

---

# Generalized Random Forests using Fixed-Point Trees

---

David Fleischer <sup>\*1</sup> David A. Stephens <sup>\*1</sup> Archer Y. Yang <sup>\*1 2</sup>

## Abstract

We propose a computationally efficient alternative to generalized random forests (GRFs) for estimating heterogeneous effects in large dimensions. While GRFs rely on a gradient-based splitting criterion, which in large dimensions is computationally expensive and unstable, our method introduces a fixed-point approximation that eliminates the need for Jacobian estimation. This gradient-free approach preserves GRF's theoretical guarantees of consistency and asymptotic normality while significantly improving computational efficiency. We demonstrate that our method achieves a speedup of multiple times over standard GRFs without compromising statistical accuracy. Experiments on both simulated and real-world data validate our approach. Our findings suggest that the proposed method is a scalable alternative for localized effect estimation in machine learning and causal inference applications.

## 1. Introduction

In many real-world machine learning (ML) applications, practitioners seek to estimate how quantities of interest vary across different feature subgroups rather than assuming uniform effects. For example, medical interventions and policy treatments often have heterogeneous impacts across subpopulations, making localized estimation crucial for improving outcomes (Imai & Ratkovic, 2013; Knaus et al., 2021; Murdoch et al., 2019; Lee et al., 2020). Similarly, individualized recommendation systems adapt to user-specific features to enhance performance (Kohavi et al., 2013).

A key example of localized estimation arises in causal inference, where modern applications prioritize individualized treatment effects over average treatment effects (Neyman,

<sup>1</sup>Department of Mathematics and Statistics, McGill University, Montreal, Canada <sup>2</sup>Mila - Quebec AI Institute, Montreal, Quebec, Canada. Correspondence to: Archer Y. Yang <archer.yang@mcgill.ca>.

*Proceedings of the 42<sup>nd</sup> International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

1923; Rubin, 1974). The double machine learning framework (Chernozhukov et al., 2018) unifies various ML-based causal estimation methods, including lasso (Belloni et al., 2017), random forests (Athey et al., 2019; Cevid et al., 2022), boosting (Powers et al., 2018), deep learning (Johansson et al., 2016; Shalit et al., 2017), and general-purpose meta-algorithms (Nie & Wager, 2021; Künnel et al., 2019), all of which focus on capturing variation over feature space.

Generalized random forests (GRFs) (Athey et al., 2019; Wager & Athey, 2018) have emerged as a powerful tool for such tasks, leveraging adaptive partitioning with problem-specific moment conditions instead of standard loss-based splits. GRFs apply broadly to a wide range of important statistical models – local linear regression (Friedberg et al., 2020), survival analysis and missing data problems (Cui et al., 2023), nonparametric quantile regression, heterogeneous treatment effect estimation, and nonlinear instrumental variables regression (Athey & Imbens, 2016; Athey et al., 2019). Unlike local linear models (Fan et al., 1995; Fan & Gijbels, 1996; Friedberg et al., 2020) or kernel-based models (Staniswalis, 1989; Severini & Staniswalis, 1994; Lewbel, 2007; Speckman, 1988; Robinson, 1988) which suffer from the curse of dimensionality (Robins & Ritov, 1997), the tree-based approach of GRF offers a more scalable solution.

However, GRFs' gradient-based approach (Athey et al., 2019) becomes computationally expensive and unstable in large dimensions due to the reliance on Jacobian estimators for tree splitting. To address this, we propose a gradient-free approach based on fixed-point iteration, eliminating the need for Jacobian estimation while retaining GRF's theoretical guarantees of consistency and asymptotic normality. Our method significantly improves computational efficiency while maintaining statistical accuracy, achieving significant speedups in experiments on simulated and real-world datasets.

## 2. Background and Related Work

Given data  $(X_i, O_i) \in \mathcal{X} \times \mathcal{O}$ , GRF estimates a target function  $\theta^*(x)$ , defined as the solution to an estimating equation of the form

$$0 = \mathbb{E}_{O|X} [\psi_{\theta^*(x), \nu^*(x)}(O) | X = x], \quad (1)$$

for all  $x \in \mathcal{X}$ , where  $\psi$  is a score function that identifies the true  $(\theta^*(x), \nu^*(x))$  as the root of (1), and  $\nu^*(x)$  is an optional nuisance function. GRF can be understood from a nearest-neighbor perspective as approximating  $\theta^*(x)$  through a locally parametric  $\theta^*$  within small neighborhoods of test point  $x$ . Suppose  $L(x) \subset \{X_i\}_{i=1}^n$  is a subset of training observations of the covariates found in a region around  $x \in \mathcal{X}$  over which  $\theta^*(x)$  can be well-approximated by a local parameter. Observations  $X_i \in L(x)$  serve as local representatives for  $x$  in estimating  $\theta^*(x)$  such that, given sufficiently many training samples in a small enough neighborhood of  $x$ , an empirical version of (1) over  $X_i \in L(x)$  defines an estimator  $\hat{\theta}_{L(x)}$  that approaches  $\theta^*(x)$ ,

$$(\hat{\theta}_{L(x)}, \hat{\nu}_{L(x)}) \in \arg \min_{\theta, \nu} \left\| \sum_{i=1}^n \frac{\mathbb{1}(X_i \in L(x))}{|L(x)|} \cdot \psi_{\theta, \nu}(O_i) \right\|. \quad (2)$$

In GRF, the set of local representatives  $L(x)$  is determined by tree-based partitions which divide the input space into disjoint regions, or leaves. The training samples  $X_i$  that fall in the same leaf as  $x$  form the subset  $L(x)$ . However, single trees are known to have high variance with respect to small changes in the training data (Amit & Geman, 1997; Breiman, 1996; 2001; Dietterich, 2000), leading to estimates (2) that do not generalize well to values of  $x$  that are not part of the training set. GRF improves its estimates by leveraging an estimating function that averages many estimating functions of the form (2). Specifically, let  $L_b(x)$  denote the set of training covariates that fall in the same leaf as  $x$ , identified by a tree trained on an independent subsample of the data, indexed by  $b = 1, \dots, B$ . The GRF estimator is obtained by aggregating the individual estimating functions (2) across a forest of  $B$  independently trained trees, i.e. the solution to the following forest-averaged estimating equation:

$$(\hat{\theta}(x), \hat{\nu}(x)) \in \arg \min_{\theta, \nu} \left\| \frac{1}{B} \sum_{b=1}^B \left( \sum_{i=1}^n \alpha_{bi}(x) \psi_{\theta, \nu}(O_i) \right) \right\|. \quad (3)$$

where  $\alpha_{bi}(x) := \frac{\mathbb{1}(X_i \in L_b(x))}{|L_b(x)|}$ . Define observational weights  $\alpha_i(x)$  that measure the relative frequency with which training sample  $X_i$  falls in the same leaf as  $x$ , averaged over  $B$  trees:

$$\alpha_i(x) := \frac{1}{B} \sum_{b=1}^B \alpha_{bi}(x), \quad (4)$$

for  $i = 1, \dots, n$ . Then, the solution  $(\hat{\theta}(x), \hat{\nu}(x))$  to the forest-averaged model (3) is equivalent to solving the following locally weighted estimating equation

$$(\hat{\theta}(x), \hat{\nu}(x)) \in \arg \min_{\theta, \nu} \left\| \sum_{i=1}^n \alpha_i(x) \psi_{\theta, \nu}(O_i) \right\|. \quad (5)$$

Athey et al. (2019) present (5) as the definition of the GRF estimator, motivated in part by the mature analyses of local kernel methods (Newey, 1994) alongside more recent work on tree-based partitioning and estimating equations (Athey & Imbens, 2016; Zeileis & Hornik, 2007; Zeileis et al., 2008). The GRF algorithm for estimating  $\theta^*(x)$  can be summarized as a two-stage procedure. **Stage I:** Use trees to calculate weight functions  $\alpha_i(x)$  for any test observation  $x \in \mathcal{X}$ , measuring the relative importance of the  $i$ -th training sample to estimating  $\theta^*(\cdot)$  near  $x$ . **Stage II:** Given a test observation  $x \in \mathcal{X}$ , compute estimate  $\hat{\theta}(x)$  of  $\theta^*(x)$  by solving the locally weighted empirical estimating equation (5).

Our contribution improves the computational cost of Stage I by introducing a more efficient procedure to train the trees. Training the forest is the most resource-intensive step of GRF, and the cost of each split in the existing approach scales quadratically with the dimension of  $\theta^*(x)$ . We adopt a gradient-free splitting mechanism and significantly reduce both the time and memory demands of Stage I. Crucially, solving Stage II with weights  $\alpha_i(x)$  following our streamlined Stage I produces an estimator  $\hat{\theta}(x)$  that preserves the finite-sample performance and asymptotic guarantees of GRF.

### 3. Our Method

In this section we describe the details of our accelerated algorithm for GRF. We closely follow the approach of Athey et al. (2019), and define  $\hat{\theta}(x)$  as the solution to a locally weighted problem (5) with weighting functions  $\alpha_i(x)$  of the form (4). The weight functions are induced by a collection of local subsets  $\{L_b(x)\}_{b=1}^B$ , such that each subset  $L_b(x)$  is determined by the partition rules of a tree trained on a subsample. The construction of each tree, in turn, is determined by recursive splits of the subsample based on a splitting criterion designed to identify regions of  $\mathcal{X}$  that are homogeneous with respect to  $\theta^*(x)$ . Therefore, to fully specify the weight functions  $\alpha_i(x)$ , we must describe a feasible criterion for producing a split of  $\mathcal{X}$ .

#### 3.1. The target tree-splitting criterion for Stage I

In GRF, the goal of Stage I is to use recursive tree-based splits of the training data to induce a partition over the input space. Each split starts with a parent node  $P \subset \mathcal{X}$  and results in child nodes  $C_1, C_2 \subset \mathcal{X}$ , defined by a binary, axis-aligned splitting rule of the form  $C_1 = \{X_i : X_{i,\ell} \leq t\}$  and  $C_2 = \{X_i : X_{i,\ell} > t\}$ , where  $\ell$  denotes a candidate splitting feature/axis and  $t \in \mathbb{R}$  the splitting threshold. For a parent  $P$  and any child nodes  $C_1, C_2$  of  $P$ , let  $(\hat{\theta}_P, \hat{\nu}_P)$  and  $(\hat{\theta}_{C_j}, \hat{\nu}_{C_j})$  denote local solutions analogous to (2) defined

over the samples in  $P$  and  $C_j$ , respectively:

$$(\hat{\theta}_P, \hat{\nu}_P) \in \arg \min_{\theta, \nu} \left\| \sum_{\{i: X_i \in P\}} \psi_{\theta, \nu}(O_i) \right\|, \quad (6)$$

$$(\hat{\theta}_{C_j}, \hat{\nu}_{C_j}) \in \arg \min_{\theta, \nu} \left\| \sum_{\{i: X_i \in C_j\}} \psi_{\theta, \nu}(O_i) \right\|, \quad (7)$$

for  $j = 1, 2$ . A strategy to split  $P$  into two subsets of greater homogeneity with respect to  $\theta^*(\cdot)$  is as follows: Find child nodes  $C_1$  and  $C_2$  such that the total deviation between the local solutions  $\hat{\theta}_{C_j}$  and the target  $\theta^*(X)$  is minimized, conditional on  $X \in C_j$ ,  $j = 1, 2$ . A natural measure of deviation is the squared-error loss,

$$\begin{aligned} \text{err}(C_1, C_2) &:= \sum_{j=1,2} \mathbb{P}(X \in C_j \mid X \in P) \\ &\quad \times \mathbb{E} \left[ \left\| \theta^*(X) - \hat{\theta}_{C_j} \right\|^2 \mid X \in C_j \right], \end{aligned}$$

such that the resulting split  $(C_1, C_2)$  corresponds to least-squares optimal solutions  $\hat{\theta}_{C_1}$  and  $\hat{\theta}_{C_2}$ . However,  $\text{err}(C_1, C_2)$  is intractable since  $\theta^*(\cdot)$  is unknown. GRF considers a criterion that measures heterogeneity across a pair of local solutions over a candidate split

$$\Delta(C_1, C_2) := \frac{n_{C_1} n_{C_2}}{n_P^2} \left\| \hat{\theta}_{C_1} - \hat{\theta}_{C_2} \right\|^2, \quad (8)$$

where  $n_{C_1}$ ,  $n_{C_2}$ , and  $n_P$  denote the number of observations in  $C_1$ ,  $C_2$ , and  $P$ , respectively. In particular, rather than minimizing  $\text{err}(C_1, C_2)$ , one can seek a split of  $P$  such that the cross-split heterogeneity between  $\hat{\theta}_{C_1}$  and  $\hat{\theta}_{C_2}$  is maximized. [Athey et al. \(2019\)](#) observe that  $\text{err}(C_1, C_2)$  and  $\Delta(C_1, C_2)$  are coupled according to  $\text{err}(C_1, C_2) = K(P) - \mathbb{E}[\Delta(C_1, C_2)] + o(r^2)$ , where  $r > 0$  is a small radius term tied to the sampling variance, and  $K(P)$  does not depend on the split of  $P$ . That is, splits that maximize  $\Delta(C_1, C_2)$  – which emphasize the heterogeneity of  $\hat{\theta}_{C_j}$  across a split – will asymptotically minimize  $\text{err}(C_1, C_2)$ , which aims to improve the homogeneity of  $\hat{\theta}_{C_j}$  within a split.

Although the criterion  $\Delta(C_1, C_2)$  is computable, evaluating it is very computationally expensive since it requires solving (7) to obtain  $\hat{\theta}_{C_1}, \hat{\theta}_{C_2}$  for all possible splits of  $P$ , and closed-form solutions for  $\hat{\theta}_{C_j}$  are generally not available except in special cases of  $\psi$ . Instead, GRF approximates the target  $\Delta$ -criterion based on a criterion of the form

$$\tilde{\Delta}^{\text{grad}}(C_1, C_2) := \frac{n_{C_1} n_{C_2}}{n_P^2} \left\| \tilde{\theta}_{C_1}^{\text{grad}} - \tilde{\theta}_{C_2}^{\text{grad}} \right\|^2, \quad (9)$$

where  $\tilde{\theta}_{C_j}^{\text{grad}}$  denotes a *gradient-based* approximation of  $\hat{\theta}_{C_j}$ . Specifically,  $\tilde{\theta}_{C_j}^{\text{grad}}$  is a first-order approximation interpreted

as the result of taking a gradient step away from the parent estimate in the direction towards the true child solution  $\hat{\theta}_{C_j}$ :

$$\tilde{\theta}_{C_j}^{\text{grad}} := \hat{\theta}_P - \frac{1}{n_{C_j}} \sum_{\{i: X_i \in C_j\}} \xi^\top A_P^{-1} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i), \quad (10)$$

where  $(\hat{\theta}_P, \hat{\nu}_P)$  is the local solution over the parent,  $A_P$  is any consistent estimator of the local Jacobian matrix  $\nabla_{(\theta, \nu)} \mathbb{E}[\psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) \mid X_i \in P]$ , and  $\xi^\top$  can be thought of as a term that selects a  $\theta$ -subvector from a  $(\theta, \nu)$ -vector, e.g. if  $\theta \in \mathbb{R}^K$  and  $\nu \in \mathbb{R}$ , then  $\xi^\top$  such that  $\theta = \xi^\top (\theta, \nu)^\top$  is the rectangular diagonal matrix  $\xi^\top = [\mathbb{I}_K \ 0]$ . When the scoring function  $\psi$  is continuously differentiable in  $(\theta, \nu)$ , the Jacobian estimator  $A_P$  can be computed as

$$\begin{aligned} A_P &= \nabla_{(\theta, \nu)} \frac{1}{n_P} \sum_{\{i: X_i \in P\}} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) \\ &= \frac{1}{n_P} \sum_{\{i: X_i \in P\}} \nabla_{(\theta, \nu)} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i). \end{aligned} \quad (11)$$

### 3.2. Limitations of gradient-based approximation

The use of the Jacobian estimator  $A_P$  in (10) introduces considerable computational challenges. First, each parent node  $P$  in every tree of the forest requires a distinct  $A_P$  matrix, which imposes a significant computational burden when explicitly calculating  $A_P^{-1} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i)$  to determine  $\tilde{\theta}_{C_j}^{\text{grad}}$ . Second, if the local Jacobian  $\nabla_{(\theta, \nu)} \mathbb{E}[\psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) \mid X_i \in P]$  is ill-conditioned, then the resulting  $A_P$  estimator may be nearly singular. This instability can lead to highly variable gradient-based approximations  $\tilde{\theta}_{C_j}^{\text{grad}}$  and highly variable splits of  $P$ . For example, consider the following varying-coefficient model for an outcome  $Y_i$  given regressors  $W_i = (W_{i,1}, \dots, W_{i,K})^\top$  in the presence of mediating auxiliary covariates  $X_i$ :

$$\mathbb{E}[Y_i \mid X_i = x] = \nu^*(x) + W_i^\top \theta^*(x), \quad (12)$$

where  $\nu^*(\cdot)$  is a nuisance intercept function and  $\theta^*(x) = (\theta_1^*(x), \dots, \theta_K^*(x))^\top$  are the target coefficients. Models of the form (12) encompass time- or spatially-varying coefficient frameworks, where  $(X_i, Y_i, W_i)$  represent the  $i$ -th sample associated with spatiotemporal values  $X_i$ . Such models are particularly relevant in applications like heterogeneous treatment effects; see Section 5 for a more in-depth discussion. The local estimating function  $\psi_{\theta, \nu}(Y_i, W_i)$ , identifying  $(\theta^*(x), \nu^*(x))$  through moment conditions as in (1), is given by:

$$\psi_{\theta, \nu}(Y_i, W_i) := \begin{bmatrix} (Y_i - W_i^\top \theta - \nu) \cdot W_i \\ Y_i - W_i^\top \theta - \nu \end{bmatrix}.$$

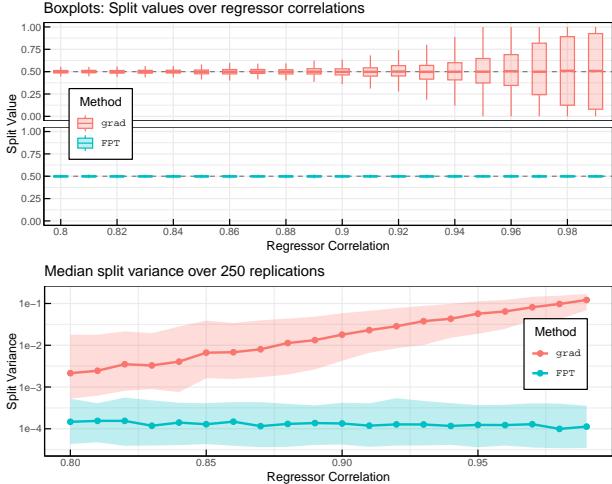


Figure 1: Splits values (top) and split variance (bottom), with 10th and 90th percentile bands, across correlations of  $W_{i,1}$  and  $W_{i,2}$ .

Consequently, the corresponding local Jacobian estimator is

$$A_P = \frac{1}{n_P} \sum_{\{i:X_i \in P\}} \nabla_{(\theta,\nu)} \psi_{\theta,\nu}(Y_i, W_i) = -\frac{1}{n_P} \sum_{\{i:X_i \in P\}} \begin{bmatrix} W_i W_i^\top & W_i^\top \\ W_i & 1 \end{bmatrix}. \quad (13)$$

When the regressors are highly correlated, the summation over the  $W_i W_i^\top$  block of the  $A_P$  matrix leads to nearly singular values of  $A_P$ , resulting in an unstable matrix inverse  $A_P^{-1}$ , and therefore unstable values of  $\hat{\theta}_{C_j}^{\text{grad}}$  and unstable splits. This issue becomes more pronounced as the number of parent samples  $n_P$  decreases, as is the case at deeper levels of the tree. These challenges highlight the limitations of relying on  $A_P$  as part of an approximation for the child solutions  $\hat{\theta}_{C_j}$ .

As an illustration, consider a simple varying coefficient model with primary regressors  $W_{i,1}, W_{i,2} \sim \mathcal{N}(0, 1)$ , auxiliary covariates  $X_i \sim \text{Unif}(0, 1)$ , and outcomes  $Y_i$  generated as

$$Y_i = \mathbf{1}(X_i > 0.5)W_{i,1} + W_{i,2} + \epsilon_i, \quad (14)$$

where  $\epsilon_i \sim \mathcal{N}(0, 1)$ . Figure 1 illustrates the distribution of 2000  $\tilde{\Delta}^{\text{grad}}$ -optimal binary splits (gradient-based tree stumps) fit over 1000 samples of the varying coefficient model (14), repeated over different regressor correlation levels  $\text{Corr}(W_{i,1}, W_{i,2}) \in \{0.80, 0.81, \dots, 0.98, 0.99\}$ . It is clear that splits based on the  $\tilde{\Delta}^{\text{grad}}$ -criterion exhibit high variability when the correlation between the regressors is large. In contrast, our proposed method, discussed in the next section, does not suffer from the same problem.

### 3.3. Fixed-point approximation

To address the limitations of gradient-based approximations, we propose a gradient-free approach based on the form of a single fixed-point iteration. Let  $\Psi_{C_j}(\theta, \nu) := \frac{1}{n_{C_j}} \sum_{\{i:X_i \in C_j\}} \psi_{\theta,\nu}(O_i)$  denote the empirical estimating function for the child solution  $(\hat{\theta}_{C_j}, \hat{\nu}_{C_j})$  such that (7) is equivalently written as:

$$(\hat{\theta}_{C_j}, \hat{\nu}_{C_j}) \in \arg \min_{\theta, \nu} \|\Psi_{C_j}(\theta, \nu)\|, \quad j = 1, 2. \quad (15)$$

Under mild regularity conditions,  $(\hat{\theta}_{C_j}, \hat{\nu}_{C_j})$  is a  $Z$ -estimator that solves the estimating equation  $\Psi_{C_j}(\theta, \nu) = \mathbf{0}$ . Reformulating this equation as a fixed-point problem, we write:

$$(\theta, \nu) = \underbrace{(\theta, \nu) - \eta \Psi_{C_j}(\theta, \nu)}_{=: f(\theta, \nu)}, \quad \eta > 0. \quad (16)$$

A necessary and sufficient condition for  $(\hat{\theta}_{C_j}, \hat{\nu}_{C_j})$  to be a solution of (15) is characterized by the fixed-point problem  $(\hat{\theta}_{C_j}, \hat{\nu}_{C_j}) = f(\hat{\theta}_{C_j}, \hat{\nu}_{C_j})$ , where  $f$  is as defined in (16). Iterative fixed-point methods (Picard, 1890; Lindelöf, 1894; Banach, 1922; Ryu & Boyd, 2016; Yang et al., 2021) solve such problems by considering an update rule of the form

$$(\theta^+, \nu^+) \leftarrow f(\theta, \nu). \quad (17)$$

The form of (17) inspires us to approximate the true child solution  $\hat{\theta}_{C_j}$  using a single fixed-point update taken from the parent solution  $\hat{\theta}_P$ :

$$\begin{aligned} \tilde{\theta}_{C_j}^{\text{FPT}} &:= \hat{\theta}_P - \eta \xi^\top \Psi_{C_j}(\hat{\theta}_P, \hat{\nu}_P) \\ &= \hat{\theta}_P - \frac{\eta}{n_{C_j}} \xi^\top \sum_{\{i:X_i \in C_j\}} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i), \end{aligned} \quad (18)$$

where the product with  $\xi^\top$  is interpreted similarly to its role in the gradient-based approximation (10) and to express the update (17) solely in terms of the target  $\theta$ -quantity. We interpret  $\tilde{\theta}_{C_j}^{\text{FPT}}$  as an approximation of  $\hat{\theta}_{C_j}$  obtained by taking a step from  $\hat{\theta}_P$  in a direction that reduces the magnitude of the local estimating function  $\Psi_{C_j}$ . Notably, the approximation  $\tilde{\theta}_{C_j}^{\text{FPT}}$  does not involve the  $A_P$  matrix, relying only on the scores  $\psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i)$  evaluated at the parent solutions. In general, removing the inverse  $A_P^{-1}$  provides computational cost savings of  $\mathcal{O}(K^3)$ . The corresponding splitting criterion, which uses the fixed-point approximations  $\tilde{\theta}_{C_j}^{\text{FPT}}$  as substitutes for  $\hat{\theta}_{C_j}$  is given by

$$\tilde{\Delta}^{\text{FPT}}(C_1, C_2) := \frac{n_{C_1} n_{C_2}}{n_P^2} \left\| \tilde{\theta}_{C_1}^{\text{FPT}} - \tilde{\theta}_{C_2}^{\text{FPT}} \right\|^2. \quad (19)$$

Revisiting the varying coefficient example from Section 3.2, we see that splits based on fixed-point approximations  $\tilde{\theta}_{C_j}^{\text{FPT}}$

are significantly more stable than those based on  $\tilde{\theta}_{C_j}^{\text{grad}}$ . Specifically, Figure 1 illustrates that splits that maximize  $\Delta^{\text{FPT}}(C_1, C_2)$  are more robust to ill-conditioning in the underlying local Jacobian  $\nabla_{(\theta, \nu)} \mathbb{E}[\psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) | X_i \in P]$ , as is the case for highly correlated regressors in the varying coefficient model (14), and leading to highly stable splits.

