Orthogonal Random Forest for Causal Inference

Miruna Oprescu [1] **Vasilis Syrgkanis** [1] **Zhiwei Steven Wu** [2]

Abstract

We propose the *orthogonal random forest*, an algorithm that combines *Neyman-orthogonality* to reduce sensitivity with respect to estimation error of nuisance parameters with generalized random forests (Athey et al., 2017)—a flexible non-parametric method for statistical estimation of conditional moment models using random forests. We provide a consistency rate and establish asymptotic normality for our estimator. We show that under mild assumptions on the consistency rate of the nuisance estimator, we can achieve the same error rate as an oracle with a priori knowledge of these nuisance parameters. We show that when the nuisance functions have a locally sparse parametrization, then a local ℓ_1-penalized regression achieves the required rate. We apply our method to estimate heterogeneous treatment effects from observational data with discrete treatments or continuous treatments, and we show that, unlike prior work, our method provably allows to control for a high-dimensional set of variables under standard sparsity conditions. We also provide a comprehensive empirical evaluation of our algorithm on both synthetic and real data.

1. Introduction

Many problems that arise in causal inference can be formulated in the language of conditional moment models: given a target feature x find a solution $\theta_0(x)$ to a system of conditional moment equations

$$\mathbb{E}\left[\psi(Z;\theta,h_0(x,W)) \mid X=x\right]=0, \tag{1}$$

given access to n i.i.d. samples from the data generating distribution, where ψ is a known score function and h_0 is an unknown nuisance function that also needs to be estimated from data. Examples include non-parametric regression, heterogeneous treatment effect estimation, instrumental variable regression, local maximum likelihood estimation and estimation of structural econometric models.[1] The study of such conditional moment restriction problems has a long history in econometrics (see e.g. Newey (1993); Ai & Chen (2003); Chen & Pouzo (2009); Chernozhukov et al. (2015)).

In this general estimation problem, the main goal is to estimate the target parameter at a rate that is robust to the estimation error of the nuisance component. This allows the use of flexible models to fit the nuisance functions and enables asymptotically valid inference. Almost all prior work on the topic has focused on two settings: i) they either assume the target function $\theta_0(x)$ takes a parametric form and allow for a potentially high-dimensional parametric nuisance function, e.g. (Chernozhukov et al., 2016; 2017; 2018), ii) or take a non-parametric stance at estimating $\theta_0(x)$ but do not allow for high-dimensional nuisance functions (Wager & Athey, 2015; Athey et al., 2017).

We propose *Orthogonal Random Forest* (ORF), a random forest-based estimation algorithm, which performs non-parametric estimation of the target parameter while permitting more complex nuisance functions with high-dimensional parameterizations. Our estimator is also asymptotically normal and hence allows for the construction of asymptotically valid confidence intervals via plug-in or bootstrap approaches. Our approach combines the notion of *Neyman orthogonality* of the moment equations with a two-stage random forest based algorithm, which generalizes prior work on *Generalized Random Forests* (Athey et al., 2017) and the double machine learning (double ML) approach proposed in (Chernozhukov et al., 2017). To support our general algorithm, we also provide a novel nuisance estimation algorithm—*Forest Lasso*—that effectively recovers high-dimensional nuisance parameters provided they have locally sparse structure. This result combines techniques from Lasso theory (Hastie et al., 2015) with concentration inequalities for U-statistics (Hoeffding, 1963).

As a concrete example and as a main application of our approach, we consider the problem of *heterogeneous treat-*

*Equal contribution [1]Microsoft Research–New England [2]University of Minnesota–Twin Cities. Correspondence to: Miruna Oprescu <moprescu@microsoft.com >, Vasilis Syrgkanis <vasy@microsoft.com>, Zhiwei Steven Wu <zsw@umn.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[1] See e.g. Reiss & Wolak (2007) and examples in Chernozhukov et al. (2016; 2018)

ment effect estimation. This problem is at the heart of many decision-making processes, including clinical trial assignment to patients, price adjustments of products, and ad placement by a search engine. In many situations, we would like to take the heterogeneity of the population into account and estimate the *heterogeneous treatment effect* (*HTE*)—the effect of a treatment T (e.g. drug treatment, price discount, and ad position), on the outcome Y of interest (e.g. clinical response, demand, and click-through-rate), as a function of observable characteristics x of the treated subject (e.g. individual patient, product, and ad). HTE estimation is a fundamental problem in causal inference from observational data (Imbens & Rubin, 2015; Wager & Athey, 2015; Athey et al., 2017), and is intimately related to many areas of machine learning, including contextual bandits, off-policy evaluation and optimization (Swaminathan et al., 2016; Wang et al., 2017; Nie & Wager, 2017), and counterfactual prediction (Swaminathan & Joachims, 2015; Hartford et al., 2016).

