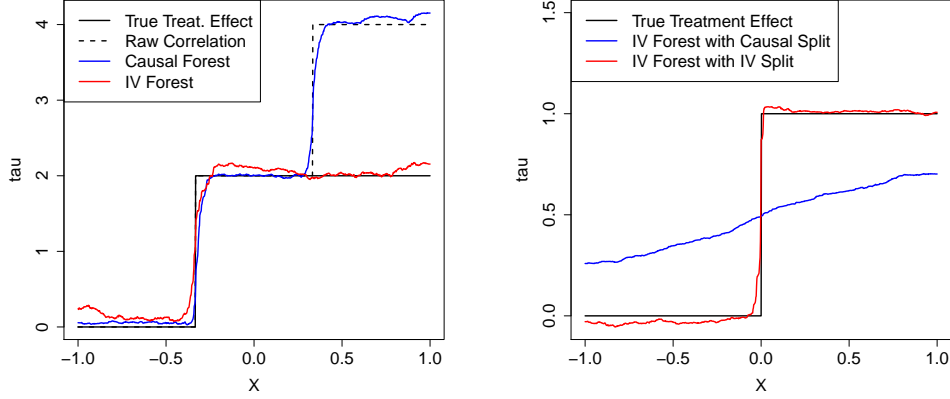


Fig 5: In both panels, we generate data as  $X_i \sim [-1, 1]^p$ , with  $n = 10,000$  and  $p = 20$ .

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

$$J_{[]} (r, \delta, L_2) := \int_0^r \sqrt{1 + \log (N_{[]} (\eta, \delta, L_2))} d\eta,$$

and we omitted the  $\lfloor \cdot \rfloor$  notation since  $(n/s)/\lfloor n/s \rfloor \leq 2$ . From our bounds on the bracketing number we get  $J_{[]} (r, \delta, L_2) \leq 4\sqrt{\kappa r} + o(\sqrt{r})$ . Thus, thanks to Lemma 8, we conclude by applying the above result with  $r = L\eta$ .

## APPENDIX C: SIMULATING INSTRUMENTAL VARIABLES FORESTS

**C.1. Evaluating the Instrumental Variables Splitting Rule.** We start our simulation analysis with a simple diagnostic of IV splitting rules, and illustrate the behavior of IV forests in Figure 5 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 5, 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 received treatment may be endogenous. We consider a case where the true causal effect of has a single jump,  $\tau(X_i) = 2 \times \mathbf{1}(\{(X_i)_1 > -1/3\})$ . Meanwhile, at  $(X_i)_1 = +1/3$ , there is a change in the correlation structure between  $W_i$  and  $\varepsilon_i$  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 received treatment  $W_i$  is exogenous will mistakenly also pick out the second spurious jump in the correlation structure of  $W_i$  and  $Y_i$ .

The second panel tests 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  $\varepsilon_i$ , such that the joint distribution of  $W_i$  and  $Y_i$  does not depend on  $X_i$ . The IV forest described above again performs well; however, the simpler causal tree splitting from Section 6 that was not designed for IV regression fails to accurately detect the jump.

**C.2. Numerical Comparisons.** We now examine the value of adaptivity in local instrumental variables regression using generalized random forests across several simulation designs. We compare the following four methods: **nearest neighbors** instrumental variables regression, which sets  $\alpha_i(x) = 1/k$  in (2) 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 via marginal regressions as in [Robinson \(1988\)](#).

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. 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` ([Kleiberg 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:<sup>1</sup>

$$(49) \quad \begin{aligned} X_i &\sim \mathcal{N}(0, I_{p \times p}), \quad \varepsilon_i \sim \mathcal{N}(0, 1), \quad Z_i \sim \text{Binom}(1/3), \\ Q_i &\sim \text{Binom}(1/(1 + e^{-\omega \varepsilon_i})), \quad W_i = Z_i \wedge Q_i, \\ Y_i &= \mu(X_i) + (W_i - 1/2)\tau(X_i) + \varepsilon_i. \end{aligned}$$

In other words, we exogenously draw features  $X_i$ , a noise term  $\varepsilon_i$  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  $\varepsilon_i$  when  $\omega > 0$ . We varied the following problem parameters. **Confounding:** We toggled the confounding parameter  $\omega$  in (49) between  $\omega = 0$  (no confounding) and  $\omega = 1$  (confounding). **Sparsity of signal:** The signal  $\tau(x)$  depended on  $\kappa_\tau$  features; we used  $\kappa_\tau \in \{2, 4\}$ .

<sup>1</sup>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  $\varepsilon_i$  are more likely to comply, and we need to use the instrument  $Z_i$  to deal with this non-compliance effect.

**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**  $n$ .<sup>2</sup>

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 received treatment. 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 performance.

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.<sup>3</sup>

**C.3. Evaluating Confidence Intervals.** We also examine the quality of the delta method confidence intervals discussed in Section 4, 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, Hastie and Efron (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. Figure 6 gives an illustration of our confidence intervals by superimposing the output from 4 different simulation runs from a single data-generating distribution.

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

<sup>2</sup>Note that these simulation setups do not perfectly match the assumptions made in Section 3, because if we map the features into the unit cube via a monotone transformation, then the signals are no longer Lipschitz. Reassuringly, this does not appear to hurt performance of our method.