### 3.4. Pseudo-outcomes

Approximations  $\tilde{\theta}_{C_j}$  of the form (10) and (18) offer an additional benefit: they enable the  $\tilde{\Delta}$ -criteria of the form (9) and (19) to be efficiently optimized through a single multivariate CART split. A CART split performed with respect to vector-valued responses  $\rho_i \in \mathbb{R}^K$  over a parent node  $P$  produces a split  $(C_1, C_2)$  that minimizes the following least-squares criterion:

$$\sum_{\{i : X_i \in C_1\}} \|\rho_i - \bar{\rho}_{C_1}\|^2 + \sum_{\{i : X_i \in C_2\}} \|\rho_i - \bar{\rho}_{C_2}\|^2, \quad (20)$$

where  $\bar{\rho}_{C_j} := \frac{1}{n_{C_j}} \sum_{\{i : X_i \in C_j\}} \rho_i$ .<sup>1</sup> Equivalently, a CART split that minimizes (20) will maximize:

$$n_{C_1} \|\bar{\rho}_{C_1}\|^2 + n_{C_2} \|\bar{\rho}_{C_2}\|^2. \quad (21)$$

The equivalence between the split that minimizes the least-squares CART criterion (20) and the split that maximizes (21) is shown in Appendix B.1.1. GRF performs its splits by adopting gradient-based *pseudo-outcomes*, defined as

$$\rho_i^{\text{grad}} := -\xi^\top A_P^{-1} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) \quad (22)$$

such that the gradient-based approximation  $\tilde{\theta}_{C_j}^{\text{grad}}$  in (10) is equivalently written:

$$\tilde{\theta}_{C_j}^{\text{grad}} = \hat{\theta}_P + \frac{1}{n_{C_j}} \sum_{\{i : X_i \in C_j\}} \rho_i^{\text{grad}} = \hat{\theta}_P + \bar{\rho}_{C_j}^{\text{grad}}.$$

In the case of fixed-point approximation, we define fixed-point pseudo-outcomes:

$$\rho_i^{\text{FPT}} := -\eta \xi^\top \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i), \quad \eta \neq 0, \quad (23)$$

such that the fixed-point approximation  $\tilde{\theta}_{C_j}^{\text{FPT}}$  in (18) is equivalently written as

$$\tilde{\theta}_{C_j}^{\text{FPT}} = \hat{\theta}_P + \frac{1}{n_{C_j}} \sum_{\{i : X_i \in C_j\}} \rho_i^{\text{FPT}} = \hat{\theta}_P + \bar{\rho}_{C_j}^{\text{FPT}}. \quad (24)$$

Substitute the above form of  $\tilde{\theta}_{C_j}^{\text{FPT}}$  into the  $\tilde{\Delta}^{\text{FPT}}$ -criterion (19) to equivalently express the criterion in terms of the FPT pseudo-outcomes:

$$\tilde{\Delta}^{\text{FPT}}(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P^2} \|\bar{\rho}_{C_1}^{\text{FPT}} - \bar{\rho}_{C_2}^{\text{FPT}}\|^2, \quad (25)$$

<sup>1</sup>The multivariate CART criterion uses a sum of squares impurity measure, as in De'ath (2002); Segal (1992).

where an analogous equivalence holds for  $\tilde{\Delta}^{\text{grad}}$  in terms of the gradient-based pseudo-outcomes. We demonstrate in Lemma B.1 (in Appendix B.1.2) that maximizing the fixed-point criterion  $\tilde{\Delta}^{\text{FPT}}(C_1, C_2)$  is equivalent to maximizing the CART criterion (21), and extend this property to any  $\tilde{\Delta}$ -style criterion induced by pseudo-outcomes that can be expressed as a split-independent linear transformation of the parent scores  $\psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i)$ .

Note that our method does not rely on iterative fixed-point procedures at all. Instead, it uses only a single step of fixed-point approximation to simplify the pseudo-outcomes. These simplified pseudo-outcomes are then passed directly to a standard CART algorithm for splitting. The numerical convergence of our method therefore relies solely on CART's established and well-known stability, not on fixed-point iteration. CART splits on pseudo-outcomes are computationally efficient. Given a parent node  $P$ , the value  $\rho_i = -B \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i)$  does not depend on a candidate split  $(C_1, C_2)$  for any matrix  $B$  that is fixed with respect to the parent. This allows much of the computation required to maximize  $\tilde{\Delta}^{\text{FPT}}(C_1, C_2)$  to be done at the parent level, and in particular avoids re-calculating the approximations  $\tilde{\theta}_{C_1}^{\text{FPT}}$  and  $\tilde{\theta}_{C_2}^{\text{FPT}}$  across the sequence of candidate splits. Once  $P$  is fixed and  $\rho_i^{\text{FPT}}$  are computed, the value of  $\tilde{\Delta}^{\text{FPT}}(C_1, C_2)$  for the first candidate split requires  $\mathcal{O}(n_P)$  time, and the value for all other candidate splits of  $P$  are queried in  $\mathcal{O}(1)$  time. While gradient-based pseudo-outcomes share this property, the use of fixed-point pseudo-outcomes eliminates the computational overhead and instability associated with estimating  $A_P$ , as discussed in Section 3.2.

We show in Lemma B.2 (Appendix B.1.3) that choosing different values of  $\eta$  does not change the outcome of the fixed-point splitting mechanism. Specifically, the optimal split identified by CART on pseudo-outcomes  $\rho_i^{\text{FPT}}$  of the form (23) does not depend on  $\eta$ . This can be heuristically understood by studying how the criterion changes as a function of the candidate splits. To illustrate, we consider a VCM model of the form (12) for bivariate regressors  $W_i$ , univariate  $X_i \in [0, 1]$ , and scalar outcomes  $Y_i$ . A detailed summary of the settings is found in Appendix D.1. The sequence of valid candidate child nodes obtained by a split over univariate  $X_i$  can be parameterized through scalar  $t$  as  $C_1(t) := \{X_i : X_i \leq t\}$  and  $C_2(t) := \{X_i : X_i > t\}$ . Let  $\Delta(t) := \Delta(C_1(t), C_2(t))$  denote the parameterized target criterion (8), and consider the behavior of  $\Delta(t)$ ,  $\tilde{\Delta}^{\text{grad}}(t)$ , and two fixed-point criteria  $\tilde{\Delta}_1^{\text{FPT}}(t)$  and  $\tilde{\Delta}_2^{\text{FPT}}(t)$  of the form (25) based on pseudo-outcomes with scale factors  $\eta = 1$  and  $\eta = 1/\sqrt{2}$ , respectively. Figure 2 illustrates the different splitting criteria values plotted against the sequence of candidate splits. The visualization clearly shows that the criteria curves for  $\Delta(t)$ ,  $\tilde{\Delta}^{\text{grad}}(t)$ , and  $\tilde{\Delta}_1^{\text{FPT}}(t)$  with  $\eta = 1$  are all very close to one

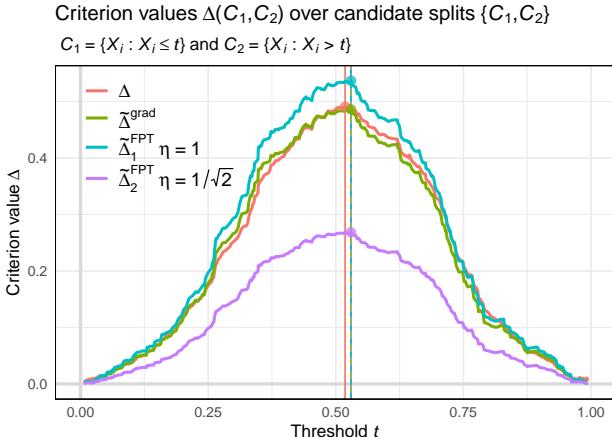


Figure 2: Criterion values across candidate splits  $(C_1(t), C_2(t))$  over threshold  $t \in [0, 1]$ . The location of the optimal split under each criterion is given by the corresponding vertical line.

another. Critically, the fixed-point criterion with  $\eta = 1/\sqrt{2}$ , i.e.  $\tilde{\Delta}_2^{\text{FPT}}(t)$ , although scaled differently, still identifies the same maximizing split as  $\tilde{\Delta}_1^{\text{FPT}}(t)$ . This is because CART chooses a split based on a rank ordering of the criterion over all candidate splits. The absolute scale of the CART criterion does not matter, and it is only criterion rankings over the candidates that determines the optimal split. Therefore, choosing a different scalar  $\eta$  does not change the outcome of the splitting process.

Based on the scale-invariance of our splitting criterion, we now detail the recursive procedure for growing our fixed-point trees pseudo-outcomes with  $\eta = 1$ .

**The fixed-point tree algorithm.** The entire fixed-point tree-growing procedure recursively applies the following two steps on a given parent node  $P$ :

- (i) **Labeling:** Solve (6) over  $P$  to obtain the parent estimate  $(\hat{\theta}_P, \hat{\nu}_P)$ . Compute the pseudo-outcomes:

$$\rho_i^{\text{FPT}} := -\xi^\top \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i), \quad (26)$$

for all  $i$  such that  $X_i \in P$ .

- (ii) **Regression:** Maximize  $\tilde{\Delta}^{\text{FPT}}(C_1, C_2)$  by performing a CART split on the pseudo-outcomes  $\rho_i^{\text{FPT}}$  over  $P$ .

### 3.5. Estimates of $\hat{\theta}(x)$ for Stage II

The fixed-point tree algorithm generates a single tree-based partition of  $\mathcal{X}$ . Repeating this process over subsamples of the training data yields a forest of trees, each specifying local leaf functions  $L_b(x)$ . These leaf functions define the local weight functions  $\alpha_i(x)$  via (4), completing Stage I of GRF. The full fixed-point tree training algorithm is described in Algorithm 1, while Algorithm 2 provides the pseudocode for the forest-wide Stage I procedure.

To compute the final GRF estimates  $\hat{\theta}(x)$  for the target  $\theta^*(x)$ , we follow the standard GRF mechanism for Stage II. After the fixed-point trees are trained in Stage I, a test observation  $x_0 \in \mathcal{X}$  is assigned to local leaves  $L_b(x_0)$ , indexed by trees  $b \in \{1, \dots, B\}$ . Each leaf  $L_b(x_0)$  contains the training observations that fall into the same leaf as  $x_0$  in tree  $b$ . Using these local leaves, the forest computes training weights  $\alpha_i(x_0)$  as in (4). The final estimate  $\hat{\theta}(x_0)$  is obtained by solving the locally weighted estimating equation (5).

Importantly, as discussed in Section 2, solving for  $\hat{\theta}(x_0)$  in Stage II is independent of the specific mechanism used in Stage I. The only requirement is that Stage I produces valid weights. This ensures that Stage II remains a standard weighted estimating equation, enabling the fixed-point tree algorithm to integrate seamlessly into GRF's two-stage framework. We refer to the complete algorithm for estimating  $\theta^*(x)$  using fixed-point trees as GRF-FPT. By preserving Stage II of GRF, the GRF-FPT estimator  $\hat{\theta}(x)$  retains GRF's theoretical guarantees of consistency and asymptotic normality while offering a computationally efficient tree-building method. Pseudocode for Stage II of the GRF-FPT algorithm is provided in Algorithm 3, located in Appendix C.3.

## 4. Theoretical Analysis

In this section, we provide a theoretical foundation for the GRF-FPT estimator  $\hat{\theta}(x)$ . For Stage I, Proposition 4.1 establishes an asymptotic equivalence between the FPT criterion and a weighted oracle criterion  $\Delta_V(C_1, C_2)$  in (27), while Lemma 4.2 demonstrates that the Specifications A.2 are met by a forest based on the  $\Delta_V$ -criterion whenever they are met by a forest based on the  $\Delta$ -criterion. Assumptions A.1 and Specifications A.2 are the sufficient conditions for the consistency and asymptotic normality of  $\hat{\theta}(x)$  in (5), and thus are used to formally justify the FPT algorithm as a mechanism for specifying an estimator of  $\theta^*(x)$ .

**Proposition 4.1.** Suppose Assumptions A.1 hold, and assume moreover Neyman orthogonal moment conditions (defined in Appendix A.4). Denote by  $r := \sup_{\{i: X_i \in P\}} \|X_i - x_P\|$  the radius of the parent  $P$ , where  $x_P$  denotes the center of mass over  $X_i \in P$ . Let  $V_{\theta\theta}(x_P)$  denote the  $\theta$ -block of  $V(x_P)$  in (37). Denote by  $\|\cdot\|_V$  the weighted Euclidean norm  $\|z\|_V := \|V_{\theta\theta}(x_P)z\|_2 = \sqrt{z^\top V_{\theta\theta}^\top(x_P)V_{\theta\theta}(x_P)z}$ . Define the weighted oracle criterion  $\Delta_V(C_1, C_2)$ :

$$\Delta_V(C_1, C_2) := \frac{n_{C_1} n_{C_2}}{n_P^2} \left\| \hat{\theta}_{C_1} - \hat{\theta}_{C_2} \right\|_V^2. \quad (27)$$

Then, treating the split as fixed with  $r^{-2} \ll n_{C_1}, n_{C_2}$  and

sufficiently small  $r > 0$ ,

$$\tilde{\Delta}^{\text{FPT}}(C_1, C_2) = \Delta_V(C_1, C_2) + o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right).$$

**Lemma 4.2.** Let  $\mathcal{T}(\Delta)$  denote a tree whose splitting mechanism seeks splits that maximize  $\Delta(C_1, C_2)$  defined in (8), and let  $\mathcal{T}(\Delta_V)$  denote a tree whose splitting mechanism seeks splits that maximize  $\Delta_V(C_1, C_2)$  defined in (27). Suppose Assumptions A.1 hold and assume moreover that  $\mathcal{T}(\Delta)$  is a tree that satisfies Specifications A.2. Then,  $\mathcal{T}(\Delta_V)$  satisfies Specifications A.2.

For Stage II, Theorem 4.3 establishes the consistency of the GRF-FPT estimator  $\hat{\theta}(x)$ :

**Theorem 4.3.** Suppose that Assumptions A.1 hold, and let  $(\hat{\theta}(x), \hat{\nu}(x))$  be estimates that solve (5) based on weights induced by a forest of trees grown under the fixed-point tree algorithm satisfying Specifications A.2. Then,  $(\hat{\theta}(x), \hat{\nu}(x))$  converges in probability to  $(\theta^*(x), \nu^*(x))$ .

The proof of Theorem 4.3 follows directly from Theorem 3 of Athey et al. (2019), which, under Assumptions A.1, establishes consistency for estimates  $(\hat{\theta}(x), \hat{\nu}(x))$  that solve (5) with weights from a forest that satisfies Specifications 1-5. Thanks to Lemma 4.2, these forest specifications must also apply to a forest grown under the FPT mechanism. Specifications 1-3 collectively impose mild boundary conditions on the splitting procedure. Meanwhile, Specification 4 requires that trees are trained on subsamples drawn without replacement (Biau et al., 2008; Scornet et al., 2015; Wager et al., 2014; Wager & Athey, 2018), and Specification 5 requires that trees must be grown using an additional subsample splitting mechanism known as honesty (Athey & Imbens, 2016; Biau, 2012; Denil et al., 2014). Appendix C.1 provides a detailed explanation of the subsampling and honest sample splitting procedure.

Finally, Theorem 4.4 establishes the asymptotic normality of the GRF-FPT estimator  $\hat{\theta}(x)$ :

**Theorem 4.4.** Under the conditions of Theorem 4.3, suppose moreover that Regularity Condition 1 holds, and that a forest is grown on subsamples of size  $s$  scaling as  $s = n^\beta$ , where  $\beta$  satisfies Regularity Condition 2. Then, there exists a sequence  $\sigma_n(x)$  such that  $(\hat{\theta}_n(x) - \theta^*(x))/\sigma_n(x) \rightsquigarrow \mathcal{N}(0, 1)$  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 of the inverse sampling ratio  $\log(n/s)$ .

The proof of Theorem 4.4 is an immediate consequence of Theorem 5 of Athey et al. (2019). Theorems 4.3 and 4.4 demonstrate that the GRF-FPT estimator is able to meet key statistical guarantees.

## 5. Applications

In this section, we explore applications of GRF-FPT for two related models: varying coefficient models and heterogeneous treatment effects. We consider an outcome model of the form introduced in Section 3.2. For each observation, let  $Y_i$  denote the observed outcome,  $W_i = (W_{i,1}, \dots, W_{i,K})^\top$  a  $K$ -dimensional regressor, and  $X_i$  a set of mediating auxiliary variables, such that

$$Y_i = \nu^*(X_i) + W_i^\top \theta^*(X_i) + \epsilon_i, \quad (28)$$

where  $\nu^*(\cdot)$  is a nuisance intercept function,  $\theta^*(x) = (\theta_1^*(x), \dots, \theta_K^*(x))^\top$  are the target effect functions local to  $X_i = x$ , under the assumptions  $\mathbb{E}[\epsilon_i | X_i = x] = 0$  and  $\mathbb{E}[\epsilon_i W_i | X_i = x] = 0$ .

**Varying coefficient models (VCM).** Given regressors  $W_i \in \mathbb{R}^K$ , models of the form (28) can be characterized as varying coefficient models (Hastie & Tibshirani, 1993). As discussed in Section 3.2, we must also assume that the regressors  $W_i$  are conditionally exogenous given  $X_i = x$ .

**Heterogeneous treatment effects (HTE).** A special case of (28) arises within the Neyman-Rubin potential outcome framework, which models the causal effect of treatment on an outcome (Neyman, 1923; Rubin, 1974). Here,  $\theta^*(x) = (\theta_1^*(x), \dots, \theta_K^*(x))^\top$  represents heterogeneous treatment effects associated with  $K$  discrete treatment levels. Let  $T_i \in \{1, \dots, K\}$  denote the observed treatment level for the  $i$ -th observation, and  $Y_i(k)$  the potential outcome that would have been observed if treatment level  $k$  had been applied. The regressors  $W_i \in \{0, 1\}^K$  in (28) are interpreted as a vector of dummy variables indicating the observed treatment level,  $W_{i,k} := \mathbf{1}(T_i = k)$ . The auxiliary variables  $X_i$  account for potential confounding effects. The conditional average treatment effect of treatment level  $k \in \{2, \dots, K\}$  relative to the baseline level  $k = 1$  is then defined as:

$$\theta_k^*(x) := \mathbb{E}[Y_i(k) - Y_i(1) | X_i = x],$$

where the baseline contrast is set to  $\theta_1^*(x) := 0$ .

Under exogeneity of the regressors, the target effects  $\theta^*(x)$  in models (28) are identified by moment conditions (1) for scoring function (Angrist & Pischke, 2009; Athey et al., 2019)

$$\psi_{\theta, \nu}(Y_i, W_i) := \begin{bmatrix} (Y_i - W_i^\top \theta - \nu) \cdot W_i \\ Y_i - W_i^\top \theta - \nu \end{bmatrix}.$$

The gradient-based pseudo-outcomes (22) are computed as

$$\rho_i^{\text{grad}} = -A_P^{-1}(W_i - \bar{W}_P) \left( Y_i - \bar{Y}_P - (W_i - \bar{W}_P)^\top \hat{\theta}_P \right), \quad (29)$$

where  $\bar{W}_P$  and  $\bar{Y}_P$  are the local means of  $W_i$  and  $Y_i$  over the observations in  $P$ . Centering  $Y_i - \bar{Y}_P$  and  $W_i - \bar{W}_P$

removes the baseline effect of the mean  $\hat{\nu}_P$  on  $\rho_i^{\text{grad}}$ , and where  $A_P$  is given by (13) as:

$$A_P = -\frac{1}{n_P} \sum_{\{i:X_i \in P\}} (W_i - \bar{W}_P)(W_i - \bar{W}_P)^\top. \quad (30)$$

Computing  $\rho_i^{\text{grad}}$  in (29) involves the OLS coefficients  $\hat{\theta}_P$  from regressing  $Y_i - \bar{Y}_P$  on  $W_i - \bar{W}_P$ , over the observations in  $P$ :

$$\hat{\theta}_P := -A_P^{-1} \frac{1}{n_P} \sum_{\{i:X_i \in P\}} (W_i - \bar{W}_P)(Y_i - \bar{Y}_P). \quad (31)$$

In comparison,  $\rho_i^{\text{FPT}}$  in (26) are computed as:

$$\begin{aligned} \rho_i^{\text{FPT}} &:= -\xi^\top \psi_{\hat{\theta}_P, \hat{\nu}_P}(Y_i, W_i), \\ &= -(W_i - \bar{W}_P) \left( Y_i - \bar{Y}_P - (W_i - \bar{W}_P)^\top \hat{\theta}_P \right), \end{aligned} \quad (32)$$

The relationship  $\rho_i^{\text{grad}} = A_P^{-1} \rho_i^{\text{FPT}}$  reveals a significant benefit of FPT pseudo-outcomes. The form of  $\rho_i^{\text{FPT}}$  eliminates the computational cost associated with the multiplication of  $A_P^{-1}$ , leading to  $\mathcal{O}(K^3)$  computational savings. Furthermore, the computation of  $\hat{\theta}_P$  in (32) no longer requires solving for  $A_P^{-1}$ . Therefore, we can further enhance computational efficiency by using an accelerated form of pseudo-outcome  $\phi_i^{\text{FPT}}$  instead of  $\rho_i^{\text{FPT}}$ :

$$\phi_i^{\text{FPT}} := -(W_i - \bar{W}_P) \left( Y_i - \bar{Y}_P - (W_i - \bar{W}_P)^\top \tilde{\theta}_P \right), \quad (33)$$

where  $\hat{\theta}_P$  is replaced by  $\tilde{\theta}_P$  in (32), which is defined as a one-step gradient descent approximation of  $\hat{\theta}_P$  taken from the origin:

$$\tilde{\theta}_P := \gamma \frac{1}{n_P} \sum_{\{i:X_i \in P\}} (W_i - \bar{W}_P)(Y_i - \bar{Y}_P). \quad (34)$$

Here,  $\gamma$  denotes the exact line search step size for the regression of  $Y_i - \bar{Y}_P$  on  $W_i - \bar{W}_P$  over  $P$ :

$$\gamma := \frac{\|(W - \bar{W}_P)^\top (Y - \bar{Y}_P)\|_2^2}{\|(W - \bar{W}_P)(W - \bar{W}_P)^\top (Y - \bar{Y}_P)\|_2^2}, \quad (35)$$

where  $W = [W_1 \cdots W_{n_P}]^\top$  and  $Y = [Y_1 \cdots Y_{n_P}]^\top$  with the notation  $W - \bar{W}_P$  and  $Y - \bar{Y}_P$  understood as row-wise centering.

The computational cost associated with  $\tilde{\theta}_P$  is comparatively small because many of the products that appear in (34) and (35) are already computed as part of  $\rho_i^{\text{FPT}}$  in (32). Meanwhile, we show in Appendix B.3 that the approximation for the FPT child estimator:

$$\bar{\theta}_{C_j}^{\text{FPT}} := \hat{\theta}_P + \frac{1}{n_{C_j}} \sum_{\{i:X_i \in C_j\}} \phi_i^{\text{FPT}},$$

is consistent for the original FPT child estimator  $\tilde{\theta}_{C_j}^{\text{FPT}}$  as  $\|\tilde{\theta}_{C_j}^{\text{FPT}} - \bar{\theta}_{C_j}^{\text{FPT}}\| = o_P(1)$ , meaning that this approximation does not alter the asymptotic behavior of our estimator. These accelerations are particularly compelling when the dimension of  $\theta^*(x)$  is large and computational efficiency is critical, as in large-scale A/B testing with multiple concurrent treatment arms or observational studies with numerous treatment levels (Kohavi et al., 2013; Bakshy et al., 2014).

## 6. Simulations

In this section, we perform empirical evaluations of the computational efficiency and estimation accuracy of the GRF-FPT method. We let GRF-FPT1 denote the FPT algorithm using the exact form of the FPT VCM/HTE pseudo-outcomes (32) and we let GRF-FPT2 denote the accelerated FPT algorithm based on the form of the FPT pseudo-outcome approximation (33) in Section 5. We compare both implementations relative to GRF-grad under VCM and HTE designs. Implementation details and links to the reproducible code are found in Appendix C.4.

**Settings.** We follow the structural model in (28). The auxiliary variables  $X_i$  are drawn from the Gaussian copula with latent covariance matrix  $\Sigma$ , where  $[\Sigma]_{j,k} = (0.3)^{|j-k|}$ . Supporting experiments for multicollinearity in  $X_i$  can be found in Appendix D.2. The outcomes  $Y_i$  follow (28) with Gaussian noise  $\epsilon_i \sim \mathcal{N}(0, 1)$ . For VCM experiments, regressors  $W_i \in \mathbb{R}^K$  are sampled from  $\mathcal{N}_K(\mathbf{0}, \mathbb{I})$ . For HTE experiments,  $W_i \in \{0, 1\}^K$  follows a multinomial distribution,  $W_i | X_i = x \sim \text{Multinomial}(1, (\pi_1(x), \dots, \pi_K(x)))$ , where  $\pi_k(x)$  is the probability of treatment level  $k \in \{1, \dots, K\}$ , characterizing a variety of different location-specific dependence structures through the setting of  $\pi_k(\cdot)$ . We set  $\nu^*(x) := 0$  and vary the target effect functions  $\theta_k^*(x)$  and treatment probabilities  $\pi_k(x)$  across different settings, fully detailed in Appendix C.4. Throughout our experiments we use subsampling ratio  $s/n = 0.5$ . Supporting experiments under different subsample ratios are found in Appendix D.2.

**Results.** The relative computational advantage of forests trained under GRF-FPT is displayed in Figure 3, while Figure 5 (in Appendix D.3) summarizes the absolute fit times across the three methods. These data show that the FPT mechanism is able to consistently offer a relative advantage, observing speedups of up to  $3.5\times$  faster than the gradient-based approach at the largest dimension  $K = 256$ . Figure 3 also shows increasing gains with increasing  $K$  and provides an empirical measurement of the theoretical scaling benefits discussed in Section 5. Moreover, the absolute fit times in Figure 5 (in Appendix D.3) illustrate that our method consistently remains faster than GRF-grad, with no clear computational or algorithmic bottleneck as a function of either  $n$  or  $K$ . Supporting experiments exploring the ef-

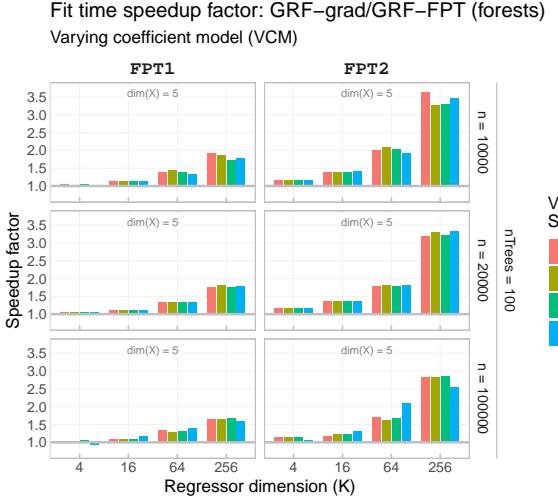


Figure 3: Speedup factor for GRF-FPT in comparison to GRF-grad for VCM timing experiments.

fects of sample sizes up to  $n = 500,000$  are presented in Appendix D.2, while Figures 7 and 8 (in Appendix D.3) show that even when  $n$  is small, GRF-FPT still observes a noticeable gain relative to GRF-grad. Additional timing benchmarks for VCM experiments and all HTE experiments are discussed in Appendix D.3.