The key challenge in HTE estimation is that the observations are typically collected by a policy that depends on confounders or control variables W, which also directly influence the outcome. Performing a direct regression of the outcome Y on the treatment T and features x, without controlling for a multitude of other potential confounders, will produce biased estimation. This leads to a regression problem that in the language of conditional moments takes the form:

$$\mathbb{E}\left[Y - \theta_0(x)T - f_0(x, W) \mid X = x\right] = 0 \qquad (2)$$

where $\theta_0(x)$ is the heterogeneous effect of the treatment T (discrete or continuous) on the outcome Y as a function of the features x and $f_0(x, W)$ is an unknown nuisance function that captures the direct effect of the control variables on the outcome. Moreover, unlike active experimentation settings such as contextual bandits, when dealing with observational data, the actual treatment or logging policy $\mathbb{E}\left[T|x, W\right] = g_0(x, W)$ that could potentially be used to de-bias the estimation of $\theta_0(x)$ is also unknown.

There is a surge of recent work at the interplay of machine learning and causal inference that studies efficient estimation and inference of treatment effects. Chernozhukov et al. (2017) propose a two-stage estimation method called *double machine learning* that first orthogonalizes out the effect of high-dimensional confounding factors using sophisticated machine learning algorithms, including Lasso, deep neural nets and random forests, and then estimates the effect of the lower dimensional treatment variables, by running a low-dimensional linear regression between the residualized treatments and residualized outcomes. They show that even if the estimation error of the first stage is not particularly accurate, the second-stage estimate can still be $n^{-1/2}$-asymptotically normal. However, their approach requires a parametric specification of $\theta_0(x)$. In contrast, another line of work that brings machine learning to causal inference provides fully flexible non-parametric HTE estimation based on random forest techniques (Wager & Athey, 2015; Athey et al., 2017; Powers et al., 2017). However, these methods heavily rely on low-dimensional assumptions.

Our algorithm ORF, when applied to the HTE problem (see Section 6) allows for the non-parametric estimation of $\theta_0(x)$ via forest based approaches while simultaneously allowing for a high-dimensional set of control variables W. This estimation problem is of practical importance when a decision maker (DM) wants to optimize a policy that depends only on a small set of variables, e.g. due to data collection or regulatory constraints or due to interpretability of the resulting policy, while at the same time controlling for many potential confounders in the existing data that could lead to biased estimates. Such settings naturally arise in contextual pricing or personalized medicine. In such settings the DM is faced with the problem of estimating a conditional average treatment effect conditional on a small set of variables while controlling for a much larger set. Our estimator provably offers a significant statistical advantage for this task over prior approaches.

In the HTE setting, the ORF algorithm follows the residual-on-residual regression approach analyzed by (Chernozhukov et al., 2016) to formulate a locally Neyman orthogonal moment and then applies our orthogonal forest algorithm to this orthogonal moment. Notably, (Athey et al., 2017) also recommend such a residual on residual regression approach in their empirical evaluation, which they refer to as "local centering", albeit with no theoretical analysis. Our results provide a theoretical foundation of the local centering approach through the lens of Neyman orthogonality. Moreover, our theoretical results give rise to a slightly different overall estimation approach than the one in (Athey et al., 2017): namely we residualize locally around the target estimation point x, as opposed to performing an overall residualization step and then calling the Generalized Random Forest algorithm on the residuals. The latter stems from the fact that our results require that the nuisance estimator achieve a good estimation rate *only* around the target point x. Hence, residualizing locally seems more appropriate than running a global nuisance estimation, which would typically minimize a non-local mean squared error. Our experimental findings reinforce this intuition (see e.g. comparison between ORF and the GRF-Res benchmark). Another notable work that combines the residualization idea with flexible heterogeneous effect estimation is that of (Nie & Wager, 2017), who formulate the problem as an appropriate residual-based square loss minimization over an arbitrary hypothesis space for the heterogeneous effect function $\theta(x)$. Formally, they show robustness, with respect to nuisance estimation errors, of the mean squared error (MSE) of the resulting estimate in expectation over the distribution X and for the case where