<sup>3</sup>In this simulation design, sparse variants of the series regression based on the lasso 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.

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

TABLE 2

Results from simulation study described in Appendix C.2, 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 (49) is 0, whether all signals are additive (add.), whether the received treatment  $W$  is confounded (conf.), the signal dimension ( $\kappa_\tau$ ), 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.

$\kappa_\tau$	additive		yes coverage	no coverage
	$p$	$n$		
2	6	2000	0.87	0.82
2	6	4000	0.91	0.87
2	6	8000	0.91	0.89
2	6	16000	0.93	0.92
2	12	2000	0.75	0.77
2	12	4000	0.82	0.77
2	12	8000	0.89	0.86
2	12	16000	0.93	0.90
2	18	2000	0.73	0.73
2	18	4000	0.82	0.79
2	18	8000	0.89	0.80
2	18	16000	0.91	0.87
4	6	2000	0.78	0.75
4	6	4000	0.82	0.76
4	6	8000	0.87	0.77
4	6	16000	0.88	0.81
4	12	2000	0.64	0.59
4	12	4000	0.73	0.63
4	12	8000	0.78	0.66
4	12	16000	0.82	0.70
4	18	2000	0.59	0.53
4	18	4000	0.63	0.54
4	18	8000	0.70	0.60
4	18	16000	0.78	0.60

$\kappa_\tau$	additive		yes coverage	no coverage
	$p$	$n$		
2	6	2000	0.96	0.95
2	6	4000	0.96	0.96
2	6	8000	0.94	0.95
2	6	16000	0.95	0.95
2	12	2000	0.96	0.97
2	12	4000	0.97	0.95
2	12	8000	0.96	0.97
2	12	16000	0.97	0.98
2	18	2000	0.98	0.96
2	18	4000	0.97	0.97
2	18	8000	0.97	0.96
2	18	16000	0.97	0.98
4	6	2000	0.96	0.96
4	6	4000	0.95	0.94
4	6	8000	0.95	0.95
4	6	16000	0.95	0.95
4	12	2000	0.96	0.95
4	12	4000	0.97	0.96
4	12	8000	0.97	0.97
4	12	16000	0.96	0.97
4	18	2000	0.96	0.97
4	18	4000	0.97	0.97
4	18	8000	0.96	0.96
4	18	16000	0.96	0.97

target: population  $\tau(x)$ target: expected  $\hat{\tau}(x)$ 

TABLE 3

*Empirical coverage of 95% confidence intervals for instrumental variables forests, averaged over 20 replications with 1,000 test points each. The left panel reports coverage of the true effects  $\tau(X_i)$  on the test set, while the right panel measures the fraction of times the expected forest prediction*

$$\mathbb{E} [\hat{\tau}(X_i) | X_i] \text{ falls within the confidence intervals.}$$

effects, the most important one in Table 3 is the sparsity of  $\tau$ . When  $\kappa_\tau = 2$ , i.e., the true signal can be expressed as a bivariate function, our confidence intervals achieve closer to nominal coverage; however, when  $\kappa_\tau = 4$ , performance declines considerably at the sample sizes  $n$  under investigation.

To gain more intuition about this result, the right panel of Table 3 reports the fraction of confidence intervals that cover the expected prediction made by the forest; in other words, it measures the accuracy with which our confidence intervals quantify the sampling uncertainty of the forest. If instrumental forests were unbiased, the left and right panels would be the same. These results suggest that low coverage numbers in the left panel are mostly due to our forests having non-negligible bias, rather than to failures of Gaussianity or of the variance estimates underlying our confidence intervals. It would be of considerable interest to develop confidence intervals for random forests that allow for asymptotically non-vanishing bias.

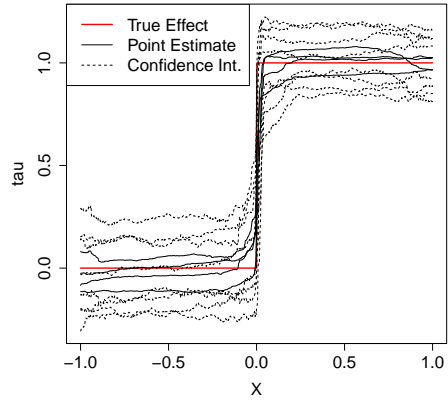


Fig 6: Illustration of 95% confidence intervals for instrumental variables forests across 4 simulation replications. We use the same simulation setting as in the right panel of Figure 5, except now with  $n = 4,000$ ,  $p = 20$ , and  $B = 10,000$  trees.

STANFORD GRADUATE SCHOOL OF BUSINESS  
655 KNIGHT WAY  
STANFORD, CA-94305, USA  
E-MAIL: [athey@stanford.edu](mailto:athey@stanford.edu)

ELASTICSEARCH BV  
800 WEST EL CAMINO REAL, SUITE 350  
MOUNTAIN VIEW, CA-94040  
E-MAIL: [julietibs@gmail.com](mailto:julietibs@gmail.com)

STANFORD GRADUATE SCHOOL OF BUSINESS  
655 KNIGHT WAY  
STANFORD, CA-94305, USA  
E-MAIL: [swager@stanford.edu](mailto:swager@stanford.edu)