To assess estimation accuracy, we evaluate the mean squared error (MSE) of  $\hat{\theta}(x)$  across 50 replications of the model and testing on a separate set of 5,000 observations. Figure 6 in Appendix D.3 confirms that GRF-FPT matches the accuracy of GRF-grad, while significantly reducing computation time. Further comparisons for both VCM and HTE settings are provided in Appendix D.3.

## 7. Real Data Application

**Data.** In this section we apply GRF-FPT to the analysis of geographically-varying effects  $\theta^*(x)$  on housing prices. The data, first appearing in Kelley Pace & Barry (1997), contains 20,640 observations of housing prices taken from the 1990 California census. Each observation corresponds to measurements aggregated over a small geographical census block, and contains measurements of 9 variables: median housing value, longitude, latitude, median housing age, total rooms, total bedrooms, population, households, and median income. We employ a VCM design of the form (28) where  $Y_i$  denotes the housing value,  $X_i$  denote the spatial coordinates, and  $W_i = (W_{i,1}, \dots, W_{i,6})^\top$  are the remaining six regressors. Details of the model and data transformations used for the California housing analysis is found in Appendix F.

**Results.** Table 7 summarizes the computational benefit of GRF-FPT applied to the California housing data. Figure 4 illustrates the six geographically-varying effect estimates

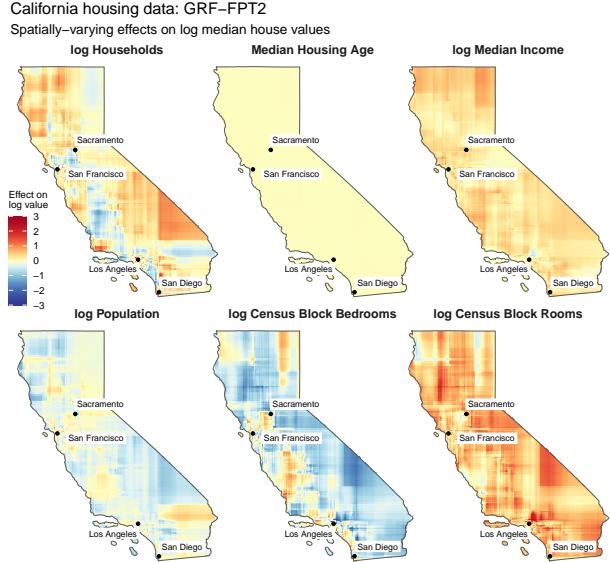


Figure 4: Geographically-varying GRF-FPT2 estimates  $\hat{\theta}(x)$ .

under GRF-FPT2, with qualitatively similar results shown in Figure 16 for GRF-FPT1 and GRF-grad in Appendix F. Figure 4 shows clearly the geographically-dependent relationship between different housing features and housing prices. In major urban centers such as LA, San Francisco, and Sacramento, housing prices tend to decrease with an increasing number of households, and may reflect overcrowding in densely populated areas. In contrast, rural regions show the opposite trend: prices rise slightly when rural areas have a larger number of housing units. This suggests that, in sparsely populated rural areas, a modest increase in households makes these places more attractive and livable. Median income, however, consistently shows a positive effect on prices across nearly all of California, while population size tends to show a negative effect, highlighting broader state-wide pressures on housing affordability.

## 8. Conclusion

Our results demonstrate that the FPT algorithm offers a substantial computational advantage over GRF-grad with comparable statistical accuracy, and highlights GRF-FPT as a powerful method for multi-dimensional estimation, particularly when estimates of the target function must be learned from the data rather than observed directly. Future work may explore extensions to larger-scale problems and alternative estimation tasks, as in unsupervised learning and structured prediction. Our findings position GRF-FPT as a scalable and robust alternative for practitioners seeking efficient localized estimation.

## Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## Acknowledgments

This work was supported by Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2024-06780) and FRQNT Team Research Project Grant (FRQ-NT 327788).

## References

- Amit, Y. and Geman, D. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, 1997.
- Angrist, J. D. and Pischke, J.-S. *Mostly harmless econometrics: An empiricist’s companion*. Princeton university press, 2009.
- Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.
- Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, 2019. doi: 10.1214/18-AOS1709. URL <https://doi.org/10.1214/18-AOS1709>.
- Bakshy, E., Eckles, D., and Bernstein, M. S. Designing and deploying online field experiments. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW ’14, pp. 283–292, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327442. doi: 10.1145/2566486.2567967. URL <https://doi.org/10.1145/2566486.2567967>.
- Banach, S. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3:133–181, 1922.
- Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.
- Biau, G. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.
- Biau, G., Devroye, L., and Lugosi, G. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(66):2015–2033, 2008.
- URL <http://jmlr.org/papers/v9/biau08a.html>.
- Breiman, L. Bagging predictors. *Machine Learning*, 24: 123–140, 1996.
- Breiman, L. Random forests. *Machine Learning*, 45:5–32, 2001.
- Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*. CRC, 1984. ISBN 9780412048418.
- Cevid, D., Michel, L., Näf, J., Bühlmann, P., and Meinshausen, N. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022. URL <http://jmlr.org/papers/v23/21-0585.html>.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL <https://doi.org/10.1111/ectj.12097>.
- Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., and Zhu, R. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):179–211, 02 2023. ISSN 1369-7412. doi: 10.1093/rssb/qkac001. URL <https://doi.org/10.1093/rssb/qkac001>.
- De’ath, G. Multivariate regression trees: a new technique for modeling species–environment relationships. *Ecology*, 83(4):1105–1117, 2002.
- Denil, M., Matheson, D., and De Freitas, N. Narrowing the gap: Random forests in theory and in practice. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 665–673, Bejing, China, 22–24 Jun 2014. PMLR. URL <https://proceedings.mlr.press/v32/denil14.html>.
- Dietterich, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40:139–157, 2000.
- Fan, J. and Gijbels, I. *Local Polynomial Modelling and Its Applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, London, 1996. doi: 10.1201/9780203748725. URL <https://www.taylorfrancis.com/books/mono/10.1201/9780203748725>.

- Fan, J., Heckman, N. E., and Wand, M. P. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90(429):141–150, 1995.
- Friedberg, R., Tibshirani, J., Athey, S., and Wager, S. Local linear forests. *Journal of Computational and Graphical Statistics*, 30(2):503–517, 2020.
- Friedman, J. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pp. 1189–1232, 2001.
- Hastie, T. and Tibshirani, R. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):757–796, 1993. ISSN 00359246. URL <http://www.jstor.org/stable/2345993>.
- Imai, K. and Ratkovic, M. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443 – 470, 2013. doi: 10.1214/12-AOAS593. URL <https://doi.org/10.1214/12-AOAS593>.
- Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029. PMLR, 2016.
- Kelley Pace, R. and Barry, R. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. ISSN 0167-7152. doi: [https://doi.org/10.1016/S0167-7152\(96\)00140-X](https://doi.org/10.1016/S0167-7152(96)00140-X). URL <https://www.sciencedirect.com/science/article/pii/S016771529600140X>.
- Knaus, M. C., Lechner, M., and Strittmatter, A. Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, 24(1):134–161, 2021.
- Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., and Pohlmann, N. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’13, pp. 1168–1176, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450321747. doi: 10.1145/2487575.2488217. URL <https://doi.org/10.1145/2487575.2488217>.
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.
- Lee, Y., Veerubhotla, K., Jeong, M. H., and Lee, C. H. Deep learning in personalization of cardiovascular stents. *Journal of Cardiovascular Pharmacology and Therapeutics*, 25(2):110–120, 2020.
- Lewbel, A. A local generalized method of moments estimator. *Economics Letters*, 94(1):124–128, 2007.
- Lindelöf, E. Sur l’application de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre. *Comptes Rendus Hebdomadaires des Séances de l’Académie des Sciences*, 116:454–457, 1894.
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.
- Newey, W. K. Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, 10(2): 1–21, 1994.
- Neyman, J. Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51, 1923. Reprinted and translated in Neyman, J. (1990). *Statistical Science*, 5(4), 463–480.
- Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.
- Picard, É. Mémoire sur la théorie des équations aux dérivées partielles et la méthode des approximations successives. *Journal de Mathématiques Pures et Appliquées*, 6:145–210, 1890.
- Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 37(11):1767–1787, 2018.
- Robins, J. M. and Ritov, Y. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semiparametric models. *Statistics in Medicine*, 16(1-3):285–319, 1997. doi: 10.1002/(SICI)1097-0258(19970215)16:3<285::AID-SIM535>3.0.CO;2-#. URL <https://pubmed.ncbi.nlm.nih.gov/9004398/>.
- Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pp. 931–954, 1988.
- Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

- Ryu, E. K. and Boyd, S. A primer on monotone operator methods (survey). *Applied and Computational Mathematics*, 15(1):3–43, 2016. Survey article.
- Scornet, E., Biau, G., and Vert, J.-P. Consistency of random forests. *The Annals of Statistics*, 43(4):1716 – 1741, 2015. doi: 10.1214/15-AOS1321. URL <https://doi.org/10.1214/15-AOS1321>.
- Segal, M. R. Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, 87(418): 407–418, 1992.
- Severini, T. A. and Staniswalis, J. G. Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, 89(426):501–511, 1994.
- Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/shalit17a.html>.
- Speckman, P. Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50(3):413–436, 1988. ISSN 00359246. URL <http://www.jstor.org/stable/2345705>.
- Staniswalis, J. G. The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84(405):276–283, 1989.
- Tibshirani, J., Athey, S., Sverdrup, E., and Wager, S. *grf: Generalized Random Forests*, 2024. URL <https://github.com/grf-labs/grf>. R package version 2.4.0.
- Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.
- Wager, S. and Walther, G. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.
- Wager, S., Hastie, T., and Efron, B. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(1):1625–1651, 2014. URL <https://jmlr.org/papers/volume15/wager14a/wager14a.pdf>.
- Yang, Y., Gu, Y., Zhao, Y., and Fan, J. Flexible regularized estimating equations: Some new perspectives. 2021. URL <https://arxiv.org/abs/2110.11074>.
- Zeileis, A. and Hornik, K. Generalized m-fluctuation tests for parameter instability. *Statistica Neerlandica*, 61(4): 488–508, 2007.
- Zeileis, A., Hothorn, T., and Hornik, K. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514, 2008.

## A. Technical Preliminaries

### A.1. Assumptions

We follow the key assumptions of Athey et al. (2019) made for the theoretical analyses of GRF. The predictor and parameter spaces are both subsets of Euclidean space such that  $x \in \mathcal{X} = [0, 1]^p$  and  $(\theta, \nu) \in \mathcal{B} \subset \mathbb{R}^K$ , where  $\mathcal{B}$  is a compact subset of  $\mathbb{R}^K$ . Under the analyses of Wager & Walther (2015), we suppose that the features of the auxiliary covariates  $X_i = (X_{i,1}, \dots, X_{i,p})^\top$  have density  $f_X$  that is bounded away from 0 and  $\infty$ , i.e.  $c \leq f_X(x) \leq C < \infty$ , for some constants  $c > 0$  and  $C < \infty$ . GRF does not require that the score function  $\psi$  is continuous in  $(\theta, \nu)$ , as is the case for quantile estimation, one does require that the expected score/moment function

$$M_{\theta, \nu}(x) := \mathbb{E}_{O|X} [\psi_{\theta, \nu}(O) | X = x], \quad (36)$$

is smoothly varying in its parameters  $(\theta, \nu)$ .

ASSUMPTION 1. For fixed  $(\theta, \nu)$ , the  $M$ -function (36) is Lipschitz continuous in  $x$ .

ASSUMPTION 2. For fixed  $x$ , the  $M$ -function is twice-differentiable in  $(\theta, \nu)$  with uniformly bounded second derivative,

$$\left\| \nabla_{(\theta, \nu)}^2 M_{\theta, \nu}(x) \right\| < \infty,$$

where  $\|\cdot\|$  denotes the appropriate tensor norm for the second derivative of  $M_{\theta, \nu}$  taken with respect to  $(\theta, \nu)$ . Let  $V(x) := \nabla_{(\theta, \nu)} M_{\theta, \nu}(x)|_{\theta=\theta^*(x), \nu=\nu^*(x)}$  denote the population Jacobian at the true  $(\theta^*(x), \nu^*(x))$ , and assume that  $V(x)$  is invertible for all  $x \in \mathcal{X}$ . We write  $V(x)$  in block form as

$$V(x) = \begin{bmatrix} V_{\theta\theta}(x) & V_{\theta\nu}(x) \\ V_{\nu\theta}(x) & V_{\nu\nu}(x) \end{bmatrix}. \quad (37)$$

ASSUMPTION 3. The score functions  $\psi_{\theta, \nu}(O_i)$  have a continuous covariance structure in the following sense: Let  $\gamma(\cdot, \cdot)$  denote the worst-case variogram:

$$\gamma \left( \begin{bmatrix} \theta_1 \\ \nu_1 \end{bmatrix}, \begin{bmatrix} \theta_2 \\ \nu_2 \end{bmatrix} \right) := \sup_{x \in \mathcal{X}} \left\{ \left\| \text{Var}_{O|X} (\psi_{\theta_1, \nu_1}(O_i) - \psi_{\theta_2, \nu_2}(O_i) | X_i = x) \right\|_F \right\},$$

then, for some  $L > 0$ ,

$$\gamma \left( \begin{bmatrix} \theta_1 \\ \nu_1 \end{bmatrix}, \begin{bmatrix} \theta_2 \\ \nu_2 \end{bmatrix} \right) \leq L \left\| \begin{bmatrix} \theta_1 \\ \nu_1 \end{bmatrix} - \begin{bmatrix} \theta_2 \\ \nu_2 \end{bmatrix} \right\|_2, \quad \text{for all } (\theta_1, \nu_1), (\theta_2, \nu_2).$$

ASSUMPTION 4. The score function  $\psi_{\theta, \nu}(O_i)$  can be written as

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

where  $\lambda$  is Lipschitz-continuous in  $(\theta, \nu)$ ,  $g : \{O_i\} \rightarrow \mathbb{R}$  a univariate summary of the observables  $O_i$ , and  $\zeta_{\theta} : \mathbb{R} \rightarrow \mathbb{R}$  any family of monotone and bounded functions.

ASSUMPTION 5. For any weights  $\alpha_i$  with  $\sum \alpha_i = 1$ , the minimizer  $(\hat{\theta}, \hat{\nu})$  of the weighted empirical estimation problem (5) satisfies:

$$\left\| \sum_{i=1}^n \alpha_i \psi_{\hat{\theta}, \hat{\nu}}(O_i) \right\|_2 \leq C \max_{1 \leq i \leq n} \{\alpha_i\}, \quad \text{for } C \geq 0.$$

ASSUMPTION 6. The score function  $\psi_{\theta, \nu}(O_i)$  is a negative subgradient of a convex function, and the moment function  $M_{\theta, \nu}(X_i)$  is the negative gradient of a strongly convex function.

## A.2. Forest specifications

The consistency and asymptotic normality results, Theorems 4.3 and 4.4, require that the forest trained following Algorithm 2 consists of trees that satisfy a certain set of specifications. These forest specifications are precisely those imposed by Athey et al. (2019) for forests of gradient-based trees, and collectively, these specifications describe fairly mild conditions on the tree splitting mechanism, as well as specific requirements for the sampling procedure.

**SPECIFICATION 1. (*Symmetric*)** Tree estimates are invariant to permutations of the training indices. In other words, the output of a tree does not depend on the order in which the training samples are indexed.

**SPECIFICATION 2. (*Balanced/* $\omega$ *-regular*)** The proportion of parent observations assigned into either child is bound below by some  $\omega > 0$ , i.e.  $n_{C_j} \geq \omega n_P$ .

**SPECIFICATION 3. (*Randomized/random-split*)** The probability of splitting along any feature/dimension of the input space is bound below by some  $\pi > 0$ .

**SPECIFICATION 4. (*Subsampling*)** Trees are trained on subsample of size  $s$ , drawn without replacement from  $n$  training samples, where  $s/n \rightarrow 0$  as  $s \rightarrow \infty$ .

**SPECIFICATION 5. (*Honesty*)** Trees are trained using the sample splitting procedure described in Appendix C.1.

## A.3. Regularity conditions

**REGULARITY CONDITION 1.** Let  $V(x)$  be as defined in Assumption 2 and let  $\rho_i^*(x)$  denote the influence function of the  $i$ -th observation with respect to the target  $\theta^*(x)$ :

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

Then,

$$\text{Var}(\rho_i^*(x) \mid X_i = x) > 0, \quad \text{for all } x \in \mathcal{X}.$$

**REGULARITY CONDITION 2.** Trees are grown on subsamples of size  $s$  scaling as  $s = n^\beta$ , for some subsample scaling exponent  $\beta$  bound according to  $\beta_{\min} < \beta < 1$ , such that

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

where  $0 < \pi, \omega < 1$  are constants defined in forest Specifications 2 and 3.

## A.4. Neyman orthogonality

To identify the underlying local parameters  $(\theta^*(x), \nu^*(x)) \in \mathbb{R}^K$  one must have a score  $\psi_{\theta, \nu}(O)$  with at least  $K = K_\theta + K_\nu$  components, where here we use  $K_\theta$  and  $K_\nu$  to denote the dimensions of the component subvectors  $\theta^*(x) \in \mathbb{R}^{K_\theta}$  and  $\nu^*(x) \in \mathbb{R}^{K_\nu}$ . Conceptually, a score  $\psi_{\theta, \nu}(O)$  can be partitioned into the components that identify the  $\theta$ -coordinates, denoted by  $\psi_1$ , and those that identify the  $\nu$ -coordinates, denoted by  $\psi_2$ , and thus the moment functions  $M_{\theta, \nu}(x)$  in (36) can also be partitioned the same way:

$$\psi_{\theta, \nu}(O) = \begin{bmatrix} \psi_1(\theta, \nu; O) \\ \psi_2(\theta, \nu; O) \end{bmatrix}, \quad M_{\theta, \nu}(x) = \begin{bmatrix} M_1(\theta, \nu; x) \\ M_2(\theta, \nu; x) \end{bmatrix} = \begin{bmatrix} \mathbb{E}[\psi_1(\theta, \nu; O) \mid X = x] \\ \mathbb{E}[\psi_2(\theta, \nu; O) \mid X = x] \end{bmatrix}.$$

The corresponding Jacobian matrix of  $M_{\theta, \nu}(x)$  taken with respect to  $(\theta, \nu)$  and evaluated at the truth  $(\theta^*(x), \nu^*(x))$  is

$$V(x) = \nabla_{(\theta, \nu)} M(\theta, \nu; x) \Big|_{\theta=\theta^*(x), \nu=\nu^*(x)} = \begin{bmatrix} V_{\theta\theta}(x) & V_{\theta\nu}(x) \\ V_{\nu\theta}(x) & V_{\nu\nu}(x) \end{bmatrix},$$

where here the subscripts in the block expressions of  $V(x)$  indicate the coordinates with which the gradient is taken, and in all cases are evaluated at the truth  $(\theta^*(x), \nu^*(x))$ :

$$\begin{aligned} V_{\theta\theta}(x) &= \nabla_\theta M_1(\theta, \nu; x)|_{\theta=\theta^*(x), \nu=\nu^*(x)}, \\ V_{\theta\nu}(x) &= \nabla_\nu M_1(\theta, \nu; x)|_{\theta=\theta^*(x), \nu=\nu^*(x)}, \\ V_{\nu\theta}(x) &= \nabla_\theta M_2(\theta, \nu; x)|_{\theta=\theta^*(x), \nu=\nu^*(x)}, \\ V_{\nu\nu}(x) &= \nabla_\nu M_2(\theta, \nu; x)|_{\theta=\theta^*(x), \nu=\nu^*(x)}. \end{aligned}$$

In this context, the assumption of Neyman orthogonal moment conditions is more completely labeled as Neyman orthogonality for the estimation of  $\theta^*(x)$  with respect to the nuisance  $\nu^*(x)$ , and can be summarized as an assumption that the moment conditions for  $\theta^*(x)$  are insensitive to first-order changes in  $\nu$  around the truth  $\nu^*(x)$  whenever  $\theta = \theta^*(x)$ . For GRF, this means that one assumes (1) satisfies  $M_1(\theta^*(x), \nu^*(x); x) = \mathbf{0}$ , and in other words, the partial derivatives of the moment functions for  $\theta^*(x)$  with respect to  $\nu$  are zero at  $(\theta^*(x), \nu^*(x))$ :

$$V_{\theta\nu}(x) = \mathbf{0}.$$

### A.5. Example: Neyman orthogonality for VCM and HTE

Consider the VCM/HTE model with data  $(Y_i, W_i, X_i)$  related according to

$$\mathbb{E}[Y_i | X_i = x] = \nu^*(x) + W_i^\top \theta^*(x),$$

such that, as discussed in Section 3.2, the score function  $\psi_{\theta, \nu}$  that identifies the underlying  $(\theta^*(x), \nu^*(x))$  is

$$\psi_{\theta, \nu}(Y_i, W_i) := \begin{bmatrix} (Y_i - W_i^\top \theta - \nu)W_i \\ Y_i - W_i^\top \theta - \nu \end{bmatrix},$$

and the corresponding local Jacobian  $V(x)$  has block form

$$V(x) = -\mathbb{E} \left[ \begin{bmatrix} W_i W_i^\top & W_i^\top \\ W_i & 1 \end{bmatrix} \middle| X_i = x \right] = - \begin{bmatrix} \mathbb{E}[W_i W_i^\top | X_i = x] & \mathbb{E}[W_i^\top | X_i = x] \\ \mathbb{E}[W_i | X_i = x] & 1 \end{bmatrix}.$$

Therefore, for Neyman orthogonality to hold one requires that  $\mathbb{E}[W_i | X_i = x] = \mathbf{0}$ .

## B. Derivations and Proofs

### B.1. Proofs for Section 3.4

#### B.1.1. MULTIVARIATE CART CRITERIA

Let  $\rho_i \in \mathbb{R}^K$  be vector-valued responses associated with covariates  $X_i \in P$ . A standard CART split  $(C_1, C_2)$  of  $P$  minimizes the conventional least-squares criterion:

$$\sum_{\{i : X_i \in C_1\}} \|\rho_i - \bar{\rho}_{C_1}\|^2 + \sum_{\{i : X_i \in C_2\}} \|\rho_i - \bar{\rho}_{C_2}\|^2, \quad (38)$$

where  $\bar{\rho}_{C_j} := n_{C_j}^{-1} \sum_{\{i : X_i \in C_j\}} \rho_i$  is the local prediction over child node  $C_j$ . We verify that a split  $(C_1, C_2)$  minimizes (38) if and only if it maximizes

$$n_{C_1} \|\bar{\rho}_{C_1}\|^2 + n_{C_2} \|\bar{\rho}_{C_2}\|^2. \quad (39)$$

*Proof.* Each sum in (38) can be expanded as

$$\begin{aligned} \sum_{\{i : X_i \in C_j\}} \|\rho_i - \bar{\rho}_{C_j}\|^2 &= \sum_{\{i : X_i \in P\}} \|\rho_i - \bar{\rho}_{C_j}\|^2 \cdot \mathbf{1}(X_i \in C_j), \\ &= \sum_{\{i : X_i \in P\}} \left( \|\rho_i\|^2 - 2\rho_i^\top \bar{\rho}_{C_j} + \|\bar{\rho}_{C_j}\|^2 \right) \cdot \mathbf{1}(X_i \in C_j), \\ &= \sum_{\{i : X_i \in P\}} \|\rho_i\|^2 \cdot \mathbf{1}(X_i \in C_j) - n_{C_j} \|\bar{\rho}_{C_j}\|^2. \end{aligned}$$

Therefore, the least-squares criterion CART (38) is equivalently written as

$$\begin{aligned}
 \sum_{j=1,2} \sum_{\{i : X_i \in C_j\}} \|\rho_i - \bar{\rho}_{C_j}\|^2 &= \sum_{j=1,2} \left( \sum_{\{i : X_i \in P\}} \|\rho_i\|^2 \cdot \mathbb{1}(X_i \in C_j) - n_{C_j} \|\bar{\rho}_{C_j}\|^2 \right), \\
 &= \sum_{j=1,2} \left( \sum_{\{i : X_i \in P\}} \|\rho_i\|^2 \cdot \mathbb{1}(X_i \in C_j) \right) - \left( n_{C_1} \|\bar{\rho}_{C_1}\|^2 + n_{C_2} \|\bar{\rho}_{C_2}\|^2 \right), \\
 &= \sum_{\{i : X_i \in P\}} \|\rho_i\|^2 - \left( n_{C_1} \|\bar{\rho}_{C_1}\|^2 + n_{C_2} \|\bar{\rho}_{C_2}\|^2 \right).
 \end{aligned}$$

The first term does not depend on the choice of split, and therefore the split that minimizes (38) is equivalent to the split that maximizes (39).  $\blacksquare$

### B.1.2. SPLITS VIA CART ON PSEUDO-OUTCOMES

The following result is a generalization to the claim made in Section 3.4 that a CART split on pseudo-outcomes  $\rho_i^{\text{FPT}}$  will produce a split that maximizes the  $\tilde{\Delta}^{\text{FPT}}$ -criterion, and is sufficiently general to cover gradient-based pseudo-outcomes  $\rho_i^{\text{grad}}$  and the corresponding  $\tilde{\Delta}^{\text{grad}}$ -criterion.

**Lemma B.1.** *Suppose we can write*

$$\tilde{\theta}_{C_j} = a + \frac{1}{n_{C_j}} \sum_{\{i : X_i \in C_j\}} \rho_i, \quad \rho_i = -B\psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i), \quad (40)$$

where  $a$  and  $B$  denote appropriately sized vectors and matrices whose values do not depend on the candidate child node  $C_j$ . Under Assumptions A.1, the split  $(C_1, C_2)$  that maximizes

$$\tilde{\Delta}(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P^2} \left\| \tilde{\theta}_{C_1} - \tilde{\theta}_{C_2} \right\|^2,$$

is exactly the split chosen by CART for vector-valued responses  $\rho_i$  fit over covariates  $X_i \in P$ .