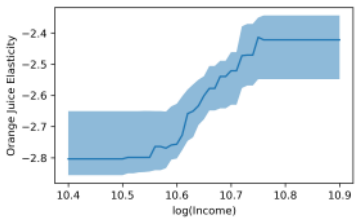

Figure 1: ORF estimates for the effect of orange juice price on demand from a high-dimensional dataset. We depict the estimated heterogeneity in elasticity by income level. The shaded region depicts the 1%-99% confidence interval obtained via bootstrap.

the hypothesis space is a reproducing kernel Hilbert space (RKHS). Our work differs primarily by: i) focusing on sup-norm estimation error at any target point x as opposed to MSE, ii) using forest based estimation as opposed to finding a function in an RKHS, iii) working with the general orthogonal conditional moment problems, and iv) providing asymptotic normality results and hence valid inference.

We provide a comprehensive empirical comparison of ORF with several benchmarks, including three variants of GRF. We show that by setting the parameters according to what our theory suggests, ORF consistently outperforms all of the benchmarks. Moreover, we show that bootstrap based confidence intervals provide good finite sample coverage.

Finally, to motivate the usage of the ORF, we applied our technique to Dominick's dataset, a popular historical dataset of store-level orange juice prices and sales provided by University of Chicago Booth School of Business. The dataset is comprised of a large number of covariates W, but economics researchers might only be interested in learning the elasticity of demand as a function of a few variables x such as income or education. We applied our method (see Appendix G for details) to estimate orange juice price elasticity as a function of income, and our results, depicted in Figure 1, unveil the natural phenomenon that lower income consumers are more price-sensitive.

2. Estimation via Local Orthogonal Moments

We study non-parametric estimation of models defined via conditional moment restrictions, in the presence of nuisance functions. Suppose we have a set of $2n$ observations $Z_1, \dots, Z_{2n}$ drawn independently from some underlying distribution $\mathcal{D}$ over the observation domain $\mathcal{Z}$. Each observation Z_i contains a feature vector $X_i \in \mathcal{X} := [0,1]^d$.

Given a target feature $x \in \mathcal{X}$, our goal is to estimate a parameter vector $\theta_0(x) \in \mathbb{R}^p$ that is defined via a local moment condition, i.e. for all $x \in \mathcal{X}$, $\theta_0(x)$ is the unique solution with respect to θ of:

$$\mathbb{E}\left[\psi(Z; \theta, h_0(x, W)) \mid X = x\right] = 0, \tag{3}$$

where $\psi\colon \mathcal{Z} \times \mathbb{R}^p \times \mathbb{R}^\ell \to \mathbb{R}^p$ is a score function that maps an observation Z, parameter vector $\theta(x) \in \Theta \subset \mathbb{R}^p$, and nuisance vector $h(x,w)$ to a vector-valued score $\psi(z; \theta(x), h(x,w))$ and $h_0 \in H \subseteq \left(\mathbb{R}^d \times \mathbb{R}^L \to \mathbb{R}^\ell\right)$ is an unknown nuisance function that takes as input X and a subvector W of Z, and outputs a nuisance vector in $\mathbb{R}^\ell$. For any feature $x \in \mathcal{X}$, parameter $\theta \in \Theta$, and nuisance function $h \in H$, we define the *moment* function as:

$$m(x; \theta, h) = \mathbb{E}\left[\psi(Z; \theta, h(X, W)) \mid X = x\right] \tag{4}$$

We assume that the dimensions p, ℓ, d are constants, while the dimension L of W can be growing with n.

We will analyze the following two-stage estimation process.