**Proof of Lemma B.1.** The scores  $\psi_{\theta, \nu}(O_i)$  satisfy subgradient conditions by Assumption 6, and therefore the parent solutions  $(\hat{\theta}_P, \hat{\nu}_P)$  satisfy the first-order conditions

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

Hence,

$$\begin{aligned}
 \mathbf{0} &= \sum_{\{i : X_i \in P\}} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) = \sum_{\{i : X_i \in C_1\}} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) + \sum_{\{i : X_i \in C_2\}} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i), \\
 &= -B \left( \sum_{\{i : X_i \in C_1\}} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) + \sum_{\{i : X_i \in C_2\}} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i) \right), \\
 &= \sum_{\{i : X_i \in C_1\}} \rho_i + \sum_{\{i : X_i \in C_2\}} \rho_i.
 \end{aligned}$$

Each sum in the previous expression is equivalently written as  $\sum \rho_i = n_{C_j}(\tilde{\theta}_{C_j} - a)$ . Hence,

$$\begin{aligned}
 \mathbf{0} &= \sum_{\{i : X_i \in C_1\}} \rho_i + \sum_{\{i : X_i \in C_2\}} \rho_i, \\
 &= n_{C_1}(\tilde{\theta}_{C_1} - a) + n_{C_2}(\tilde{\theta}_{C_2} - a), \\
 \iff a &= \frac{n_{C_1}}{n_P} \tilde{\theta}_{C_1} + \frac{n_{C_2}}{n_P} \tilde{\theta}_{C_2}.
 \end{aligned}$$

Writing  $\bar{\rho}_{C_j} := \frac{1}{n_{C_j}} \sum_{\{i: X_i \in C_j\}} \rho_i$ , one has:

$$\begin{aligned}\bar{\rho}_{C_1} &= \tilde{\theta}_{C_1} - a, \\ &= \tilde{\theta}_{C_1} - \frac{n_{C_1}}{n_P} \tilde{\theta}_{C_1} - \frac{n_{C_2}}{n_P} \tilde{\theta}_{C_2}, \\ &= \frac{n_{C_2}}{n_P} (\tilde{\theta}_{C_1} - \tilde{\theta}_{C_2}),\end{aligned}$$

and

$$\frac{n_{C_1}}{n_P} \|\bar{\rho}_{C_1}\|^2 = \frac{n_{C_1} n_{C_2}^2}{n_P^3} \left\| \tilde{\theta}_{C_1} - \tilde{\theta}_{C_2} \right\|^2.$$

Applying analogous arguments with respect to  $C_2$ , one has the symmetric result:

$$\frac{n_{C_2}}{n_P} \|\bar{\rho}_{C_2}\|^2 = \frac{n_{C_2} n_{C_1}^2}{n_P^3} \left\| \tilde{\theta}_{C_1} - \tilde{\theta}_{C_2} \right\|^2.$$

Therefore,

$$\begin{aligned}\frac{1}{n_P} (n_{C_1} \|\bar{\rho}_{C_1}\|^2 + n_{C_2} \|\bar{\rho}_{C_2}\|^2) &= \frac{n_{C_1} n_{C_2}^2}{n_P^3} \left\| \tilde{\theta}_{C_1} - \tilde{\theta}_{C_2} \right\|^2 + \frac{n_{C_2} n_{C_1}^2}{n_P^3} \left\| \tilde{\theta}_{C_1} - \tilde{\theta}_{C_2} \right\|^2, \\ &= \frac{n_{C_1} n_{C_2}}{n_P^2} \left\| \tilde{\theta}_{C_1} - \tilde{\theta}_{C_2} \right\|^2, \\ &= \tilde{\Delta}(C_1, C_2).\end{aligned}$$

Based on the arguments in Appendix B.1.1, a split  $(C_1, C_2)$  maximizes  $n_{C_1} \|\bar{\rho}_{C_1}\|^2 + n_{C_2} \|\bar{\rho}_{C_2}\|^2$  if and only if it is a CART split performed on the  $\rho_i$  over  $P$ . That is,  $\tilde{\Delta}(C_1, C_2)$  is precisely maximized by a single CART split on  $\rho_i = -B\psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i)$  fit over covariates  $X_i \in P$ , as desired. ■

### B.1.3. SCALE INVARIANCE OF CART SPLITS

**Lemma B.2** (Argmax equivalence of FPT criteria). *The optimal split identified by CART on pseudo-outcomes  $\rho_i^{FPT}$  of the form (23) does not depend on the scale factor  $\eta$ , for any  $\eta \neq 0$ .*

**Proof of Lemma B.2.** Denote by  $\rho_i^{(\eta)}$  FPT pseudo-outcomes based on an arbitrary scale factor  $\eta \neq 0$  of the form (23):

$$\rho_i^{(\eta)} := -\eta \xi^\top \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i), \quad (41)$$

and let  $\bar{\psi}_{C_j}$  denote the child-leaf average score evaluated at the parent solution  $(\hat{\theta}_P, \hat{\nu}_P)$ :

$$\bar{\psi}_{C_j} := \frac{1}{n_{C_j}} \sum_{\{i: X_i \in C_j\}} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i),$$

such that the corresponding child-leaf pseudo-outcome averages  $\bar{\rho}_{C_j}^{(\eta)} := \frac{1}{n_{C_j}} \sum_{\{i: X_i \in C_j\}} \rho_i^{(\eta)}$  are equivalently written as

$$\bar{\rho}_{C_j}^{(\eta)} = -\eta \xi^\top \bar{\psi}_{C_j}$$

Let  $\tilde{\Delta}_\eta^{FPT}(C_1, C_2)$  denote the FPT criterion of the form (25) based on pseudo-outcomes (41):

$$\tilde{\Delta}_\eta^{FPT}(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P^2} \left\| \bar{\rho}_{C_1}^{(\eta)} - \bar{\rho}_{C_2}^{(\eta)} \right\|^2 = \frac{n_{C_1} n_{C_2}}{n_P^2} \left\| \eta \xi^\top (\bar{\psi}_{C_1} - \bar{\psi}_{C_2}) \right\|^2.$$

One has:

$$\left\| \eta \xi^\top (\bar{\psi}_{C_1} - \bar{\psi}_{C_2}) \right\|^2 = \eta^2 \left\| \xi^\top (\bar{\psi}_{C_1} - \bar{\psi}_{C_2}) \right\|^2,$$

and hence the  $\tilde{\Delta}_\eta^{\text{FPT}}$ -criteria obey the scaling relation:

$$\tilde{\Delta}_\eta^{\text{FPT}}(C_1, C_2) = \eta^2 \cdot \tilde{\Delta}_1^{\text{FPT}}(C_1, C_2), \quad (42)$$

where  $\tilde{\Delta}_1^{\text{FPT}}$  denotes the FPT criterion induced by pseudo-outcomes  $\rho_i^{(1)}$  based on unit scale factor  $\eta = 1$ . The relation (42) implies that any nonzero split-independent rescaling  $\rho_i^{(\eta)} = \eta \rho_i^{(1)}$  will induce a splitting criterion  $\tilde{\Delta}_\eta^{\text{FPT}}(C_1, C_2)$  with the same maximizer as  $\tilde{\Delta}_1^{\text{FPT}}(C_1, C_2)$ :

$$\arg \max_{(C_1, C_2)} \left\{ \tilde{\Delta}_\eta^{\text{FPT}}(C_1, C_2) \right\} = \arg \max_{(C_1, C_2)} \left\{ \eta^2 \cdot \tilde{\Delta}_1^{\text{FPT}}(C_1, C_2) \right\} = \arg \max_{(C_1, C_2)} \left\{ \tilde{\Delta}_1^{\text{FPT}}(C_1, C_2) \right\}.$$

Intuitively, a CART split is chosen by ranking the criterion values among the candidate splits and selecting the maximizing split  $(C_1, C_2)$ . Therefore, the FPT splitting mechanism is unaffected by the scale factor  $\eta$  used to specify fixed-point pseudo-outcomes (23). The absolute scale of the  $\tilde{\Delta}^{\text{FPT}}$ -criterion does not matter when searching for the optimal split, and only the criterion rankings across the candidate splits determine the final partition. ■

## B.2. Proofs for Section 4

### Notation and definitions.

- Let  $o_P(a, b, c) := o_P(\max\{a, b, c\})$ , with an analogous abbreviation for  $O_P(\cdot)$ .
- For a fixed parent node  $P$ , denote by  $x_P$  the center of mass of the  $X_i \in P$ , and let  $r := \sup_{\{i: X_i \in P\}} \|X_i - x_P\|$  denote the radius of the parent  $P$ . Throughout, we consider an asymptotic regime where  $n_{C_j} \rightarrow \infty$  and  $r \rightarrow 0$ , corresponding to leaves over  $\mathcal{X}$  of vanishing radius. Further,  $r$  and  $n_{C_j}$  are related under the conditions of GRF Proposition 1, namely,  $r^{-2} \ll n_{C_j}$  and hence  $n_{C_j} r^2 \rightarrow \infty$  and  $1/\sqrt{n_{C_j}} = o(r)$ .
- Let  $\theta_{C_j}^*$  denote the true parameter expectation over the child node:

$$\theta_{C_j}^* := \mathbb{E}[\theta^*(X) \mid X \in C_j], \quad j = 1, 2, \quad (43)$$

and let  $\tilde{\theta}_{C_j}^*(x_P)$  denote an oracle version of the gradient-based leaf statistic:

$$\tilde{\theta}_{C_j}^*(x_P) := \theta^*(x_P) - \frac{1}{n_{C_j}} \sum_{\{i: X_i \in C_j\}} \xi^\top V(x_P)^{-1} \psi_{\theta^*(x_P), \nu^*(x_P)}(O_i),$$

where  $V(x)$  is the underlying local Jacobian in Assumption 2. Equivalently, in terms of the oracle pseudo-outcome/influence function  $\rho_i^*(\cdot)$  defined in Regularity Condition 1,

$$\tilde{\theta}_{C_j}^*(x_P) := \theta^*(x_P) + \frac{1}{n_{C_j}} \sum_{\{i: X_i \in C_j\}} \rho_i^*(x_P).$$

The following are technical lemmas used for the proof of Proposition 4.1.

**Lemma B.3.** Suppose Assumptions A.1 and Specifications A.2 hold. Then,

$$\Delta(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P^2} \|\theta_{C_1}^* - \theta_{C_2}^*\|^2 + o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right).$$

**Proof of Lemma B.3.** Write the difference  $\hat{\theta}_{C_j} - \theta_{C_j}^*$  as

$$\hat{\theta}_{C_j} - \theta_{C_j}^* = \underbrace{\left( \hat{\theta}_{C_j} - \tilde{\theta}_{C_j}^*(x_P) \right)}_{T_1} + \underbrace{\left( \tilde{\theta}_{C_j}^*(x_P) - \mathbb{E}[\tilde{\theta}_{C_j}^*(x_P) \mid X \in C_j] \right)}_{T_2} + \underbrace{\left( \mathbb{E}[\tilde{\theta}_{C_j}^*(x_P) \mid X \in C_j] - \theta_{C_j}^* \right)}_{T_3}.$$

Under standard LLN arguments, the second term satisfies  $T_2 = O_P(1/\sqrt{n_{C_j}})$ , and in an asymptotic regime with  $r^{-2} \ll n_{C_j}$  one has  $T_2 = o_P(r)$ . Meanwhile, the first and third terms appear in the proofs of Propositions 2 and 1 of Athey et al. (2019), respectively, and satisfy  $T_1 = o_P(r, 1/\sqrt{n_{C_j}})$  and  $T_3 = \mathcal{O}(r^2) \implies T_3 = o(r)$ . It follows

$$\hat{\theta}_{C_j} - \theta_{C_j}^* = o_P(r, 1/\sqrt{n_{C_j}}),$$

and in particular

$$\hat{\theta}_{C_1} - \hat{\theta}_{C_2} = \theta_{C_1}^* - \theta_{C_2}^* + o_P\left(r, \frac{1}{\sqrt{n_{C_1}}}, \frac{1}{\sqrt{n_{C_2}}}\right).$$

Write  $A = \theta_{C_1}^* - \theta_{C_2}^*$  and let  $E$  be any term satisfying  $E = o_P(r, 1/\sqrt{n_{C_1}}, 1/\sqrt{n_{C_2}})$  such that  $\Delta(C_1, C_2)$  is equivalently written  $\Delta(C_1, C_2) = (n_{C_1} n_{C_2}/n_P^2) \cdot \|A + E\|^2$ . Consider the difference

$$\begin{aligned} \Delta(C_1, C_2) - \frac{n_{C_1} n_{C_2}}{n_P^2} \|\theta_{C_1}^* - \theta_{C_2}^*\|^2 &= \frac{n_{C_1} n_{C_2}}{n_P^2} \left( \|A + E\|^2 - \|A\|^2 \right), \\ &= \frac{n_{C_1} n_{C_2}}{n_P^2} \left( 2\langle A, E \rangle + \|E\|^2 \right). \end{aligned}$$

Under Specification 2 there exists a fixed proportion  $\omega > 0$  such that  $n_{C_1}, n_{C_2} \geq \omega n_P$ , and hence  $n_{C_1} n_{C_2}/n_P^2 \geq \omega(1 - \omega)$  and also  $n_{C_1} n_{C_2}/n_P^2 \leq 1/4$  for all  $n_{C_1} + n_{C_2} = n_P$ . Therefore  $n_{C_1} n_{C_2}/n_P^2 = \mathcal{O}(1)$ . Meanwhile,  $\|E\|^2 = o_P(r^2, 1/n_{C_1}, 1/n_{C_2})$  is true by definition of  $E$ , and under our assumptions one may follow the arguments of Athey et al. (2019) Proposition 1 to see that  $A = \theta_{C_1}^* - \theta_{C_2}^* = \mathcal{O}(r)$ . Thus,

$$\langle A, E \rangle = \mathcal{O}(r) \cdot o_P\left(r, \frac{1}{\sqrt{n_{C_1}}}, \frac{1}{\sqrt{n_{C_2}}}\right) = o_P\left(r^2, \frac{r}{\sqrt{n_{C_1}}}, \frac{r}{\sqrt{n_{C_2}}}\right),$$

and therefore

$$\Delta(C_1, C_2) - \frac{n_{C_1} n_{C_2}}{n_P^2} \|\theta_{C_1}^* - \theta_{C_2}^*\|^2 = o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right),$$

as desired.  $\blacksquare$

**Lemma B.4.** Suppose the conditions of Lemma B.3 hold, and assume moreover Neyman orthogonal moment conditions such that the underlying Jacobian  $V(x)$  defined in Assumption 2 with block form (37). Then,

$$\tilde{\Delta}_\eta^{\text{FPT}}(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P^2} \eta^2 \|V_{\theta\theta}(x_P)(\theta_{C_1}^* - \theta_{C_2}^*)\|^2 + o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right),$$

where  $\tilde{\Delta}_\eta^{\text{FPT}}$  defined in Lemma B.2 denotes the FPT criterion with arbitrary scale factor  $\eta \neq 0$ .

**Proof of Lemma B.4.** From the proof of Lemma B.2 one finds that  $\tilde{\Delta}_\eta^{\text{FPT}}(C_1, C_2)$  is equivalently written

$$\Delta_\eta^{\text{FPT}}(C_1, C_2) := \frac{n_{C_1} n_{C_2}}{n_P^2} \eta^2 \|\xi^\top (\bar{\psi}_{C_1} - \bar{\psi}_{C_2})\|^2, \quad \bar{\psi}_{C_j} := \frac{1}{n_{C_1}} \sum_{\{i: X_i \in C_j\}} \psi_{\hat{\theta}_P, \hat{\nu}_P}(O_i).$$

Under standard LLN arguments the average scores  $\bar{\psi}_{C_j}$  satisfy

$$\bar{\psi}_{C_j} = \mathbb{E}[\psi_{\hat{\theta}_P, \hat{\nu}_P}(O) \mid X \in C_j] + O_P(1/\sqrt{n_{C_j}}). \quad (44)$$

One applies iterated expectation to see

$$\mathbb{E}[\psi_{\hat{\theta}_P, \hat{\nu}_P}(O) \mid X \in C_j] = \mathbb{E}\left[\mathbb{E}\left[\psi_{\hat{\theta}_P, \hat{\nu}_P}(O) \mid X\right] \mid X \in C_j\right] = \mathbb{E}[M_{\hat{\theta}_P, \hat{\nu}_P}(X) \mid X \in C_j],$$

and hence

$$\bar{\psi}_{C_j} = \mathbb{E}[M_{\hat{\theta}_P, \hat{\nu}_P}(X) \mid X \in C_j] + O_P(1/\sqrt{n_{C_j}}). \quad (45)$$

**Expansion of  $M_{\hat{\theta}_P, \hat{\nu}_P}(X)$ .** Under Assumption 2 one considers the Taylor expansion of  $M_{\hat{\theta}_P, \hat{\nu}_P}(X)$  about  $(\theta, \nu) = (\theta^*(x_P), \nu^*(x_P))$ :

$$M_{\hat{\theta}_P, \hat{\nu}_P}(X) = M_{\theta^*(x_P), \nu^*(x_P)}(X) + [\nabla_{(\theta, \nu)} M_{\theta^*(x_P), \nu^*(x_P)}(X)] \begin{bmatrix} \hat{\theta}_P - \theta^*(x_P) \\ \hat{\nu}_P - \nu^*(x_P) \end{bmatrix} + O_P \left( \left\| \begin{bmatrix} \hat{\theta}_P - \theta^*(x_P) \\ \hat{\nu}_P - \nu^*(x_P) \end{bmatrix} \right\|^2 \right).$$

The consistency of the parent solutions  $(\hat{\theta}_P, \hat{\nu}_P)$  for  $(\theta^*(x_P), \nu^*(x_P))$  is established by Athey et al. (2019), and in particular  $(\hat{\theta}_P, \hat{\nu}_P) - (\theta^*(x_P), \nu^*(x_P)) = O_P(r, 1/\sqrt{n_P})$ . The asymptotic regime  $r^{-2} \ll n_P$  implies  $1/\sqrt{n_P} = o(r)$  and therefore the higher order quadratic term is equivalently expressed:

$$M_{\hat{\theta}_P, \hat{\nu}_P}(X) = M_{\theta^*(x_P), \nu^*(x_P)}(X) + [\nabla_{(\theta, \nu)} M_{\theta^*(x_P), \nu^*(x_P)}(X)] \begin{bmatrix} \hat{\theta}_P - \theta^*(x_P) \\ \hat{\nu}_P - \nu^*(x_P) \end{bmatrix} + O_P(r^2),$$

and therefore

$$\begin{aligned} \mathbb{E} [M_{\hat{\theta}_P, \hat{\nu}_P}(X) | X \in C_j] &= \mathbb{E} [M_{\theta^*(x_P), \nu^*(x_P)}(X) | X \in C_j] \\ &\quad + \mathbb{E} [\nabla_{(\theta, \nu)} M_{\theta^*(x_P), \nu^*(x_P)}(X) | X \in C_j] \begin{bmatrix} \hat{\theta}_P - \theta^*(x_P) \\ \hat{\nu}_P - \nu^*(x_P) \end{bmatrix} + O_P(r^2). \end{aligned}$$

One has  $\nabla_{(\theta, \nu)} M_{\theta^*(x_P), \nu^*(x_P)}(X) = V(x_P) + O_P(r)$  because  $M_{\theta, \nu}(x)$  is Lipschitz in  $x$ , and the expansion in the previous display becomes:

$$\mathbb{E} [M_{\hat{\theta}_P, \hat{\nu}_P}(X) | X \in C_j] = \mathbb{E} [M_{\theta^*(x_P), \nu^*(x_P)}(X) | X \in C_j] + V(x_P) \begin{bmatrix} \hat{\theta}_P - \theta^*(x_P) \\ \hat{\nu}_P - \nu^*(x_P) \end{bmatrix} + O_P(r^2). \quad (46)$$

**Expansion of  $M_{\theta^*(x_P), \nu^*(x_P)}(X)$ .** Following similar arguments, the term  $M_{\theta^*(x_P), \nu^*(x_P)}(X)$  is expanded about  $(\theta, \nu) = (\theta^*(X), \nu^*(X))$  as:

$$\begin{aligned} M_{\theta^*(x_P), \nu^*(x_P)}(X) &= M_{\theta^*(X), \nu^*(X)}(X) + V(X) \begin{bmatrix} \theta^*(x_P) - \theta^*(X) \\ \nu^*(x_P) - \nu^*(X) \end{bmatrix} + O_P(r^2), \\ &= V(X) \begin{bmatrix} \theta^*(x_P) - \theta^*(X) \\ \nu^*(x_P) - \nu^*(X) \end{bmatrix} + O_P(r^2), \end{aligned}$$

where  $M_{\theta^*(X), \nu^*(X)}(X) = \mathbf{0}$  holds because  $(\theta^*(X), \nu^*(X))$  are defined as satisfying the GRF moment conditions (1) local to  $X$ . One takes the conditional expectation of the previous display:

$$\mathbb{E} [M_{\theta^*(x_P), \nu^*(x_P)}(X) | X \in C_j] = \mathbb{E} \left[ V(X) \begin{bmatrix} \theta^*(x_P) - \theta^*(X) \\ \nu^*(x_P) - \nu^*(X) \end{bmatrix} \middle| X \in C_j \right] + O_P(r^2).$$

Whenever  $X \in C_j$  one has  $\|X - x_P\| = \mathcal{O}(r)$ , and the same Lipschitz arguments can be applied to see  $V(X) = V(x_P) + O_P(r)$  conditional on  $X \in C_j$ , and the previous display simplifies:

$$\mathbb{E} [M_{\theta^*(x_P), \nu^*(x_P)}(X) | X \in C_j] = V(x_P) \begin{bmatrix} \theta^*(x_P) - \theta_{C_j}^* \\ \nu^*(x_P) - \nu_{C_j}^* \end{bmatrix} + O_P(r^2), \quad (47)$$

where  $\theta_{C_j}^* := \mathbb{E}[\theta^*(X) | X \in C_j]$  and  $\nu_{C_j}^* := \mathbb{E}[\nu^*(X) | X \in C_j]$ . Substitute (47) into the conditional expectation (46):

$$\begin{aligned} \mathbb{E} [M_{\hat{\theta}_P, \hat{\nu}_P}(X) | X \in C_j] &= V(x_P) \begin{bmatrix} \theta^*(x_P) - \theta_{C_j}^* \\ \nu^*(x_P) - \nu_{C_j}^* \end{bmatrix} + V(x_P) \begin{bmatrix} \hat{\theta}_P - \theta^*(x_P) \\ \hat{\nu}_P - \nu^*(x_P) \end{bmatrix} + O_P(r^2), \\ &= V(x_P) \begin{bmatrix} \hat{\theta}_P - \theta_{C_j}^* \\ \hat{\nu}_P - \nu_{C_j}^* \end{bmatrix} + O_P(r^2). \end{aligned}$$

Therefore, the child node score averages  $\bar{\psi}_{C_j}$  in (45) satisfy

$$\bar{\psi}_{C_j} = V(x_P) \begin{bmatrix} \hat{\theta}_P - \theta_{C_j}^* \\ \hat{\nu}_P - \nu_{C_j}^* \end{bmatrix} + O_P(r^2, 1/\sqrt{n_{C_j}}),$$

and the difference  $\bar{\psi}_{C_1} - \bar{\psi}_{C_2}$  satisfies

$$\begin{aligned} \bar{\psi}_{C_1} - \bar{\psi}_{C_2} &= V(x_P) \begin{bmatrix} \hat{\theta}_P - \theta_{C_1}^* \\ \hat{\nu}_P - \nu_{C_1}^* \end{bmatrix} - V(x_P) \begin{bmatrix} \hat{\theta}_P - \theta_{C_2}^* \\ \hat{\nu}_P - \nu_{C_2}^* \end{bmatrix} + O_P\left(r^2, \frac{1}{\sqrt{n_{C_1}}}, \frac{1}{\sqrt{n_{C_2}}}\right), \\ &= -V(x_P) \begin{bmatrix} \theta_{C_1}^* - \theta_{C_2}^* \\ \nu_{C_1}^* - \nu_{C_2}^* \end{bmatrix} + O_P\left(r^2, \frac{1}{\sqrt{n_{C_1}}}, \frac{1}{\sqrt{n_{C_2}}}\right). \end{aligned}$$

We assume  $\eta$  is a fixed scalar and  $\xi^\top$  a fixed matrix, it follows:

$$\eta \xi^\top (\bar{\psi}_{C_1} - \bar{\psi}_{C_2}) = -\eta \xi^\top V(x_P) \begin{bmatrix} \theta_{C_1}^* - \theta_{C_2}^* \\ \nu_{C_1}^* - \nu_{C_2}^* \end{bmatrix} + O_P\left(r^2, \frac{1}{\sqrt{n_{C_1}}}, \frac{1}{\sqrt{n_{C_2}}}\right). \quad (48)$$

The fixed matrix  $\xi^\top$  selects the coordinates of the target effect as  $\xi^\top(\theta, \nu)^\top = \theta$ , and hence the product  $\xi^\top V(x_P)$  simplifies:

$$\xi^\top V(x_P) = \xi^\top \begin{bmatrix} V_{\theta\theta}(x_P) & V_{\theta\nu}(x_P) \\ V_{\nu\theta}(x_P) & V_{\nu\nu}(x_P) \end{bmatrix} = [V_{\theta\theta}(x_P) \quad V_{\theta\nu}(x_P)].$$

Under Neyman orthogonality one has  $V_{\theta\nu}(x_P) = \mathbf{0}$ , implying that  $\xi^\top V(x_P) = [V_{\theta\theta}(x_P) \quad \mathbf{0}]$ , and (48) becomes

$$\eta \xi^\top (\bar{\psi}_{C_1} - \bar{\psi}_{C_2}) = -\eta V_{\theta\theta}(x_P)(\theta_{C_1}^* - \theta_{C_2}^*) + O_P\left(r^2, \frac{1}{\sqrt{n_{C_1}}}, \frac{1}{\sqrt{n_{C_2}}}\right). \quad (49)$$

**Asymptotic analysis.** Let  $E$  be any term satisfying  $E = O_P(r^2, 1/\sqrt{n_{C_1}}, 1/\sqrt{n_{C_2}})$ . In our asymptotic regime with  $r^{-2} \ll n_{C_j} \implies 1/\sqrt{n_{C_j}} = o(r)$ , one has

$$E = O_P\left(r^2, \frac{1}{\sqrt{n_{C_1}}}, \frac{1}{\sqrt{n_{C_2}}}\right) \implies E = o_P\left(r, \frac{1}{\sqrt{n_{C_1}}}, \frac{1}{\sqrt{n_{C_2}}}\right)$$

and therefore (49) satisfies

$$\eta \xi^\top (\bar{\psi}_{C_1} - \bar{\psi}_{C_2}) = -\eta V_{\theta\theta}(x_P)(\theta_{C_1}^* - \theta_{C_2}^*) + o_P\left(r, \frac{1}{\sqrt{n_{C_1}}}, \frac{1}{\sqrt{n_{C_2}}}\right). \quad (50)$$