1. *First stage.* Compute a nuisance estimate $\hat{h}$ for h_0 using data $\{Z_{n+1}, \dots, Z_{2n}\}$ with some guarantee on the conditional root mean squared error:[2]

$$\mathcal{E}(\hat{h}) = \sqrt{\mathbb{E}\left[\|\hat{h}(x, W) - h_0(x, W)\|^2 \mid X = x\right]}$$

2. *Second stage.* Compute a set of similarity weights $\{a_i\}$ over the data $\{Z_1, \dots, Z_n\}$ that measure the similarity between their feature vectors X_i and the target x. Compute the estimate $\hat{\theta}(x)$ using the nuisance estimate $\hat{h}$ via the plug-in weighted moment condition:

$$\hat{\theta}(x) \text{ solves: } \textstyle\sum_{i=1}^n a_i \psi(Z_i; \theta, \hat{h}(X_i, W_i)) = 0 \tag{5}$$

In practice, our framework permits the use of any method to estimate the nuisance function in the first stage. However, since our description is a bit too abstract let us give a special case, which we will also need to assume for our normality result. Consider the case when the nuisance function h takes the form $h(x,w) = g(w; \nu(x))$, for some known function g but unknown function $\nu : \mathcal{X} \to \mathbb{R}^{d_\nu}$ (with d_ν potentially growing with n), i.e. locally around each x the function h is a parametric function of w. Moreover, the parameter $\nu_0(x)$ of the true nuisance function h_0 is identified as the minimizer of a local loss:

$$\nu_0(x) = \operatorname{argmin}_{\nu \in \mathcal{V}} \mathbb{E}\left[\ell(Z; \nu) \mid X = x\right] \tag{6}$$

Then we can estimate $\nu_0(x)$ via a locally weighted and penalized empirical loss minimization algorithm. In particular in Section 5 we will consider the case of local ℓ_1-penalized estimation that we will refer to as *forest lasso* and which provides formal guarantees in the case where $\nu_0(x)$ is sparse.

The key technical condition that allows us to reliably perform the two-stage estimation is the following *local orthogonality* condition, which can be viewed as a localized version of the *Neyman orthogonality* condtion (Neyman,

[2] Throughout the paper we denote with $\|\cdot\|$ the euclidean norm and with $\|\cdot\|_p$ the p-norm.

1979; Chernozhukov et al., 2017) around the neighborhood of the target feature x. Intuitively, the condition says that the score function ψ is insensitive to local perturbations in the nuisance parameters around their true values.

Definition 2.1 (Local Orthogonality)**.** Fix any estimator $\hat{h}$ for the nuisance function. Then the Gateaux derivative with respect to h, denoted $D_\psi[\hat{h} - h_0 \mid x]$, is defined as:

$$\mathbb{E}\left[\nabla_h \psi(Z, \theta_0(x), h_0(x, W))(\hat{h}(x, W) - h_0(x, W)) \mid x\right]$$

where ∇_h denotes the gradient of ψ with respect to the final ℓ arguments. We say that the moment conditions are *locally orthogonal* if for all x: $D_\psi[\hat{h} - h_0 \mid x] = 0$.

3. Orthogonal Random Forest

We describe our main algorithm *orthogonal random forest* (ORF) for calculating the similarity weights in the second stage of the two stage estimation. In the next section we will see that we will be using this algorithm for the estimation of the nuisance functions, so as to perform a local nuisance estimation. At a high level, ORF can be viewed as an orthogonalized version of GRF that is more robust to the nuisance estimation error. Similar to GRF, the algorithm runs a tree learner over B random *subsamples* S_b *(without replacement) of size* $s < n$, to build B trees such that each tree indexed by b provides a tree-based weight a_{ib} for each observation Z_i in the input sample. Then the ORF weight a_i for each sample i is the average over the tree-weights a_{ib}.

The tree learner starts with a root node that contains the entire $\mathcal{X}$ and recursively grows the tree to split $\mathcal{X}$ into a set of leaves until the number of observations in each leaf is not too small. The set of neighboods defined by the leaves naturally gives a simlarity measure between each observation and the target x. Following the same approach of (Tibshirani et al., 2018; Wager & Athey, 2015), we maintain the following tree properties in the process of building a tree.