Write  $A = V_{\theta\theta}(\theta_{C_1}^* - \theta_{C_2}^*)$  such that  $\tilde{\Delta}_\eta^{\text{FPT}}(C_1, C_2)$  is equivalently written  $\tilde{\Delta}_\eta^{\text{FPT}}(C_1, C_2) = (n_{C_1} n_{C_2} / n_P^2) \cdot \eta^2 \|A + E\|^2$ . Consider the difference

$$\begin{aligned} \tilde{\Delta}_\eta^{\text{FPT}}(C_1, C_2) - \frac{n_{C_1} n_{C_2}}{n_P^2} \eta^2 \|V_{\theta\theta}(\theta_{C_1}^* - \theta_{C_2}^*)\|^2 &= \frac{n_{C_1} n_{C_2}}{n_P^2} \eta^2 (\|A + E\|^2 - \|A\|^2), \\ &= \frac{n_{C_1} n_{C_2}}{n_P^2} \eta^2 (2\langle A, E \rangle + \|E\|^2). \end{aligned}$$

One repeats the same arguments used in the final asymptotic analysis of Lemma B.3 to show

$$2\langle A, E \rangle + \|E\|^2 = o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right),$$

and thus

$$\tilde{\Delta}_\eta^{\text{FPT}}(C_1, C_2) - \frac{n_{C_1} n_{C_2}}{n_P^2} \eta^2 \|V_{\theta\theta}(x_P)(\theta_{C_1}^* - \theta_{C_2}^*)\|^2 = o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right),$$

as desired. ■

**Proof of Proposition 4.1.** First, under Assumptions A.1 the  $\theta$ -block  $V_{\theta\theta}(x_P)$  of the local Jacobian  $V(x_P)$  is strictly positive definite and thus  $\|\cdot\|_V$  defines a true norm. From the proof of Lemma B.3:

$$\hat{\theta}_{C_1} - \hat{\theta}_{C_2} = \theta_{C_1}^* - \theta_{C_2}^* + o_P\left(r, \frac{1}{\sqrt{n_{C_1}}}, \frac{1}{\sqrt{n_{C_2}}}\right).$$

The matrix  $V(x_P)$  is non-random and fixed given  $P$  and  $\eta$  is a fixed scalar. It follows:

$$\eta V_{\theta\theta}(x_P)(\hat{\theta}_{C_1} - \hat{\theta}_{C_2}) = \eta V_{\theta\theta}(x_P)(\theta_{C_1}^* - \theta_{C_2}^*) + o_P\left(r, \frac{1}{\sqrt{n_{C_1}}}, \frac{1}{\sqrt{n_{C_2}}}\right).$$

Up to a negative factor, the expression on the right is precisely the same as (50) in the proof of Lemma B.4, and thus one repeats the arguments to arrive at

$$\left\| \eta V_{\theta\theta}(x_P)(\hat{\theta}_{C_1} - \hat{\theta}_{C_2}) \right\|_2^2 = \left\| \eta V_{\theta\theta}(x_P)(\theta_{C_1}^* - \theta_{C_2}^*) \right\|_2^2 + o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right),$$

and hence

$$\Delta_{\eta V}(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P^2} \eta^2 \left\| V_{\theta\theta}(x_P)(\theta_{C_1}^* - \theta_{C_2}^*) \right\|_2^2 + o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right).$$

The right hand side is precisely the same as in the statement of Lemma B.4 established for  $\tilde{\Delta}_\eta^{\text{FPPT}}(C_1, C_2)$ , and thus

$$\tilde{\Delta}_\eta^{\text{FPPT}}(C_1, C_2) - \Delta_{\eta V}(C_1, C_2) = o_P\left(r^2, \frac{1}{n_{C_1}}, \frac{1}{n_{C_2}}\right), \quad (51)$$

as desired. ■

**Proof of Lemma 4.2.** Firstly, Specifications 4 (subsampling) and 5 (honesty) describe conditions imposed on the sampling mechanism and are not affected by the form of the splitting criterion. It remains to verify whether Specification 1 (symmetry), Specification 2 (balanced/ $\omega$ -regular), and Specification 3 (randomized/random-split) are satisfied by  $\mathcal{T}(\Delta_V)$ .

1. **Specification 1: Symmetry.** A tree is said to be symmetric if its estimates are invariant under permutations of the tree's training samples. Conditional on a sequence of criterion values computed over splits of  $P$ , the CART mechanism of selecting the best split by scanning over the collection of candidates does not depend on the parent samples at all. This means that asymmetry could only enter through the criterion values. Therefore, a sufficient condition for symmetry in the tree estimates with respect to permutations of the tree samples is whether the criterion  $\Delta_V(C_1, C_2)$  is symmetric. Conditional on the child solutions  $\hat{\theta}_{C_1}, \hat{\theta}_{C_2}$ , the map  $(\hat{\theta}_{C_1}, \hat{\theta}_{C_2}) \mapsto \Delta_V(C_1, C_2)$  does not depend on the parent samples at all, and therefore asymmetry could only enter through child solutions  $\hat{\theta}_{C_j}$ . However, both criteria use precisely the same child solutions  $\hat{\theta}_{C_j}$  in (7), and therefore  $\Delta_V(C_1, C_2)$  will be symmetric whenever  $\Delta(C_1, C_2)$  is symmetric (specifically, whenever  $\psi_{\theta,\nu}(O_i)$  is symmetric with respect to permutations of  $i$ ).
2. **Specification 2: Balanced/ $\omega$ -regular.** This condition is enforced by GRF by adding an additional stopping condition to the gradient-based version of Algorithm 1. Specifically, GRF stops a recursive splitting path if a proposed  $\Delta$ -optimal split were to send fewer than  $\omega n_P$  of the parent samples into either child. Simply stated, GRF enforces balanced splits by defining the set of valid candidate splits as those with at least  $\omega n_P$  parent samples in each child. This is left unchanged by our method.
3. **Specification 3: Randomized/random-split.** The asymptotic theory of GRF requires that, at each node, each variable is selected for a split with some lower bound probability  $\pi > 0$ . In order to satisfy the minimum split probability GRF uses the feature sampling device of Denil et al. (2014) which, at each step, considers only  $\min\{\max\{\text{Poisson}(m), 1\}, p\}$  randomly selected features as candidate variables, where  $p = \dim(\mathcal{X})$  and  $m$  is a GRF tuning parameter. In other words, GRF defines the set of valid candidate splits such that the set of valid splitting dimensions is itself a random variable. This mechanism is left unchanged under our method.

No column of  $V(x_P)$  is all-zero  $V_{\cdot,k}(x_P) \neq \mathbf{0}$  because  $V(x_P)$  is strictly positive definite symmetric, and therefore  $\Delta_V(C_1, C_2)$  will not systematically ignore signals along parameter dimensions  $\theta_k$  that can be detected by the  $\Delta$ -criterion. Finally,

$$\hat{\theta}_{C_1} - \hat{\theta}_{C_2} \neq \mathbf{0} \implies V_{\theta\theta}(x_P)(\hat{\theta}_{C_1} - \hat{\theta}_{C_2}) \neq \mathbf{0},$$

because  $V(x)$  is strictly positive definite symmetric by Assumption 2. Therefore  $\Delta(C_1, C_2) > 0 \implies \Delta_V(C_1, C_2) > 0$  meaning that the  $\Delta_V$ -criterion is non-degenerate and will always be able to select at least one feature whenever the  $\Delta$ -criterion can select a feature.

Therefore, all five specifications are met, and one concludes that  $\mathcal{T}(\Delta_V)$  must satisfy the forest Specifications A.2 whenever they are satisfied by  $\mathcal{T}(\Delta)$ .  $\blacksquare$

### B.3. Asymptotic equivalence of the pseudo-outcome approximation for VCM/HTE models

In this section we establish the asymptotic equivalence of the further acceleration of the fixed-point method proposed in Section 5 for VCM/HTE models. The accelerated algorithm is based on FPT pseudo-outcomes that use an approximation  $\tilde{\theta}_P$  for the actual parent solutions  $\hat{\theta}_P$  in (31). Specifically, the parent-leaf approximations  $\tilde{\theta}_P$  are found by a single gradient descent step towards  $\hat{\theta}_P$  taken from the origin (54). Let  $\rho_i^{\text{FPT}}$  denote the original FPT pseudo-outcomes for VCM/HTE models:

$$\rho_i^{\text{FPT}} := -(W_i - \bar{W}_P) \left( Y_i - \bar{Y}_P - (W_i - \bar{W}_P)^\top \hat{\theta}_P \right), \quad (52)$$

where the solution  $\hat{\theta}_P$  for the local model over the parent  $P$  are precisely the OLS coefficients from the regression the centered outcomes  $Y_i - \bar{Y}_P \in \mathbb{R}$  on the centered regressors  $W_i - \bar{W}_P \in \mathbb{R}^K$ . In contrast, let  $\phi_i^{\text{FPT}}$  denote approximations of  $\rho_i^{\text{FPT}}$  pseudo-outcomes that are of the form

$$\phi_i^{\text{FPT}} := -(W_i - \bar{W}_P) \left( Y_i - \bar{Y}_P - (W_i - \bar{W}_P)^\top \tilde{\theta}_P \right), \quad (53)$$

where  $\tilde{\theta}_P$  approximates  $\hat{\theta}_P$  as:

$$\tilde{\theta}_P := \gamma \cdot \frac{1}{n_P} \sum_{\{i: X_i \in P\}} (W_i - \bar{W}_P)(Y_i - \bar{Y}_P) = \gamma \cdot \frac{1}{n_P} W_P^\top Y_P. \quad (54)$$

Here,  $W_P \in \mathbb{R}^{n_P \times K}$  and  $Y_P \in \mathbb{R}^{n_P}$  denote the centered data matrices,  $W_P := [W_i - \bar{W}_P]_{i: X_i \in P}$  and  $Y_P := [Y_i - \bar{Y}_P]_{i: X_i \in P}$ , and the scalar  $\gamma > 0$  denotes the exact line search step size corresponding to the regression of the centered outcomes on the centered regressors:

$$\gamma := \frac{\|W_P^\top Y_P\|_2^2}{\|W_P W_P^\top Y_P\|_2^2}. \quad (55)$$

**Lemma B.5.** *Let  $\tilde{\theta}_{C_j}$  denote the FPT estimator of the form (24) for the child solution  $\hat{\theta}_{C_j}$  for VCM/HTE models. One can express  $\tilde{\theta}_{C_j}$  in terms of the corresponding fixed-point pseudo-outcomes:*

$$\tilde{\theta}_{C_j} := \hat{\theta}_P + \frac{1}{n_{C_j}} \sum_{i: X_i \in C_j} \rho_i^{\text{FPT}}.$$

Similarly, denote by  $\bar{\theta}_{C_j}$  the FPT estimator of  $\hat{\theta}_{C_j}$  induced by pseudo-outcomes approximations  $\phi_i^{\text{FPT}}$ :

$$\bar{\theta}_{C_j} := \hat{\theta}_P + \frac{1}{n_{C_j}} \sum_{\{i: X_i \in C_j\}} \phi_i^{\text{FPT}}.$$

Then, under the assumptions of Proposition 4.1,  $\bar{\theta}_{C_j}$  is consistent for  $\tilde{\theta}_{C_j}$  as:

$$\|\tilde{\theta}_{C_j} - \bar{\theta}_{C_j}\| = o_P(1).$$

*Proof.* A direct calculation reveals that the difference between the original FPT pseudo-outcomes  $\rho_i^{\text{FPT}}$  in (52) and the approximations  $\phi_i^{\text{FPT}}$  in (53) satisfy

$$\begin{aligned} \rho_i^{\text{FPT}} - \phi_i^{\text{FPT}} &= -(W_i - \bar{W}_P) \left( \left[ Y_i - \bar{Y}_P - (W_i - \bar{W}_P)^\top \hat{\theta}_P \right] - \left[ Y_i - \bar{Y}_P - (W_i - \bar{W}_P)^\top \tilde{\theta}_P \right] \right), \\ &= (W_i - \bar{W}_P)(W_i - \bar{W}_P)^\top (\hat{\theta}_P - \tilde{\theta}_P). \end{aligned} \quad (56)$$

Therefore, the difference between the original FPT child estimator  $\tilde{\theta}_{C_j}$  and the approximation  $\bar{\theta}_{C_j}$  satisfies

$$\begin{aligned}\tilde{\theta}_{C_j} - \bar{\theta}_{C_j} &= \frac{1}{n_{C_j}} \sum_{\{i: X_i \in C_j\}} (\rho_i^{\text{FPT}} - \phi_i^{\text{FPT}}), \\ &= \frac{1}{n_{C_j}} \sum_{\{i: X_i \in C_j\}} (W_i - \bar{W}_P)(W_i - \bar{W}_P)^\top (\hat{\theta}_P - \tilde{\theta}_P), \\ &= S_{C_j}(\hat{\theta}_P - \tilde{\theta}_P),\end{aligned}$$

where we denote  $S_{C_j} := \frac{1}{n_{C_j}} \sum_{\{i: X_i \in C_j\}} (W_i - \bar{W}_P)(W_i - \bar{W}_P)^\top$ . Therefore,

$$\|\tilde{\theta}_{C_j} - \bar{\theta}_{C_j}\| = \|S_{C_j}(\hat{\theta}_P - \tilde{\theta}_P)\| \leq \|S_{C_j}\|_F \|\hat{\theta}_P - \tilde{\theta}_P\|. \quad (57)$$

Under GRF's regularity conditions, in a limit where  $n_{C_j} \rightarrow \infty$  and the parent radius  $r := \sup_{\{i: X_i \in P\}} \|X_i - \bar{X}_P\|$  goes to zero  $r \rightarrow 0$ , we have  $S_{C_j} \xrightarrow{p} Q$  for some positive semidefinite symmetric matrix  $Q$ , and hence  $\|S_{C_j}\|_F = O_P(1)$ . Meanwhile, by definition (31) for the OLS coefficients  $\hat{\theta}_P$  and definition (54) for the one-step approximations  $\bar{\theta}_P$ , we have

$$\begin{aligned}\|\hat{\theta}_P - \tilde{\theta}_P\| &= \|(-A_P^{-1} \cdot n_P^{-1} W_P^\top Y_P) - (\gamma \cdot n_P^{-1} W_P^\top Y_P)\|, \\ &= \|(-A_P^{-1} - \gamma \mathbb{I}) \cdot n_P^{-1} W_P^\top Y_P\|, \\ &\leq \| -A_P^{-1} - \gamma \mathbb{I} \|_F \|n_P^{-1} W_P^\top Y_P\|, \\ &= \left\| [n_P^{-1} W_P^\top W_P]^{-1} - \gamma \mathbb{I} \right\|_F \|n_P^{-1} W_P^\top Y_P\|,\end{aligned} \quad (58)$$

where  $-A_P = n_P^{-1} W_P^\top W_P$  follows from the definition of  $A_P$  as an estimator of the Jacobian  $\nabla\psi$ , e.g. (13) in the context of VCM/HTE models. Under the Lipschitz continuity Assumptions 1 & 3, one has the standard stochastic bound for the cross term  $n_P^{-1} W_P^\top Y_P$ :

$$\|n_P^{-1} W_P^\top Y_P\| = O_P\left(r, \frac{1}{\sqrt{n_P}}\right), \quad (59)$$

while the difference  $[n_P^{-1} W_P^\top W_P]^{-1} - \gamma \mathbb{I}$  is stochastically bound as

$$\left\| [n_P^{-1} W_P^\top W_P]^{-1} - \gamma \mathbb{I} \right\|_F = O_P(1),$$

because  $n_P^{-1} W_P^\top W_P \xrightarrow{p} \text{Cov}(W_i \mid X_i \in P)$  is non-singular under Assumption 2. Coupling these stochastic bounds together according to (58) gives

$$\|\hat{\theta}_P - \tilde{\theta}_P\| = O_P\left(r, \frac{1}{\sqrt{n_P}}\right),$$

and trivially, because  $n_{C_j} < n_P$ ,

$$\|\hat{\theta}_P - \tilde{\theta}_P\| = O_P\left(r, \frac{1}{\sqrt{n_{C_j}}}\right). \quad (60)$$

Under Proposition 1 of GRF one assumes  $r^{-2} \ll n_{C_1}, n_{C_2}$  and thus, in an asymptotic regime where  $n_{C_j} \rightarrow \infty$  and  $r \rightarrow 0$ , one has  $1/\sqrt{n_{C_j}} = o(r)$ , and hence:

$$\|\hat{\theta}_P - \tilde{\theta}_P\| = o_P(1). \quad (61)$$

Returning to (57), the consistency of the parent approximation  $\hat{\theta}_P$  as (61) implies that the approximation  $\bar{\theta}_{C_j}$  is itself consistent for the original FPT child estimator  $\tilde{\theta}_{C_j}$ :

$$\|\tilde{\theta}_{C_j} - \bar{\theta}_{C_j}\| = o_P(1), \quad (62)$$

as desired. ■

## C. Implementation Details

### C.1. Honest subsampling

In this section we present the honest subsampling mechanism. Trees are used to form partitions of the input space such as to specify weight functions  $\alpha_i(x)$ , defined as

$$\alpha_i(x) := \frac{1}{B} \sum_{b=1}^B \alpha_{bi}(x), \quad \text{for } \alpha_{bi}(x) := \frac{\mathbb{1}(X_i \in L_b(x))}{|L_b(x)|}, \quad i = 1, \dots, n, \quad (63)$$

where  $L_b(x)$  denotes a subset training samples that fall alongside  $x$  according to the partition of tree  $b$ . The honesty mechanism ensures that no observation in leaf  $L_b(x)$  was used to build the partition rules of tree  $b$ . This is achieved by separating an initial subsample into two subsets: One for building the partition rules, and the other allocated as samples to the local leaves  $L_b(x)$  according to the trained rules. Below, we give a detailed outline of how subsampling and honest sample splitting is used to train a forest of trees, then show that weight function  $\alpha_i(x)$  given by honest trees according to (63) is conditionally independent of  $O_i$  given  $X_i$ .

#### Honest subsampling for GRF

For tree  $b \in \{1, \dots, B\}$ ,

1. (*Subsampling*). Draw an initial subsample  $\mathcal{I}^{(b)}$  of size  $s := |\mathcal{I}^{(b)}|$  from the training set (without replacement).
2. (*Honest splitting*). Split  $\mathcal{I}^{(b)}$  into disjoint sets  $\mathcal{J}_1^{(b)}$  and  $\mathcal{J}_2^{(b)}$  of size  $|\mathcal{J}_1^{(b)}| = \lfloor s/2 \rfloor$  and  $|\mathcal{J}_2^{(b)}| = \lceil s/2 \rceil$ .
  - (a) Train tree  $T(\mathcal{J}_1^{(b)})$  based on the first subsample  $\{(X_i, O_i) : i \in \mathcal{J}_1^{(b)}\}$ . Let  $\mathcal{R}_1^{(b)}, \dots, \mathcal{R}_M^{(b)}$  denote the partition of  $\mathcal{X}$  induced by  $T(\mathcal{J}_1^{(b)})$  such that
 
$$\mathcal{R}_m^{(b)} := \left\{ x \in \mathcal{X} : x \text{ satisfies the partition rules for leaf } m \text{ of } T(\mathcal{J}_1^{(b)}) \right\}.$$
  - (b) Subset the samples from the second subsample  $\{X_i : i \in \mathcal{J}_2^{(b)}\}$  according to the trained rules of  $T(\mathcal{J}_1^{(b)})$ , i.e. the samples of  $\mathcal{J}_2^{(b)}$  in the leaves are determined by the rules of  $T(\mathcal{J}_1^{(b)})$ .

For any  $x \in \mathcal{X}$ , the local leaf  $L_b(x)$  that appears in (63) is defined as the specific subset of  $\mathcal{J}_2^{(b)}$  samples belonging to the same leaf of tree  $T(\mathcal{J}_1^{(b)})$  as  $x$ ,

$$L_b(x) = \{X_i \in \mathcal{R}_m^{(b)} : i \in \mathcal{J}_2^{(b)} \text{ and } x \in \mathcal{R}_m^{(b)}\},$$

**Conditional independence of  $\alpha_i(x)$  and  $O_i$  given  $X_i$ .** By definition, the partition rules of tree  $T(\mathcal{J}_1^{(b)})$  depend only on the  $\mathcal{J}_1^{(b)}$  subsample. The rules of a tree operate only on covariate values, and therefore the task of subsetting  $\{X_i : i \in \mathcal{J}_2^{(b)}\}$  into leaves according to the rules of  $T(\mathcal{J}_1^{(b)})$  requires knowledge of the  $X_i$  values from the  $\mathcal{J}_2^{(b)}$  subsample but not necessarily the  $O_i$ . Based on this understanding, we will show that  $\alpha_i(x)$  is conditionally independent of  $O_i$  given  $X_i$ . Based on (63), it is sufficient to show

$$\mathbb{E}[\alpha_{bi}(x) | O_i, X_i] = \mathbb{E}[\alpha_{bi}(x) | X_i].$$

Case 1. Suppose  $i \notin \mathcal{J}_2^{(b)}$ . By definition  $L_b(x) \subset \{X_j : j \in \mathcal{J}_2^{(b)}\}$ . It is immediate that  $\mathbb{1}(\{X_i \in L_b(x)\}) = 0$ , and therefore  $\alpha_{bi}(x) = 0$ , and trivially

$$\mathbb{E}[\alpha_{bi}(x) | O_i, X_i] = \mathbb{E}[\alpha_{bi}(x) | X_i] = 0, \quad \text{for all } i \notin \mathcal{J}_2^{(b)}.$$

Case 2. Suppose  $i \in \mathcal{J}_2^{(b)}$ . We show that each component used to specify  $\alpha_{bi}(x)$  in (63) is conditionally independent of  $O_i$  given  $X_i$ :

- Tree  $T(\mathcal{J}_1^{(b)})$  is trained using only the  $\mathcal{J}_1^{(b)}$  subsample. This does not depend on  $O_i$ , for all  $i \in \mathcal{J}_2^{(b)}$ , conditionally on  $X_i$  or otherwise.

- The rules of tree  $T(\mathcal{J}_1^{(b)})$  operate only on input values. Therefore, conditionally on  $X_i$  for all  $i \in \mathcal{J}_2^{(b)}$ , the leaves of the  $\mathcal{J}_2^{(b)}$  subsample specified by tree  $T(\mathcal{J}_1^{(b)})$  do not depend on the value of  $O_i$ .
  - Leaf  $L_b(x)$  is the specific subset of the  $\mathcal{J}_2^{(b)}$  samples satisfying the same partition rules of  $T(\mathcal{J}_1^{(b)})$  as  $x$ . Given the leaves have been specified by the previous step, this depends only on  $x$ .

Therefore, the individual component functions  $\alpha_{bi}(x) = \mathbb{1}(\{X_i \in L_b(x)\})/|L_b(x)|$  are conditionally independent of  $O_i$  given  $X_i$ ,

$$\mathbb{E}[\alpha_{bi}(x) \mid O_i, X_i] = \mathbb{E}[\alpha_{bi}(x) \mid X_i], \quad \text{for all } i \in \mathcal{J}_2^{(b)}.$$

**Demonstration of honest subsampling.** Let  $\{(X_i, O_i)\}_{i=1}^n$  denote a training set of  $n = 20$  observations, where each  $X_i = (X_{i,1}, X_{i,2})$  is over  $\mathcal{X} \equiv \mathbb{R}^2$ . We will use a forest of a single tree ( $B = 1$ ) to specify the functional form of weights  $\alpha_i(x)$ .

1. (Subsampling). Draw an initial subsample  $\mathcal{I}$  of size  $s = 10$ .
  2. (Honest splitting). Split  $\mathcal{I}$  into two disjoint sets  $\mathcal{J}_1$  and  $\mathcal{J}_2$ , each with  $s/2 = 5$  samples.

The diagram illustrates the decomposition of a dataset into two clusters,  $\mathcal{J}_1$  and  $\mathcal{J}_2$ , based on the variable  $i \in \mathcal{I}$ .

**Initial Dataset:**

$i$	$X_{i,1}$	$X_{i,2}$	$O_i$
1			
⋮			
20			

**Transformation:** An arrow labeled  $\mathcal{I} \rightarrow$  indicates the mapping from the original dataset to the decomposed datasets.

**Cluster  $\mathcal{J}_1$ :**

$i \in \mathcal{I}$	$X_{i,1}$	$X_{i,2}$	$O_i$
2			
3			
5			
8			
10			
11			
14			
15			
16			
20			

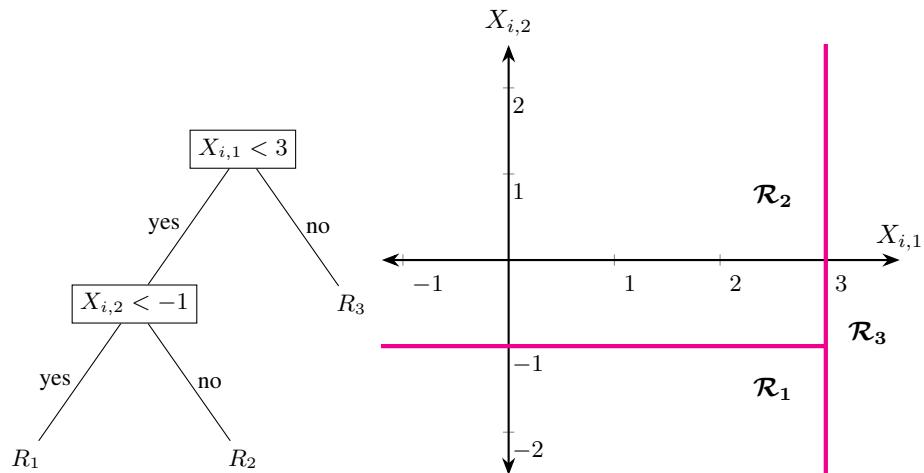
**Cluster  $\mathcal{J}_2$ :**

$i \in \mathcal{I}$	$X_{i,1}$	$X_{i,2}$	$O_i$
5	1	0	
10	2	-2	
11	0	1	
14	1	-2	
16	2	2	

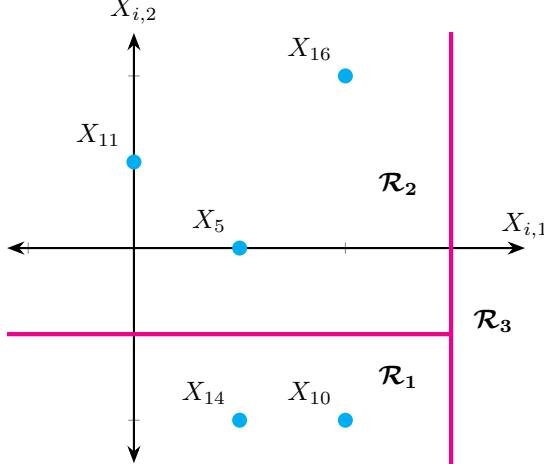
**Final Output:**

$i \in \mathcal{J}_1$	$X_{i,1}$	$X_{i,2}$	$O_i$
2			
3			
8			
15			
20			

- (a) Train a tree using the data from the first subsample  $\mathcal{J}_1$ , inducing a partition of  $\mathcal{X} \equiv \mathbb{R}^2$ . Suppose the fitted tree has the following structure:



- (b) Use the trained partition rules to subset the  $\mathcal{J}_2$  subsample into separate leaves.