Specification 1 (Forest Regularity)**.** *The tree satisfies*

- ***Honesty****: we randomly partition the input sample S into two subsets S^1, S^2, then uses S^1 to place splits in the tree, and uses S^2 for estimation.*
- ***ρ-balanced****: each split leaves at least a fraction ρ of the observations in S^2 on each side of the split for some parameter of $\rho \leq 0.2$.*
- ***Minimum leaf size*** *r: there are between r and $2r - 1$ observations from S^2 in each leaf of the tree.*
- ***π-random-split****: at every step, marginalizing over the internal randomness of the learner, the probability that the next split occurs along the j-th feature is at least π/d for some $0 < \pi \leq 1$, for all $j = 1, ..., d$.*[3]

[3] e.g., this can be achieved by uniformly randomizing the splitting variable with probability π or via a Poisson sampling scheme where a random subset of the variables of size m is chosen to consider for candidate splits, with $m \sim \text{Poisson}(\lambda)$.

The key modification to GRF's tree learner is our incorporation of orthogonal nuisance estimation in the splitting criterion. While the splitting criterion does not factor into our theoretical analysis (similar to (Tibshirani et al., 2018)), we find it to be an effective practical heuristic.

Splitting criterion with orthogonalization. At each internal node P we perform a two-stage estimation over $(P \cap S^1)$, i.e. the set of examples in S^1 that reach node P: 1) compute a nuisance estimate $\hat{h}_P$ using only data $P \cap S^1$ (e.g. by estimating a parameter $\hat{\nu}_P$ that minimizes $\sum_{i \in (P \cap S^1)} \ell(Z_i; \nu) + \lambda \|\nu\|_1$ and setting $\hat{h}_P(\cdot) = g(\cdot; \hat{\nu}_P)$), and then 2) form estimate $\hat{\theta}_P$ using $\hat{h}_P$:[4]

$$\hat{\theta}_P \in \text{argmin}_{\theta \in \Theta} \left\| \sum_{i \in (P \cap S^1)} \psi(Z_i; \theta, \hat{h}_P(W_i)) \right\|$$

We now generate a large random set of candidate axis-aligned splits (satisfying Specification 1 and we want to find the split into two children C_1 and C_2 such that if we perform the same two-stage estimation separately at each child, the new estimates $\hat{\theta}_{C_1}$ and $\hat{\theta}_{C_2}$ take on very different values, so that the heterogeneity of the two children nodes is maximized. Performing the two-stage estimation of $\hat{\theta}_{C_1}$ and $\hat{\theta}_{C_2}$ for all candidate splits is too computationally expensive. Instead, we will approximate these estimates by taking a Newton step from the parent node estimate $\hat{\theta}_P$: for any child node C given by a candidate split, our proxy estimate is:

$$\tilde{\theta}_C = \hat{\theta}_P - \frac{1}{|C \cap S^1|} \sum_{i \in C_j \cap S^1} A_P^{-1} \psi(Z_i; \hat{\theta}_P, \hat{h}_P(X_i, W_i))$$

where $A_P = \frac{1}{|P \cap S^1|} \sum_{i \in P \cap S_b^1} \nabla_\theta \psi(Z_i; \hat{\theta}_P, \hat{h}_P(X_i, W_i))$. We select the candidate split that maximizes the following proxy heterogeneity score: for each coordinate $t \in [p]$ let

$$\tilde{\Delta}_t(C_1, C_2) = \sum_{j=1}^2 \frac{1}{|C_j \cap S^1|} \left(\sum_{i \in C_j \cap S^1} \rho_{t,i} \right)^2 \quad (7)$$

where $\rho_{t,i} = A_P^{-1} \psi_t(Z_i; \hat{\theta}_P, \hat{h}_P(X_i, W_i))$. We then create a single heterogeneity score per split as a convex combination that puts weight η on the mean and $(1 - \eta)$ on the maximum score across coordinates. η is chosen uniformly at random in $[0, 1]$ at each iteration of splitting. Hence, some splits focus on heterogeneity on average, while others focus on creating heterogeneity on individual coordinates.

ORF weights and estimator. For each tree indexed $b \in [B]$ based on subsample S_b, let $L_b(x) \subseteq \mathcal{X}$ be the leaf that contains the target feature x. We assign *tree weight* and *ORF weight* to each observation i:

$$a_{ib} = \frac{\mathbf{1}[(X_i \in L_b(x)) \wedge (Z_i \in S_b^2)]}{|L_b(x) \cap S_b^2|}, \quad a_i = \frac{1}{B} \sum_{b=1}^B a_{ib}$$

[4] In our implementation we actually use a cross-fitting approach, where we use half of $P \cap S^1$ to compute a nuisance function to apply to the other half and vice versa.