The tree trained on the  $\mathcal{J}_1$  subsample will subset the  $\mathcal{J}_2$  subsample as

$$\{X_i : i \in \mathcal{J}_2\} = \{X_{10}, X_{14}\} \cup \{X_5, X_{11}, X_{16}\} \cup \emptyset,$$

where we include the trivial union with  $\emptyset$  to note that the tree assigns none of the  $\mathcal{J}_2$  samples to the partition of  $\mathbb{R}^2$  where  $X_{i,1} \geq 3$ .

The leaf  $L_b(x)$  is the specific subset of the  $\mathcal{J}_2$  subsample such that  $X_i \in \mathcal{J}_2$  satisfy the same partition rules as  $x$ . Given a test point  $x = x_0$ , there are three possible scenarios for  $L_b(x_0)$  that correspond to the three regions  $R_1, R_2, R_3 \subset \mathbb{R}^2$  in which the test point  $x_0$  can appear.

**Region 1.** If  $x_0 \in R_1$  then  $L_b(x_0) = \{X_{10}, X_{14}\}$  and

$$\alpha_{bi}(x_0) = \frac{\mathbf{1}(\{X_i \in L_b(x_0)\})}{|L_b(x_0)|} = \begin{cases} \frac{1}{2} & \text{if } i \in \{10, 14\}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,  $\alpha_i(x_0) = \frac{1}{2}$  for  $i = 10, 14$  and zero for  $i \in \{1, \dots, 20\} \setminus \{10, 14\}$ .

**Region 2.** If  $x_0 \in R_2$  then  $L_b(x_0) = \{X_5, X_{11}, X_{16}\}$  and

$$\alpha_{bi}(x_0) = \frac{\mathbf{1}(\{X_i \in L_b(x_0)\})}{|L_b(x_0)|} = \begin{cases} \frac{1}{3} & \text{if } i \in \{5, 11, 16\}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,  $\alpha_i(x_0) = \frac{1}{3}$  for  $i = 5, 11, 16$  and zero for  $i \in \{1, \dots, 20\} \setminus \{5, 11, 16\}$ .

**Region 3.** If  $x_0 \in R_3$  then  $L_b(x_0) = \emptyset$ . This is a degenerate case such that

$$\alpha_{bi}(x_0) = \frac{\mathbf{1}(\{X_i \in L_b(x_0)\})}{|L_b(x_0)|}$$

is undefined, leading to a non-identifiability problem whenever  $x_0 \in R_3$ . When this occurs, Tibshirani et al. (2024) recommends calculating  $\alpha_i(x_0)$  based on only the trees with non-empty  $L_b(x_0)$ . Let  $\mathcal{B} := \{b \in \{1, \dots, B\} : |L_b(x_0)| > 0\}$  denote the indices of non-empty leaves associated with  $x_0$ . Then, the GRF weight functions based on this recommendation can be written as

$$\alpha_i(x_0) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \alpha_{bi}(x_0).$$

---

**Algorithm 1** The fixed-point tree algorithm

---

```

function TRAINFIXEDPOINTTREE
    Input: node  $\mathcal{N}$ 
    node  $P_0 \leftarrow \text{GETSAMPLES}(\mathcal{N})$ 
    queue  $\mathcal{Q} \leftarrow \text{INITIALIZEQUEUE}(P_0)$ 
    while NOTNULL(node  $P \leftarrow \text{POP}(\mathcal{Q})$ ) do
         $(\hat{\theta}_P, \hat{\nu}_P) \leftarrow \text{SOLVEESTIMATINGEQUATION}(P)$   $\triangleright$  Computes (6).
         $\rho^{\text{FPT}} \leftarrow \text{FPTPSEUDOOUTCOMES}(\hat{\theta}_P, \hat{\nu}_P)$   $\triangleright$  Applies (26) over  $P$ .
        split  $\Sigma \leftarrow \text{CARTSPLIT}(P, \rho^{\text{FPT}})$   $\triangleright$  Optimizes (19).
        if SPLITSUCCEEDED( $\Sigma$ ) then
            SETCHILDREN( $P$ , GETLEFTCHILD( $\Sigma$ ), GETRIGHTCHILD( $\Sigma$ ))
            ADDTOQUEUE( $\mathcal{Q}$ , GETLEFTCHILD( $\Sigma$ ))
            ADDTOQUEUE( $\mathcal{Q}$ , GETRIGHTCHILD( $\Sigma$ ))
        end if
    end while
    Output: tree with root node  $P_0$ 
end function

```

---

`POP` returns and removes the oldest element of queue a  $\mathcal{Q}$ , unless  $\mathcal{Q}$  is empty, in which case it returns `NULL`. `CARTSPLIT` runs a multivariate CART split on the pseudo-outcomes  $\rho^{\text{FPT}} := \{\rho_i^{\text{FPT}}\}_{i \in P}$ , and either returns a pair of child nodes or indicates that no split of  $P$  is possible.

---

## C.2. In-sample predictions

There is additional bias associated with making predictions based on in-sample observations  $X_i$  that may have been used either to train the tree structure or to populate the local leaves  $L_b(x)$ . The recommendation of Tibshirani et al. (2024) is along the lines of the out-of-bag mechanism used by Breiman (2001). For an in-sample observation  $x' \in \{X_i\}_{i=1}^n$ , calculate weights  $\alpha_i^{\text{oob}}(x')$  based only on those trees whose initial subsample  $\mathcal{I}^{(b)}$  does not contain  $x'$ . Then the out-of-bag weight is defined as:

$$\alpha_i^{\text{oob}}(x') := \frac{1}{|\{b : x' \notin \mathcal{I}^{(b)}\}|} \sum_{\{b : x' \notin \mathcal{I}^{(b)}\}} \alpha_{bi}^{\text{oob}}(x') \quad \text{for} \quad \alpha_{bi}^{\text{oob}}(x') := \frac{\mathbb{1}(X_i \in L_b(x'))}{|L_b(x')|}.$$

The in-sample prediction  $\hat{\theta}^{\text{oob}}(x')$  for  $x'$  is made by GRF by solving a version of the locally weighted estimating equation (5) using out of bag weights  $\alpha_i^{\text{oob}}(x')$

$$\left( \hat{\theta}^{\text{oob}}(x'), \hat{\nu}^{\text{oob}}(x') \right) \in \arg \min_{\theta, \nu} \left\| \sum_{i=1}^n \alpha_i^{\text{oob}}(x') \psi_{\theta, \nu}(O_i) \right\|,$$

which preserve the consistency and asymptotic normality of the GRF estimator at in-sample observations.

## C.3. Algorithms and Pseudocode

## C.4. Simulation Details

**Implementation details.** We implement the GRF-FPT algorithm in a fork of `grf` (Tibshirani et al., 2024) available at <https://github.com/dfleis/grf>. The functions `grf::lm_forest` and `grf::multi_arm_causal_forest` provide an easy to use interface for VCM and HTE estimation, respectively, and we allow the choice `GRF-FPT1`, `GRF-FPT2`, or `GRF-grad` to be controlled via the `method` argument. Code and data for reproducing all experiments and figures are available at <https://github.com/dfleis/grf-experiments>.

**Data-generating settings.** The different setting for the target effects  $\theta_k^*(x)$  include a sparse linear setting, a sparse logistic setting with interaction, a dense logistic setting, and a random function generator setting. Tables 2 and 3 provide the details of each regime for VCM and HTE experiments, respectively, for the data-generating model (28). These tables also summarize

---

**Algorithm 2** Stage I GRF-FPT: Training a generalized random forest using fixed-point trees

---

```

function TRAINGENERALIZEDRANDOMFORESTFPT
    Input: samples  $\mathcal{S}$ , number of trees  $B$ 
    for  $b = 1, \dots, B$  do
        set of samples  $\mathcal{I} \leftarrow \text{SUBSAMPLE}(\mathcal{S})$ 
        sets of samples  $\mathcal{J}_{\text{BUILD}}, \mathcal{J}_{\text{POPULATE}} \leftarrow \text{HONESTSPLIT}(\mathcal{I})$ 
        tree  $\mathcal{T}_b \leftarrow \text{TRAINFIXEDPOINTTREE}(\mathcal{J}_{\text{BUILD}})$ 
        leaves  $\mathcal{L}_b \leftarrow \text{POPULATELEAVES}(\mathcal{T}_b, \mathcal{J}_{\text{POPULATE}})$ 
        end for
        Output: forest  $\mathcal{F} \leftarrow \{\mathcal{L}_1, \dots, \mathcal{L}_B\}$ 
    end function

```

---

`POPULATELEAVES` creates a collection of subsets (leaves) of the  $\mathcal{J}_{\text{POPULATE}}$  samples based on the partition rules of tree  $\mathcal{T}_b$ . For weight functions  $\alpha_i(x)$ , see `GETWEIGHTS` in Algorithm 3. For Stage II, see `ESTIMATE` in Algorithm 3, where estimates  $\hat{\theta}(x)$  are made given a forest  $\mathcal{F}$ .

---

**Algorithm 3** GRF-FPT: Estimates of  $\theta^*(x)$ 


---

```

function ESTIMATE
    Input: forest  $\mathcal{F}$ , test observation  $x \in \mathcal{X}$ 
    weights  $\alpha \leftarrow \text{GETWEIGHTS}(\mathcal{F}, x)$ 
    Output:  $\hat{\theta}(x)$ , the solution to the weighted estimating equation (5) using weights  $\alpha$ 
end function

function GETWEIGHTS
    Input: forest  $\mathcal{F}$ , test observation  $x$ 
    vector of weights  $\alpha \leftarrow \text{ZEROS}(n)$  ▷ Initialize weights;  $n = |\mathcal{S}|$  used to train  $\mathcal{F}$ .
    for indices  $i : X_i \in L_b(x)$  do
         $\alpha[i] += 1/|L_b(x)|$ 
    end for
    Output: local weights  $\alpha \leftarrow \alpha/|\mathcal{F}|$  ▷ Weights (4).
end function

```

---

Stage II of the GRF-FPT algorithm. The procedure `ESTIMATE` returns an estimate of  $\theta^*(x)$  given a forest  $\mathcal{F}$  trained under Stage I and a test observation  $x$ ; see Algorithm 2.

---

the different settings used to generate the  $K$ -dimensional regressors  $W_i = (W_{i,1}, \dots, W_{i,K})^\top$ . For VCM experiments,  $W_{i,k} \sim \mathcal{N}(0, 1)$  for all  $k = 1, \dots, K$ . For HTE experiments,  $W_i | X_i = x \sim \text{Multinomial}(1, (\pi_1(x), \dots, \pi_K(x)))$ , where  $\pi_k(x)$  denotes the underlying probability the sample is observed as having treatment level  $k \in \{1, \dots, K\}$ .

**Random function generator.** The effect functions  $\theta_k^*(x) = \text{RFG}(x)$  under VCM Setting 4 (in Table 2) and HTE Setting 5 (in Table 3) follow the random function generator design of Friedman (2001). The idea is to measure the performance of the estimator under a variety of randomly generated targets. Each  $\theta_k^*(\cdot)$  is randomly generated as a linear combination of functions  $\{g_\ell(\cdot)\}_{\ell=1}^{20}$  of the form

$$\theta_k^*(x) = \sum_{\ell=1}^{20} a_\ell g_\ell(z_\ell),$$

where the coefficients  $\{a_\ell\}_{\ell=1}^{20}$  are randomly generated from a uniform distribution  $a_\ell \sim \mathcal{U}([-1, 1])$ . Each  $g_\ell(z_\ell)$  is a function of a randomly selected  $p_\ell$ -size subset of the  $p$ -dimensional variable  $x$ , where the size of each subset  $p_\ell$  is randomly chosen by  $p_\ell = \min(\lfloor 1.5 + r_\ell \rfloor, p)$ , and  $r_\ell$  is generated from an exponential distribution with mean 2,  $r_\ell \sim \text{Exp}(0.5)$ . Each  $g(z_\ell)$  uses a  $p_\ell$ -sized random subset  $z_\ell \in \mathbb{R}^{p_\ell}$  of the  $p$ -dimensional input  $x \in \mathbb{R}^p$ :

$$z_\ell := (x_{\phi_\ell(1)}, \dots, x_{\phi_\ell(p_\ell)}) \in \mathbb{R}^{p_\ell},$$

## Generalized Random Forests using Fixed-Point Trees

Parameter	Values	Parameter	Values
$K$	4; 16; 64; 256	$K$	4; 16;
$n$	10,000; 20,000; 100,000	$n$	1000; 4000
$\dim(\mathcal{X})$	5	$\dim(\mathcal{X})$	2
nTrees	100	nTrees	100; 500

Table 1: Parameter values for VCM and HTE experiments in Section 6. Target/regressor dimension  $K$ , number of observations  $n$ , dimension of the auxiliary variables  $\dim(\mathcal{X})$ , and number of trees nTrees. Experiments include a large- $n$  setting (left table) and a small- $n$  setting (right table).

VCM Setting	Effect function $\theta_k^*(x)$	$W_{i,k}$
1	$\theta_k^*(x) = \beta_{k1}x_1, \beta_{k1} \sim \mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$
2	$\theta_k^*(x) = \varsigma(\beta_{k1}x_1)\varsigma(\beta_{k2}x_2), \beta_{k1}, \beta_{k2} \sim \mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$
3	$\theta_k^*(x) = \varsigma(\beta_k^\top x)$ , for $\beta_k \sim \mathcal{N}_p(\mathbf{0}, \mathbb{I})$	$\mathcal{N}(0, 1)$
4	$\theta_k^*(x) = \text{RFG}(x)$	$\mathcal{N}(0, 1)$

Table 2: Settings for the true effects  $\theta_k^*(\cdot)$  and the regressors  $W_{i,k}$  for VCM experiments in Section 6. The function  $\varsigma(u) := 1 + (1 + e^{-20(u-1/3)})^{-1}$  is a logistic-type function in (Athey et al., 2019). The random function generator RFG( $x$ ) is described in Appendix C.4.

HTE Setting	Treatment effect $\theta_k^*(x)$	Treatment probability $\pi_k(x)$ for $W_{i,k}$
1	$\theta_k^*(x) = \beta_{k1}x_1, \beta_{k1} \sim \mathcal{N}(0, 1)$	$\pi_k(x) = 1/K$ for all $k$ .
2	$\theta_k^*(x) = \beta_{k1}x_1, \beta_{k1} \sim \mathcal{N}(0, 1)$	$\pi_k(x) = \begin{cases} x_1 & k = 1, \\ \frac{1}{K-1}(1 - x_1) & k = 2, \dots, K \end{cases}$
3	$\theta_k^*(x) = \varsigma(\beta_{k1}x_1)\varsigma(\beta_{k2}x_2)$ for $\beta_{k1}, \beta_{k2} \sim \mathcal{N}(0, 1)$	$\pi_k(x) = 1/K$ for all $k$ .
4	$\theta_k^*(x) = \varsigma(\beta_k^\top x)$ , for $\beta_k \sim \mathcal{N}_p(\mathbf{0}, \mathbb{I})$	$\pi_k(x) = \begin{cases} x_1 & k = 1, \\ \frac{1}{K-1}(1 - x_1) & k = 2, \dots, K. \end{cases}$
5	$\theta_k^*(x) = \text{RFG}(x)$	$\pi_k(x) = \frac{\exp\{\gamma_k^\top x\}}{\sum_{j=1}^K \exp\{\gamma_j^\top x\}}$ for $\gamma_k \sim \mathcal{N}_p(\mathbf{0}, \mathbb{I})$ .

Table 3: Settings for the underlying treatment effects  $\theta_k^*(\cdot)$  and treatment probabilities  $\pi_k(x)$  for HTE experiments in Section 6. The function  $\varsigma(u) := 1 + (1 + e^{-20(u-1/3)})^{-1}$  is a logistic-type function used in (Athey et al., 2019). The random function generator RFG( $x$ ) is described in Appendix C.4.

such that  $\{\phi_\ell(1), \dots, \phi_\ell(p_\ell)\}$  is a length- $p_\ell$  permutation of indices drawn from  $\{1, \dots, p\}$ , without replacement. The functions  $g_\ell(\cdot)$  are Gaussian functions of the  $p_\ell$  sampled variables:

$$g_\ell(z_\ell) := \exp\left\{-\frac{1}{2}(z_\ell - \mu_\ell)^\top \mathbf{V}_\ell(z_\ell - \mu_\ell)\right\},$$

where the mean vector  $\mu_\ell \in \mathbb{R}^{p_\ell}$  is randomly generated from a standard multivariate Gaussian,  $\mu_\ell \sim \mathcal{N}_{p_\ell}(\mathbf{0}, \mathbb{I})$ . The  $p_\ell \times p_\ell$  covariance matrix  $\mathbf{V}_\ell$  are formed through the spectral decomposition:

$$\mathbf{V}_\ell = \mathbf{U}_\ell \mathbf{D}_\ell \mathbf{U}_\ell^\top,$$

where  $\mathbf{U}_\ell$  is a random  $p_\ell \times p_\ell$  orthonormal matrix and  $\mathbf{D}_\ell := \text{diag}(d_{1,\ell}, \dots, d_{p_\ell,\ell})$  with diagonal entries  $d_{j,\ell}$  generated from a uniform distribution according to  $\sqrt{d_{j,\ell}} \sim \mathcal{U}(0.1, 2.0)$ .

## D. Additional Simulations

### D.1. Settings for the criterion value experiment in Section 3.4

The criterion value experiment in Section 3.4 was run under a varying coefficient model of the form

$$Y_i := W_i^\top \theta^*(X_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 0.5^2), \quad (64)$$

where the regressors  $W_i$  were generated as bivariate standard Gaussian samples  $W_i \sim \mathcal{N}_2(\mathbf{0}, \mathbb{I})$  and the auxiliary covariates were generated as standard uniform samples  $X_i \sim \mathcal{U}(0, 1)$ . The data-generating coefficient functions were  $\theta^*(x) := (\sin(2\pi x), x)$  and the criterion values were computed based on  $n = 1000$  samples following (64).

## D.2. Supporting experiments for Section 6

**Multicollinearity in auxiliary covariates.** We conducted a VCM experiment with highly correlated auxiliary covariate features. We ran a modified version of VCM Setting 3 by generating auxiliary covariates as  $X_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$ , where  $[\Sigma]_{j,k} = \omega^{|j-k|}$  for  $\omega \in \{0, 0.5, 0.9\}$ . Table 4 provides a clear summary of the computational performance of GRF-FPT relative to GRF-grad and statistical accuracy (MSE – multiplied by 100 for readability). All experiments were run over a forest of 10 trees and MSE estimates were computed over 50 replications of the model and evaluated on a separate set of  $n = 5,000$  samples, carried out using GRF-FPT2 and GRF-grad. These results demonstrate clearly that GRF-FPT remains robust, stable, and computationally efficient, even under high multicollinearity in  $X_i$ .

dim( $\mathcal{X}$ )	$n$	$K$	$\omega$	Speedup	$100 \times \text{MSE}_{\text{grad}}$	$100 \times \text{MSE}_{\text{FPT2}}$
5	10,000	64	0.00	2.55	16.60	16.83
5	10,000	64	0.50	2.53	15.48	15.47
5	10,000	64	0.90	2.35	10.95	11.09

Table 4: Effect of multicollinearity in the auxiliary covariates  $X_i$  on the relative computational gain of GRF-FPT2, as well as the statistical accuracy of both GRF-FPT and GRF-grad estimators.

**Subsampling ratio.** We carried out an experiment to show that the subsample proportion does not affect the computational advantage or statistical accuracy of GRF-FPT relative to GRF-grad. We varied the subsampling ratio  $s/n \in \{0.25, 0.50, 0.75\}$  under VCM Setting 3 over a forest of 10 trees carried out using GRF-FPT2 and GRF-grad. Table 5 summarizes our results, averaged over 50 replications of the model, with a test set of 5,000 samples. These results show clearly that the statistical accuracy (MSE) of GRF-FPT2 relative to GRF-grad does not depend strongly on the subsample ratio.

dim( $\mathcal{X}$ )	$n$	$K$	$s/n$	Speedup	$100 \times \text{MSE}_{\text{grad}}$	$100 \times \text{MSE}_{\text{FPT2}}$
2	10,000	64	0.25	2.77	2.86	2.90
2	10,000	64	0.50	3.10	2.91	2.90
2	10,000	64	0.75	2.98	3.21	3.19

Table 5: Effect of the subsampling ratio  $s/n$  on the relative computational gain of GRF-FPT2, as well as the statistical accuracy of both GRF-FPT and GRF-grad estimators.

**Large sample size.** We ran additional experiments to clearly show how our method scales for very large datasets. Using a forest of 10 trees, we tested our method on VCM Setting 3 with sample sizes up to  $n = 500,000$ , carried out using GRF-FPT2 and GRF-grad. The results are summarized in Table 6 and demonstrate that, even as the dataset grows very large, our method consistently remains faster than GRF-grad. While the relative speedup slightly decreases at first, it stabilizes towards a consistent advantage as  $n$  grows sufficiently large, suggesting that the advantage is not bottlenecked by  $n$  and maintains a robust advantage at scale.

dim( $\mathcal{X}$ )	$K$	$n$	Speedup
2	256	10,000	4.54
2	256	20,000	3.59
2	256	50,000	3.49
2	256	100,000	3.11
2	256	200,000	3.04
2	256	500,000	3.08

Table 6: Effect of increasing sample sizes  $n$  on the relative computational gain of GRF-FPT2.

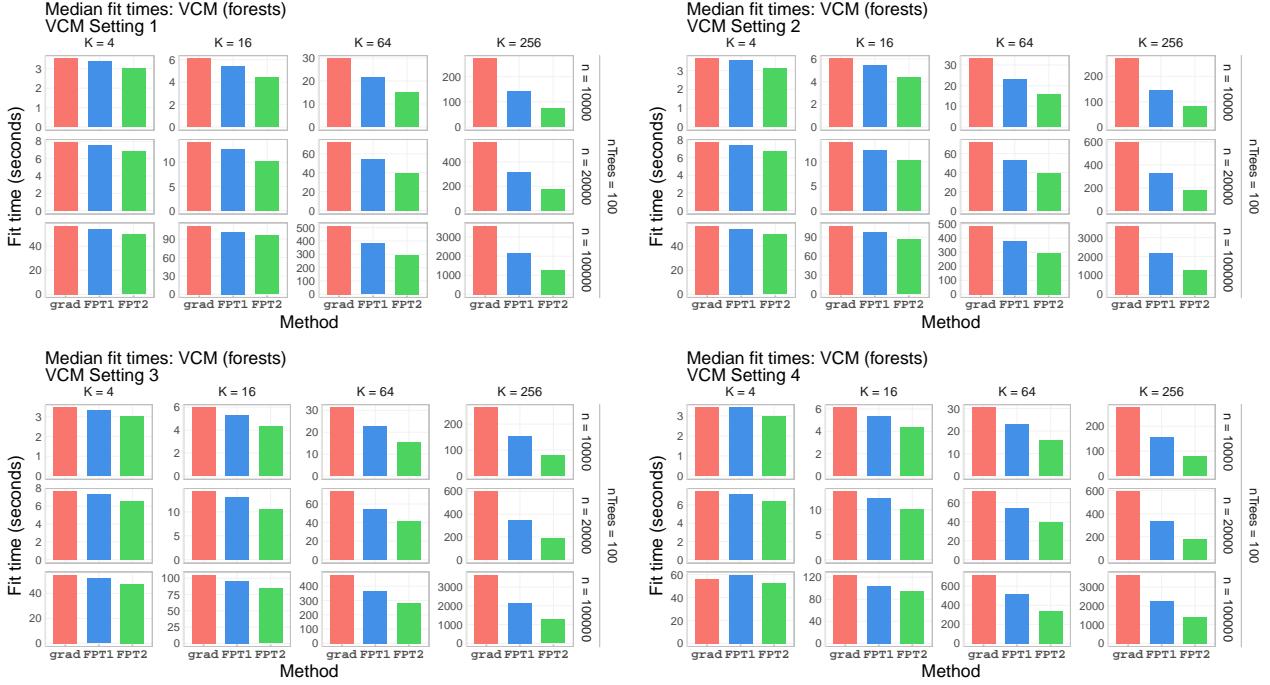


Figure 5: Absolute fit times for VCM timing experiments under the settings in Table 2 and large- $n$  settings in Table 1.

### D.3. Supporting figures for Section 6

#### D.3.1. VCM EXPERIMENTS

**Large  $n$  VCM simulations.** Figure 5 illustrates the absolute fit times for the GRF-FPT algorithms under the four VCM settings for  $\theta_k^*(x)$  described in Table 2 over the large- $n$  settings in Table 1. Across all settings and all dimensions, GRF-FPT is consistently several factors faster than GRF-grad. The speedup factor is summarized in Figure 3, which illustrates the relative speedup of GRF-FPT, calculated as the ratio of GRF-grad fit times over GRF-FPT fit times. Consistent with the observations in Section 5, we find that the speed advantage of GRF-FPT increases as the dimension of the target increases.

Figure 6 shows that this speed advantage comes while performing comparably to GRF-grad in terms of statistical accuracy. Across all settings for VCMs with  $K = 4$  dimensional targets, the MSE estimates from GRF-FPT is highly similar to the MSE estimates of GRF-grad, while for  $K = 256$  dimensional targets one sees more variation in MSE estimates across the methods. This effect likely reflects the increased variance associated with high-dimensional estimation. In some cases we see GRF-FPT1 slightly outperform both GRF-FPT2 and GRF-grad, in other cases we see GRF-grad slightly outperform both GRF-FPT methods, and in others GRF-FPT2 yields the lowest MSE. One sees that these differences are typically small. The key benefit we emphasize is that GRF-FPT is able to achieve nearly identical statistical accuracy with a substantial improvement in computational speed.

**Small  $n$  VCM simulations.** Figures 8 and 7 illustrate the absolute fit times and relative speed advantage, respectively, of GRF-FPT under the VCM design of  $\theta_k^*(x)$  over the small- $n$  settings. One sees that even when  $n$  is more modest, GRF-FPT consistently offers a computational advantage over GRF-grad, with possible outliers under VCM Setting 2 at  $K = 4$ . We believe this negative relative advantage to be caused by random fluctuations in computation and are not representative of the FPT algorithm itself, particularly in light of the fact that the negative effect vanishes when the number of trees increases from 100 to 500. As one would expect based on the large- $n$  results, the relative advantage tends to increase with increasing  $K$ , and generally stabilizes with increasing  $n$ . Figure 9 shows that the GRF-FPT speed advantage does not come at any material cost in statistical accuracy, with similar performance to GRF-grad across all settings.

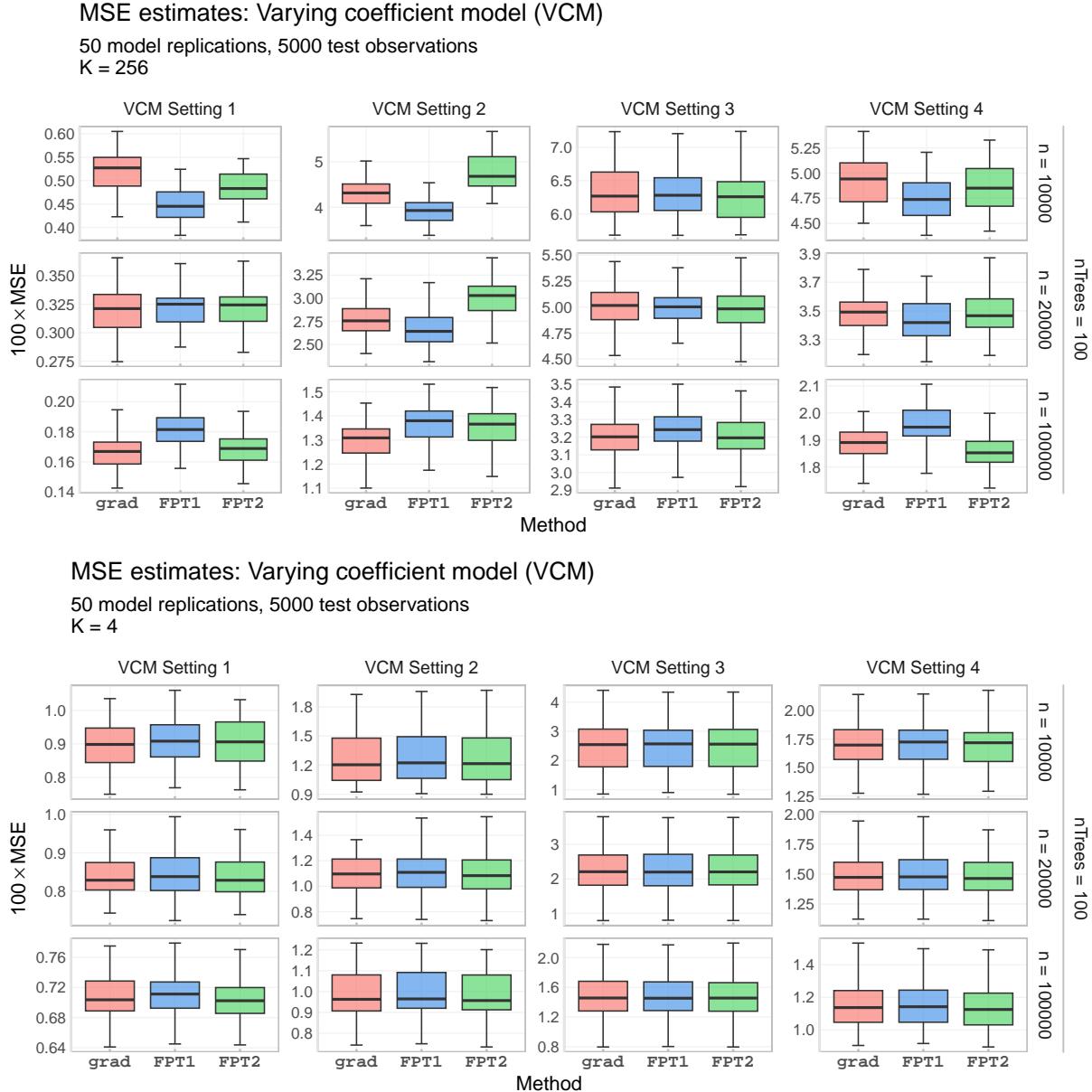


Figure 6: Estimates of  $\text{MSE}[\|\theta^*(X) - \hat{\theta}(X)/K\|_2^2]$  for VCM for  $K = 256$  dimensional (top) and  $K = 4$  dimensional targets (bottom) under the large- $n$  settings in Table 1.

Fit time speedup factor: GRF–grad/GRF–FPT (forests)  
Varying coefficient model (VCM)

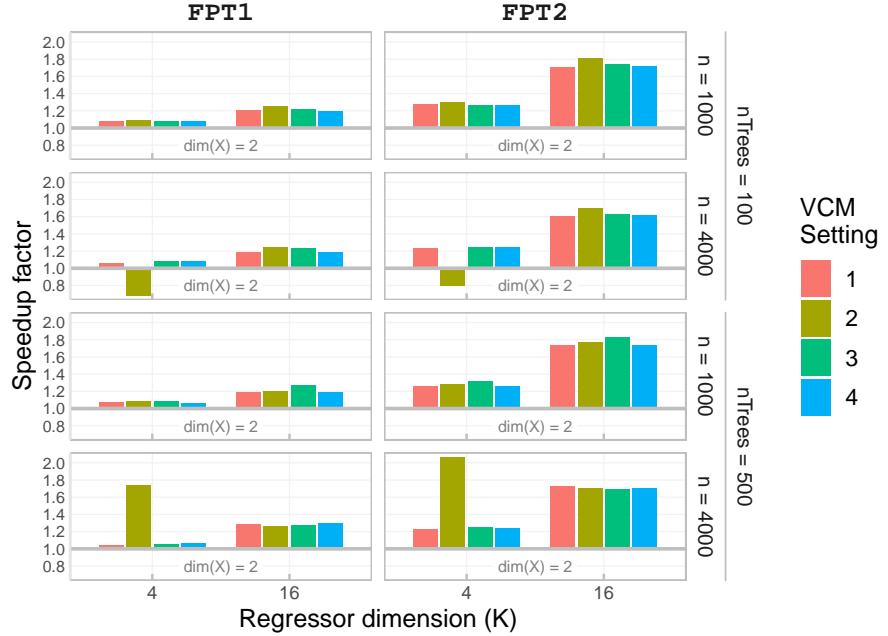


Figure 7: Speedup factor for GRF-FPT in comparison to GRF-grad for VCM timing experiments under the small- $n$  settings in Table 1.

#### D.4. HTE experiments

**Large  $n$  HTE simulations.** Figure 11 illustrates the absolute fit times for the GRF-FPT algorithm under the five HTE settings of  $\theta_k^*(x)$  and  $\pi_k(x)$  described in Table 3 over the large- $n$  settings in Table 1. We find that GRF-FPT is consistently faster than GRF-grad. The speedup factor of GRF-FPT relative to GRF-grad is summarized in Figure 10, calculated as the ratio of GRF-grad fit times over GRF-FPT fit times. As was seen for VCM experiments, the speed advantage of GRF-FPT scales with the dimensionality  $K$  of the target. One sees from both Figures 10 and 11 that GRF-FPT's computational advantage is less dramatic than under the VCM experiments. This can be understood based on the fact that the VCM regressors  $W_i$  are continuous while the HTE regressors represent binary indicators. Continuous regressors provide more granularity when fitting the child statistics  $\tilde{\theta}_{C_j}$ , and as a result provide a larger set of candidate splits over the covariates. Nevertheless, one sees in Figure 10 that the FPT splitting mechanism is still up to  $1.5 \times$  faster under the largest regressor setting  $K = 256$ , with a more modest, but persistent savings across all settings.

The statistical benchmarks for our HTE experiments are shown in Figure 12. Consistent with the VCM experiments, one sees that the computational advantage of GRF-FPT does not come at the cost of its statistical accuracy.

**Small  $n$  HTE simulations.** Figures 13 and 14 summarize the relative speed advantage and absolute fit times for the GRF-FPT algorithm under the small- $n$  HTE design. Consistent with the large- $n$  HTE experiments the FPT2 mechanism sees a stable computational advantage across all settings, with an increasing effect in increasing  $K$ , while the FPT1 mechanism displays a persistent advantage for  $K = 16$  and comparable computational performance for  $K = 4$ . The more modest relative advantage for the small- $n$  experiments is itself consistent with the VCM small- $n$  experiments, owing in large part due to the smaller values of  $K$ . Figure 15 compares the statistical performance of GRF-FPT to GRF-grad, with no material difference between either GRF-FPT1, GRF-FPT2, or GRF-grad's estimation accuracy.

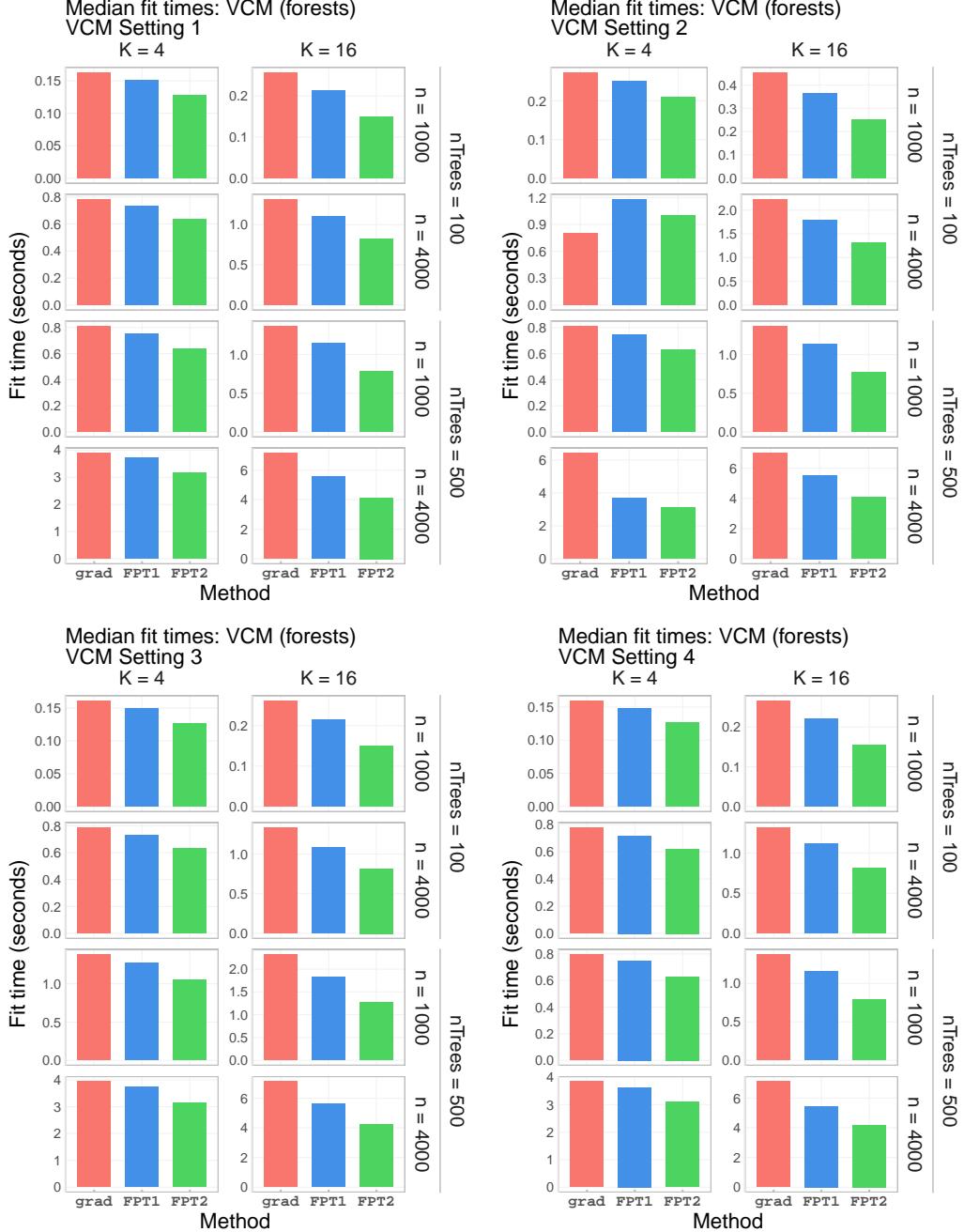
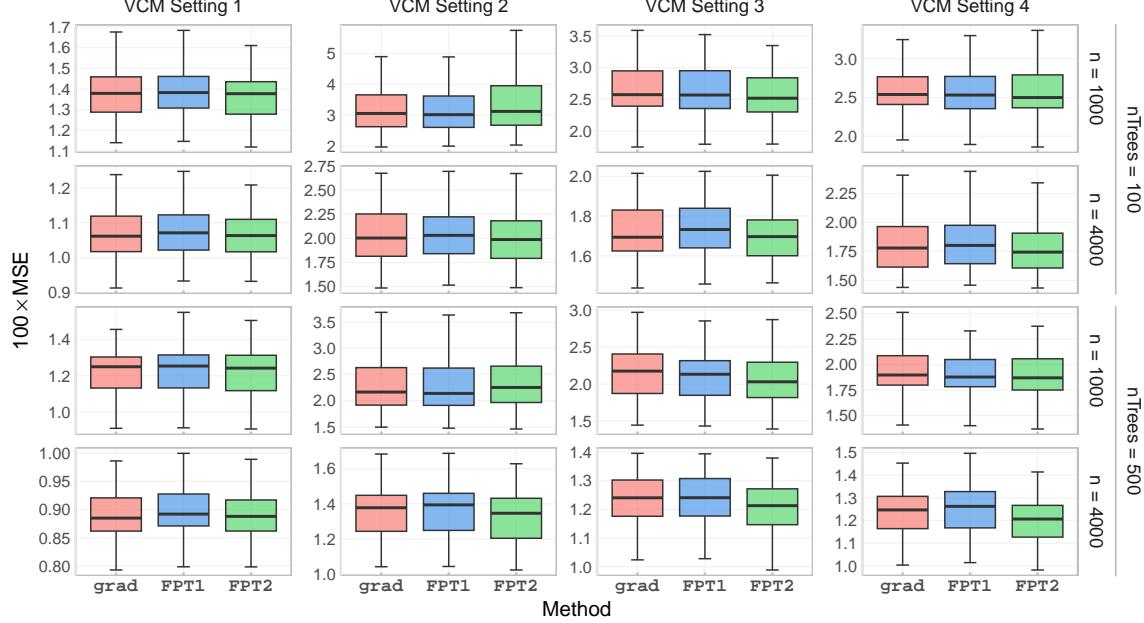


Figure 8: Absolute fit times for VCM timing experiments under the settings in Table 2 and small- $n$  settings in Table 1.

## MSE estimates: Varying coefficient model (VCM)

50 model replications, 5000 test observations  
 $K = 16$



## MSE estimates: Varying coefficient model (VCM)

50 model replications, 5000 test observations  
 $K = 4$

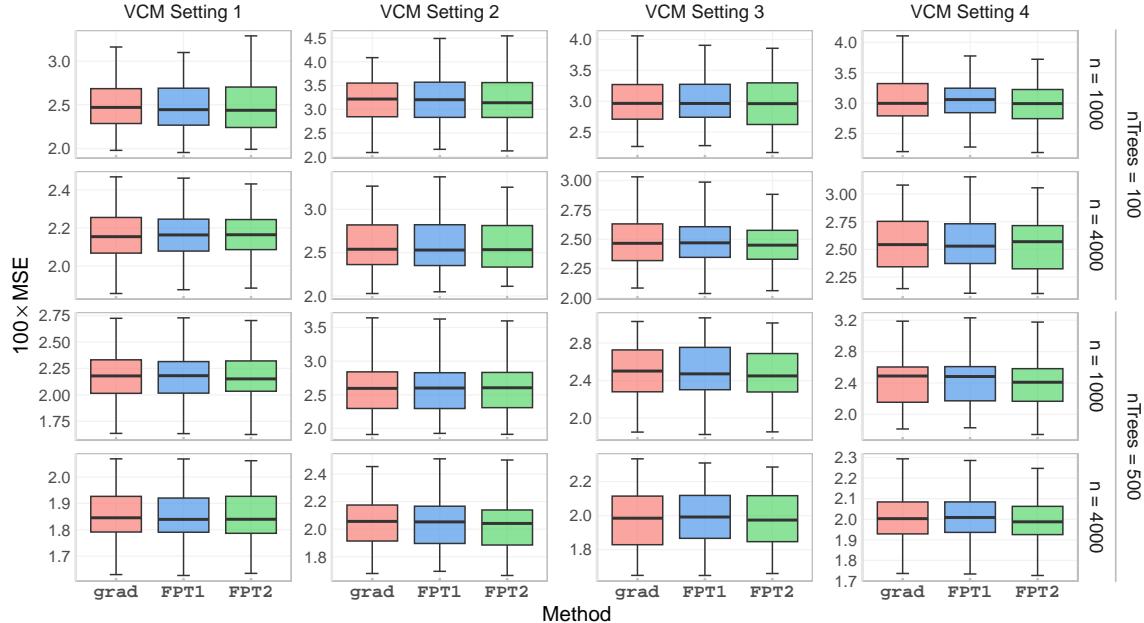


Figure 9: Estimates of  $\text{MSE}[\|\theta^*(X) - \hat{\theta}(X)/K\|_2^2]$  for VCM for  $K = 16$  dimensional (top) and  $K = 4$  dimensional targets (bottom) under the small- $n$  settings in Table 1.

Fit time speedup factor: GRF–grad/GRF–FPT (forests)  
Heterogeneous treatment effects (HTE)

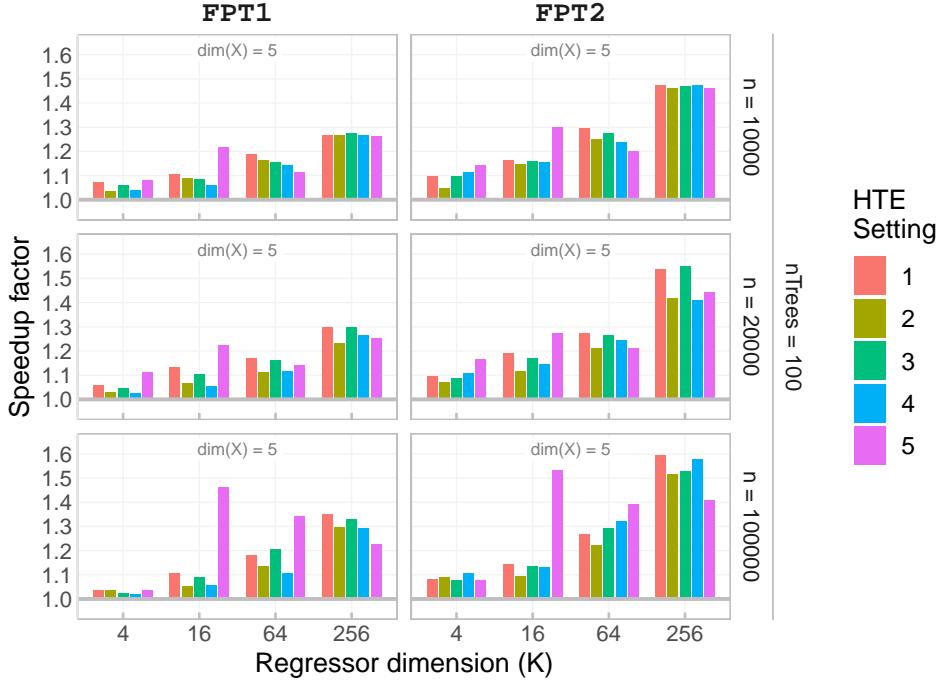


Figure 10: Speedup factor for GRF-FPT in comparison to GRF-grad for HTE timing experiments under the large- $n$  setting in Table 1.

## E. Additional Examples

### E.1. Pseudo-outcomes for nonparametric regression

Consider the task of estimating the conditional mean function  $\theta^*(x) := \mathbb{E}[Y|X = x]$ . The target  $\theta^*(x)$  is identified by a moment condition of the form (1) with scoring function  $\psi_\theta(Y_i) := Y_i - \theta$ , the residual associated with using  $\theta$  as the local estimate with respect to the  $i$ -th sample. The local solution  $\hat{\theta}_P$  over  $P$  is the mean observed response over the parent,

$$\hat{\theta}_P = \bar{Y}_P := \frac{1}{n_P} \sum_{\{i: X_i \in P\}} Y_i.$$

The fixed-point pseudo-outcomes are simple the (negative) residuals that result from fitting (6) over  $P$

$$\rho_i^{\text{FPT}} = -(Y_i - \hat{\theta}_P) = -(Y_i - \bar{Y}_P).$$

The gradient of the score function is  $\nabla_\theta \psi_\theta(y) = -1$ , and hence  $A_P = -1$ . Therefore, up to a constant factor, the gradient-based pseudo-outcomes  $\rho_i^{\text{grad}}$  for conditional mean estimation reduce to their fixed-point counterparts  $\rho_i^{\text{FPT}}$ ,

$$\rho_i^{\text{grad}} = -A_P^{-1} \psi_{\hat{\theta}_P}(Y_i) = Y_i - \bar{Y}_P = -\rho_i^{\text{FPT}}.$$

In this special case, we recover the conventional splitting algorithm used for univariate responses (Breiman et al., 1984; Breiman, 2001) or for multivariate responses (De’ath, 2002; Segal, 1992). Trees grown using  $\rho_i^{\text{FPT}}$ ,  $\rho_i^{\text{grad}}$ , or  $Y_i$  will be identical to one another because CART splits are scale and translation invariant with respect to the response.

More generally, for targets  $\theta^*(x)$  beyond the conditional mean, the form of  $\rho_i^{\text{grad}}$  will be equivalent to  $\rho_i^{\text{FPT}}$  whenever the target function  $\theta^* : \mathcal{X} \rightarrow \Theta$  is a map from the input space  $\mathcal{X}$  a one-dimensional parameter space  $\Theta$ .

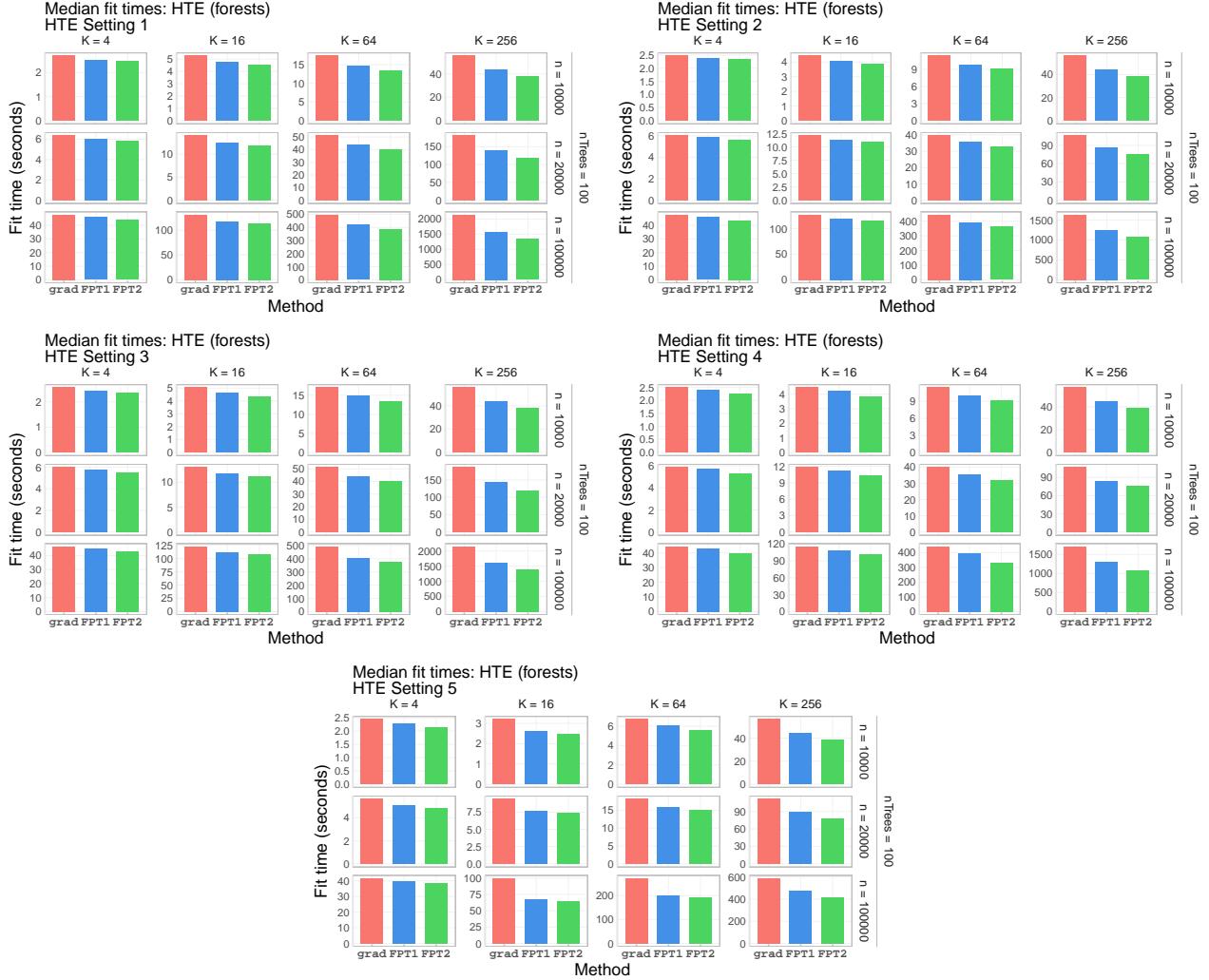


Figure 11: Absolute fit times for HTE timing experiments under the settings in Table 3 and large- $n$  settings in Table 1.

## F. Real Data Comparison: California Housing

**Data.** The California housing data appeared in Kelley Pace & Barry (1997) and can be directly obtained from the Carnegie Mellon StatLib repository (<https://lib.stat.cmu.edu/datasets/houses.zip>). The data includes 20640 observations, where each observation corresponds to measurements over an individual census block group in California taken from the 1990 census. A census block is the smallest geographical area for which the U.S. Census Bureau publishes sample data, typically with a population between 600 and 3000 people per block. Each observation from the California housing data set contains measurements of 9 variables: median housing value (dollars), longitude, latitude, median housing age (years), total rooms (count, aggregated over the census block), total bedrooms (count, aggregated over the census block), population (count), households (count), median income (dollars).

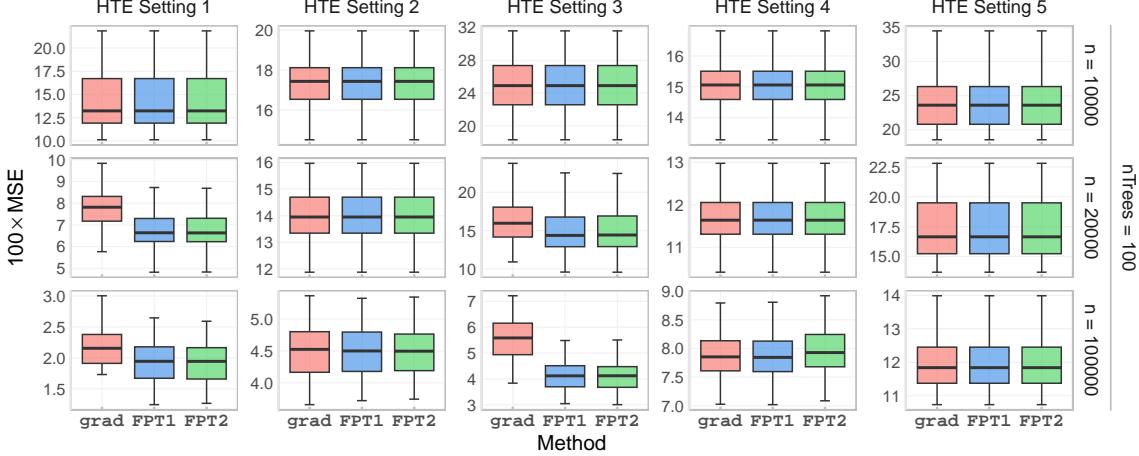
**Model.** We consider a varying coefficient model of the form

$$Y_i = \nu^*(X_i) + \theta_1^*(X_i)W_{i,1} + \cdots + \theta_6^*(X_i)W_{i,6} + \epsilon_i \quad (65)$$

where we suppose that our effects are local to spatial coordinates  $x := (\text{latitude}_i, \text{longitude}_i)$ ,  $Y_i$  denotes the log median housing value of the census block, and the primary regressors  $W_i = (W_{i,1}, \dots, W_{i,6})$  passed to the model were median housing age, log(total rooms), log(total bedrooms), log(population), log(households), and log(median income). Here, each  $\theta_k^*(x)$  denotes the geographically-varying effect of the corresponding regressor  $W_{i,k}$ , for  $k = 1, \dots, 6$ . The empirical distribution of the transformed regressors passed to each of the GRF models is seen in Figure 17.

## MSE estimates: Heterogeneous treatment effects (HTE)

50 model replications, 5000 test observations  
 $K = 256$



## MSE estimates: Heterogeneous treatment effects (HTE)

50 model replications, 5000 test observations  
 $K = 4$

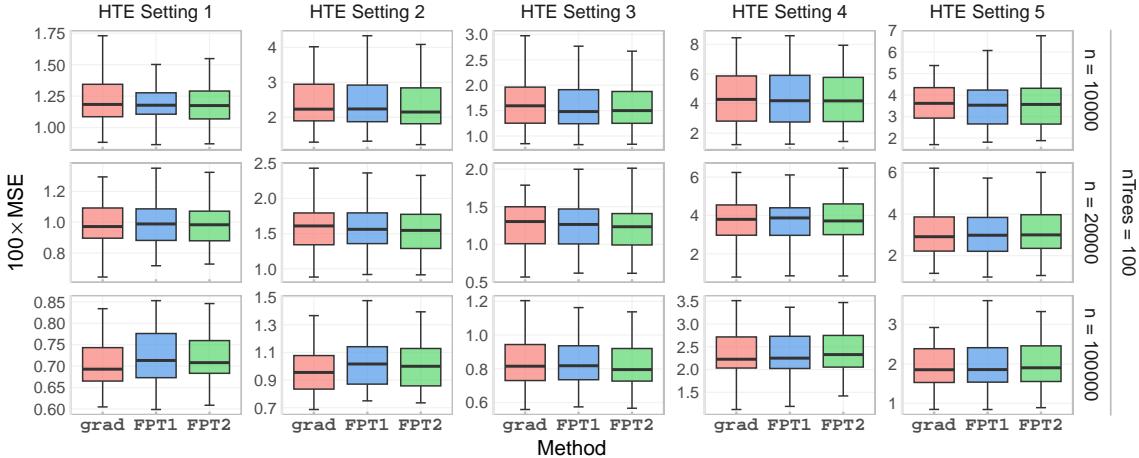


Figure 12: Estimates of MSE  $\mathbb{E}[\|\theta^*(X) - \hat{\theta}(X)/K\|_2^2]$  for HTE for  $K = 256$  dimensional (top) and  $K = 4$  dimensional targets (bottom) under the large- $n$  settings in Table 1.

**Algorithms.** We target GRF estimates  $\hat{\theta}(x) = (\hat{\theta}_1(x), \dots, \hat{\theta}_6(x))^\top$  of  $\theta^*(x) = (\theta_1^*(x), \dots, \theta_6^*(x))^\top$  based on the GRF-**FPT1** and GRF-**FPT2** algorithms described in Section 6, and compare those to GRF-grad. All forests were fit using the `grf::lm_forest` function, which trains the Stage I forest and optionally solves for the Stage II estimates  $\hat{\theta}(x)$  for varying coefficient models (65). All versions fit a forest of 2000 trees, the default settings of the original R implementation (Tibshirani et al., 2024), a subsample ratio of 0.5, and a target minimum node size of 5 observations.

Algorithm	Training time (sec.)	Speedup factor
GRF-grad	19.1	
GRF- <b>FPT1</b>	15.4	1.24
GRF- <b>FPT2</b>	12.6	1.52

Table 7: Fit times to train a forest of 2000 trees on the California housing data.

Fit time speedup factor: GRF–grad/GRF–FPT (forests)  
 Heterogeneous treatment effects (HTE)

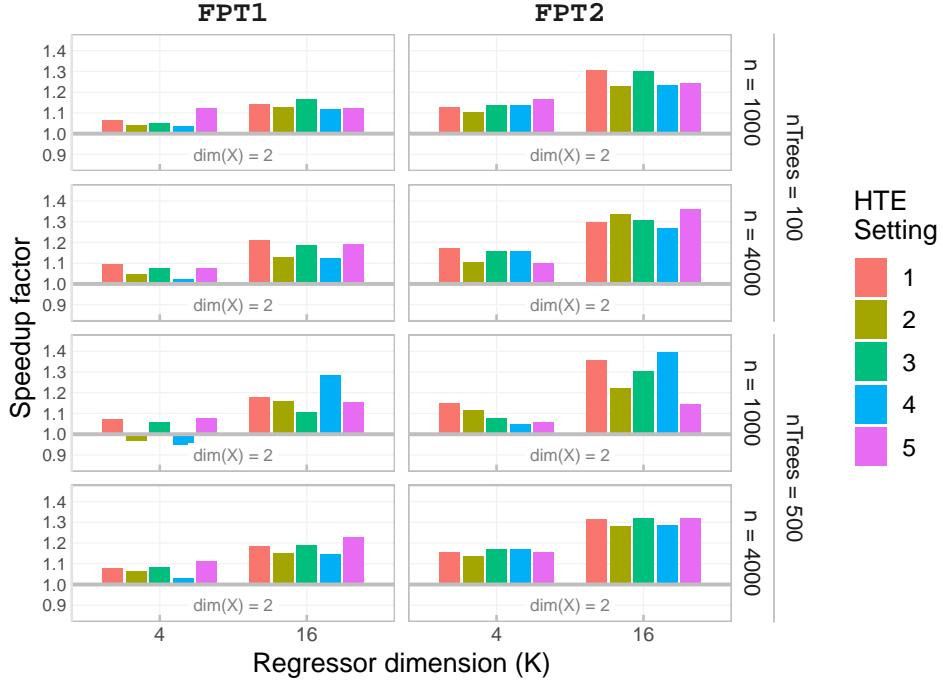


Figure 13: Speedup factor for GRF-FPT in comparison to GRF-grad for HTE timing experiments under the small- $n$  settings in Table 1.

**Results.** Table 7 summarizes the computational benefit of GRF-FPT applied to the California housing data. Figures 4 illustrates the local estimates  $\hat{\theta}(x)$  made by GRF-FPT2, while Figure 16 illustrates the fits under GRF-FPT1 and GRF-grad.

## Generalized Random Forests using Fixed-Point Trees

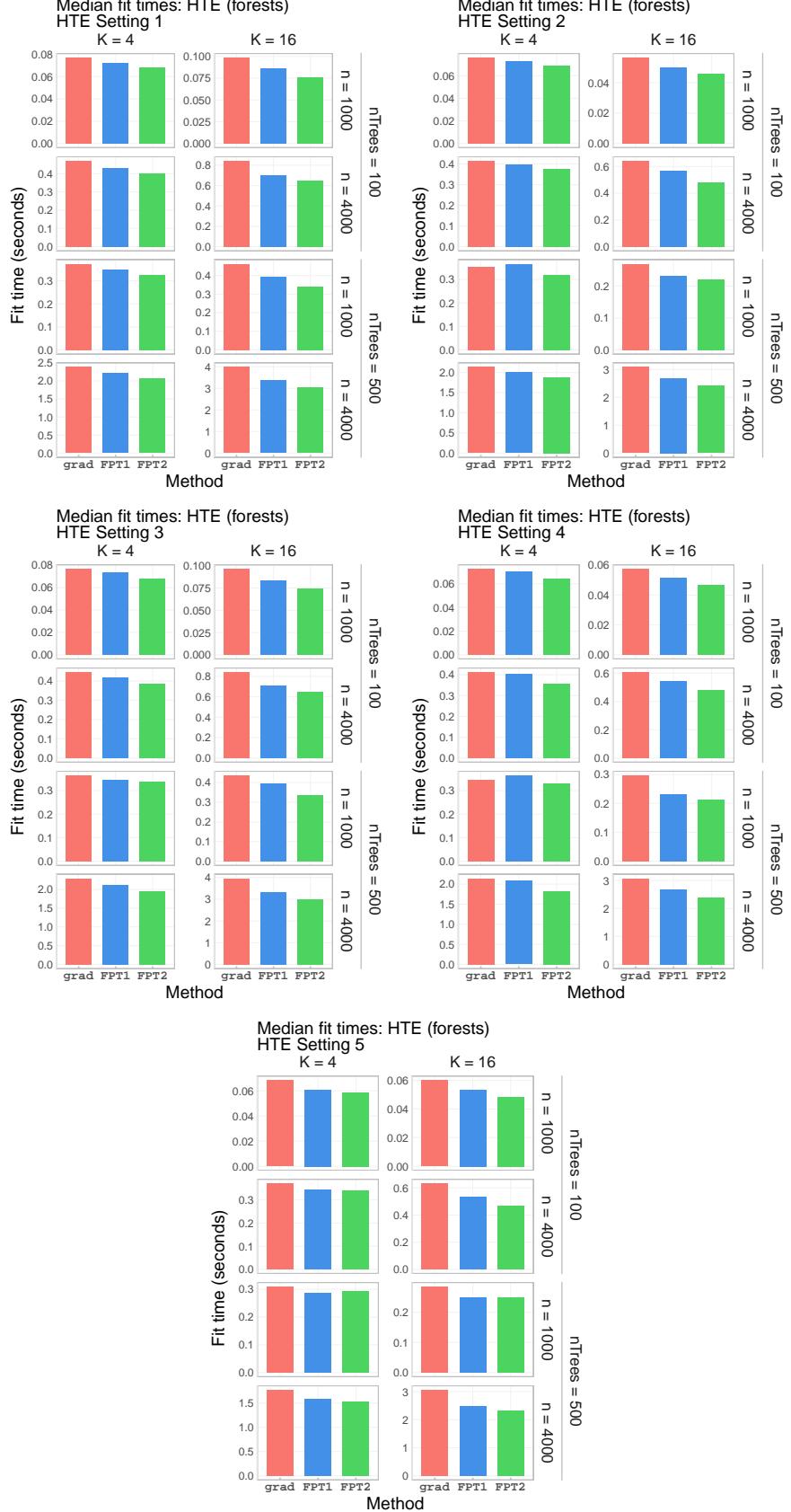
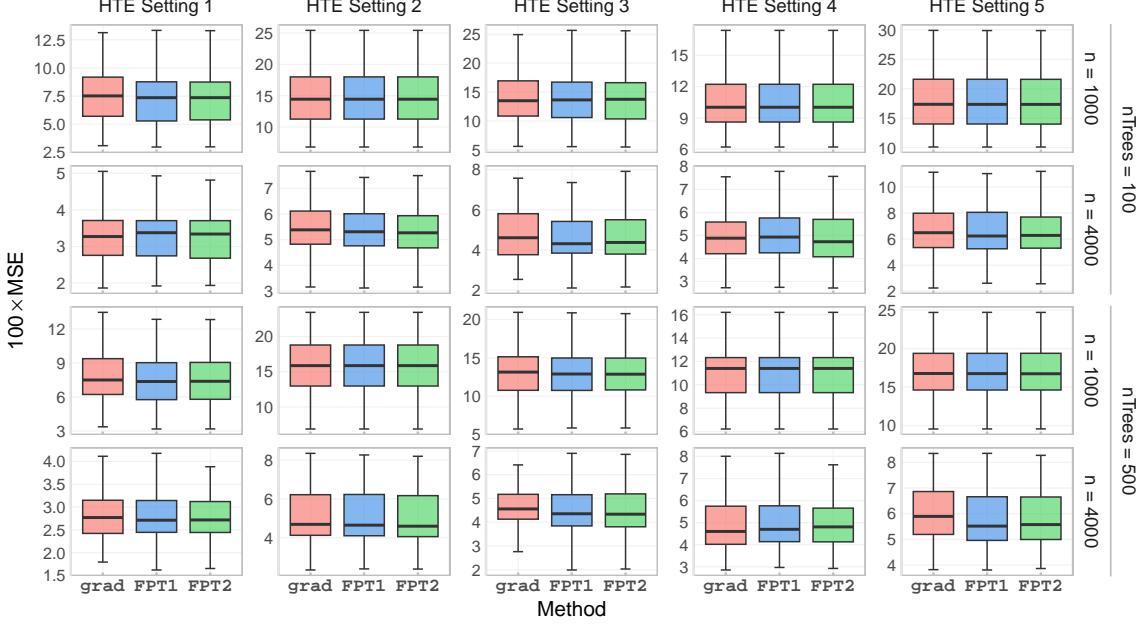


Figure 14: Absolute fit times for HTE timing experiments under the settings in Table 3 and small- $n$  settings in Table 1.

## MSE estimates: Heterogeneous treatment effects (HTE)

50 model replications, 5000 test observations  
 $K = 16$



## MSE estimates: Heterogeneous treatment effects (HTE)

50 model replications, 5000 test observations  
 $K = 4$

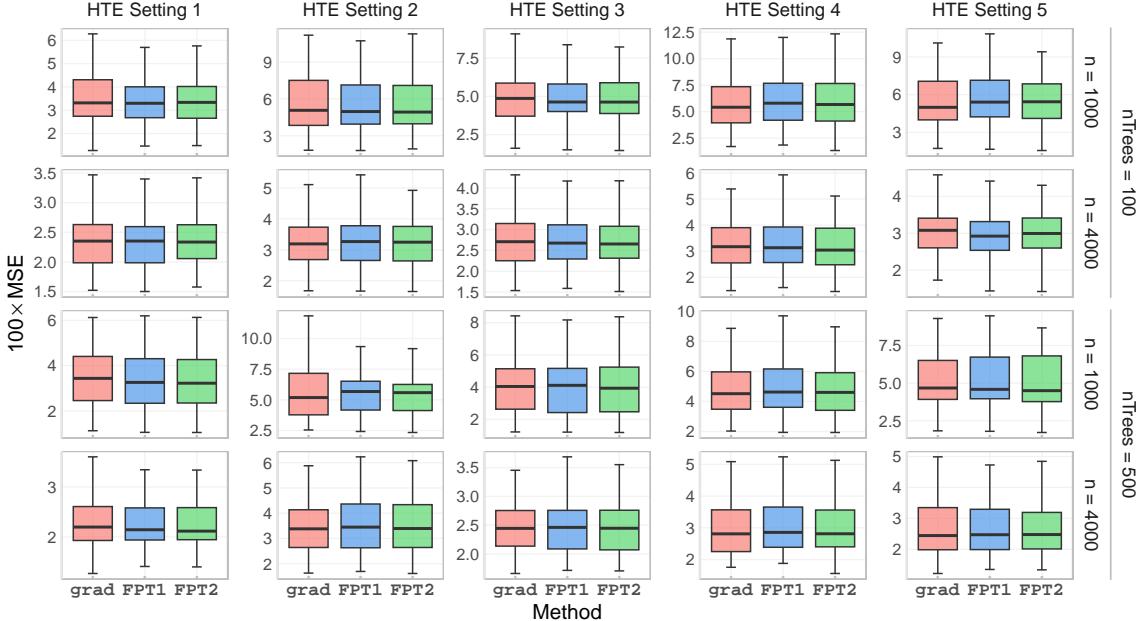


Figure 15: Estimates of  $\text{MSE} \mathbb{E}[\|\theta^*(X) - \hat{\theta}(X)/K\|_2^2]$  for HTE for  $K = 16$  dimensional (top) and  $K = 4$  dimensional targets (bottom) under the small- $n$  settings in Table 1.

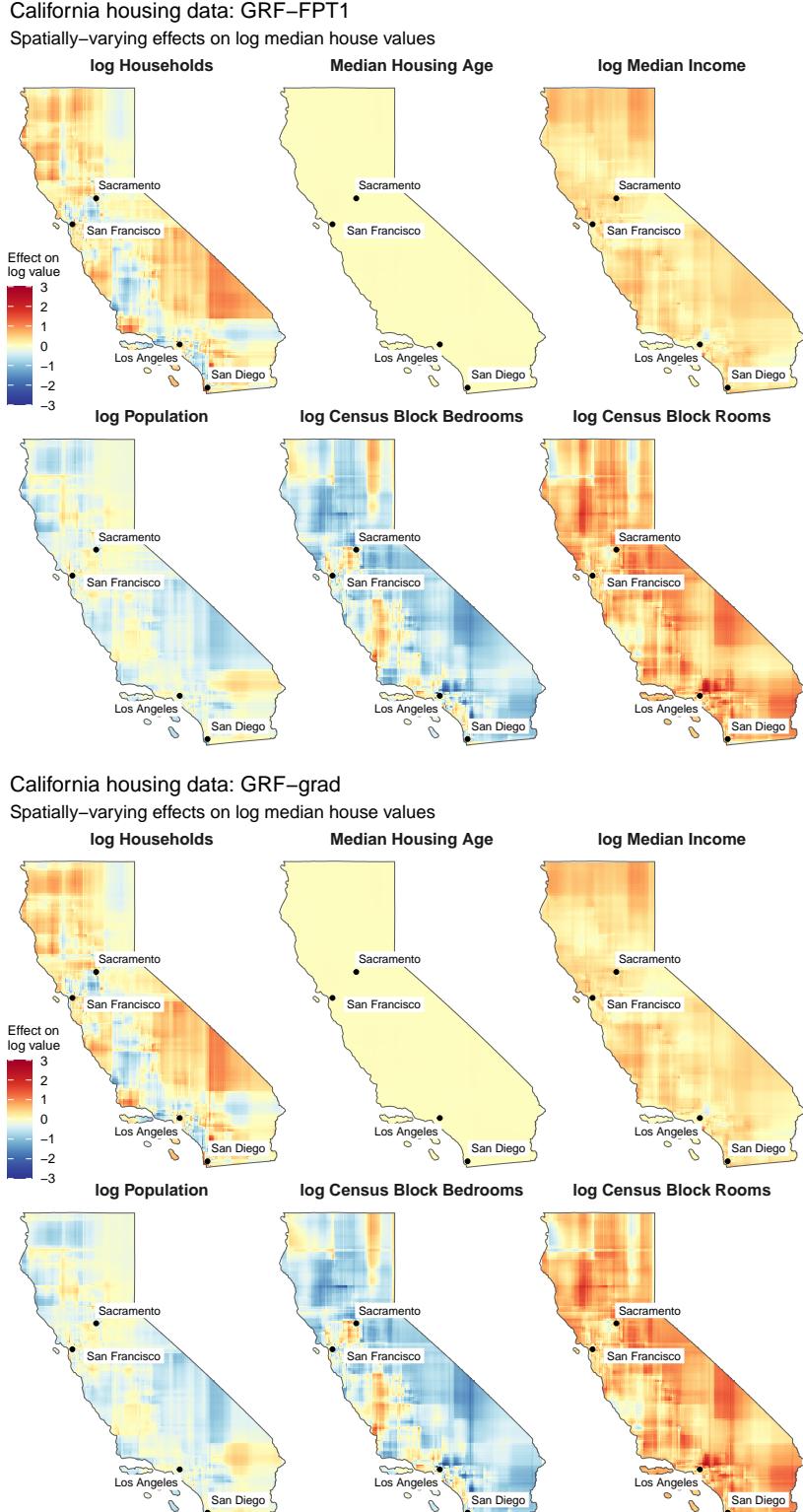


Figure 16: Geographically-varying local estimates  $\hat{\theta}(x) = (\hat{\theta}_1(x), \dots, \hat{\theta}_6(x))$ , fit under GRF-FPT1 (top) and GRF-grad (bottom). Results for GRF-FPT2 are presented in Figure 4 found in Section 7.

## California housing data: Regressor distribution

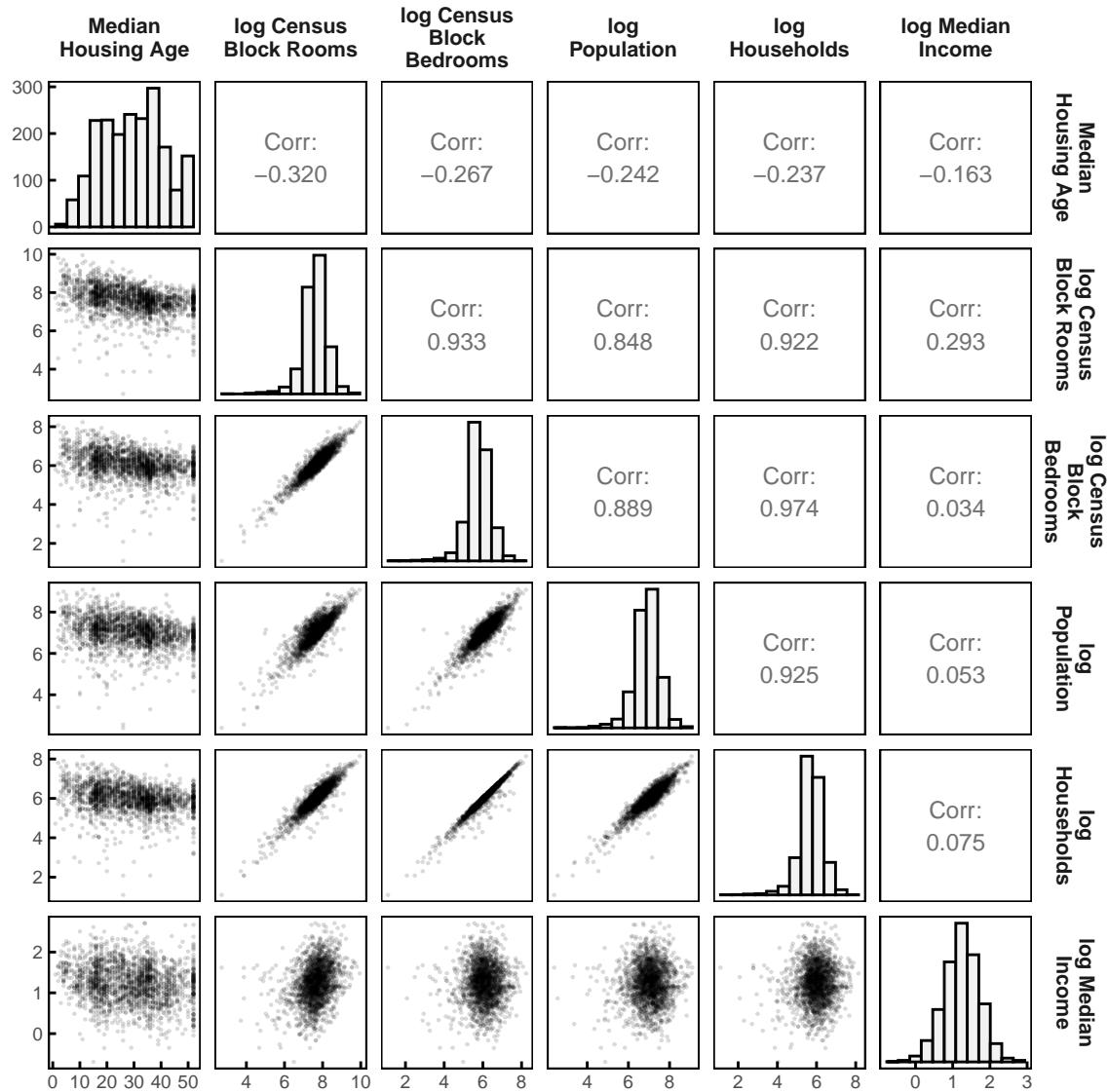


Figure 17: Empirical distribution of the regressors from the California housing data passed to GRF.