Wager & Athey (2015) show that under the structural specification of the trees, the tree weights are non-zero only around a small neighborhood of x; a property that we will leverage in our analysis.

Theorem 3.1 (Kernel shrinkage (Wager & Athey, 2015))**.** *Suppose the minimum leaf size parameter $r = O(1)$, the tree is ρ-balanced and π-random-split and the distribution of X admits a density in $[0,1]^d$ that is bounded away from zero and infinity. Then the tree weights satisfy $\mathbb{E}\left[\sup\{\|x - x_i\| : a_{ib} > 0\}\right] = O(s^{-\frac{1}{2\alpha d}})$, with $\alpha = \frac{\log(\rho^{-1})}{\pi \log((1-\rho)^{-1})}$ and s the size of the subsamples.*

4. Convergence and Asymptotic Analysis

The *ORF estimate* $\hat{\theta}$ is computed by solving the weighted moment condition in Equation (5), using the *ORF weights* as described in the previous section. We now provide theoretical guarantees for $\hat{\theta}$ under the following assumption on the moment, score fuction and the data generating process.

Assumption 4.1. *The moment condition and the score function satisfy the following:*

1. ***Local Orthogonality.*** *The moment condition satisfies local orthogonality.*
2. ***Identifiability.*** *The moments $m(x;\theta,h_0) = 0$ has a unique solution $\theta_0(x)$.*
3. ***Smooth Signal.*** *The moments $m(x;\theta,h)$ are $O(1)$-Lipschitz in x for any $\theta \in \Theta, h \in H$.*
4. ***Curvature.*** *The Jacobian $\nabla_\theta m(x;\theta_0(x),h_0)$ has minimum eigenvalue bounded away from zero.*
5. ***Smoothness of scores.*** *For every $j \in [p]$ and for all θ and h, the eigenvalues of the expected Hessian $\mathbb{E}\left[\nabla^2_{(\theta,h)}\psi_j(Z;\theta,h(W)) \mid x, W\right]$ are bounded above by a constant $O(1)$. For any Z, the score $\psi(Z;\theta,\xi)$ is $O(1)$-Lipschitz in θ for any ξ and $O(1)$-Lipschitz in ξ for any θ. The gradient of the score with respect to θ is $O(1)$-Lipschitz in ξ.*
6. ***Boundedness.*** *The parameter set Θ has constant diameter. There exists a bound $\psi_{\max}$ such that for any observation Z, the first-stage nuisance estimate $\hat{h}$ satisfies $\|\psi(Z;\theta,\hat{h})\|_\infty \leq \psi_{\max}$ for any $\theta \in \Theta$.*
7. ***Full Support*** *X. The distribution of X admits a density that is bounded away from zero and infinity.*

All the results presented in the remainder of the paper will assume these conditions and we omit stating so in each of the theorems. Any extra conditions required for each theorem will be explicitly provided. Note that except for the local orthogonality condition, all of the assumptions are imposing standard boundedness and regularity conditions of the moments.

Theorem 4.2 (L^q-Error Bound)**.** *Suppose that: $\mathbb{E}\left[\mathcal{E}(\hat{h})^{2q}\right]^{1/2q} \leq \chi_{n,2q}$. Then:*

$$\mathbb{E}\left[\|\hat{\theta} - \theta_0\|^q\right]^{1/q} = O\left(\frac{1}{s^{\frac{1}{2\alpha d}}} + \sqrt{\frac{s\log(\frac{n}{s})}{n}} + \chi^2_{n,2q}\right)$$

Theorem 4.3 (High Probability Error Bound)**.** *Suppose that the score is the gradient of a convex loss and let $\sigma > 0$ denote the minimum eigenvalue of the jacobian M. Moreover, suppose that the nuisance estimate satisfies that w.p. $1-\delta$: $\mathcal{E}(\hat{h}) \leq \chi_{n,\delta}$. Then w.p. $1-2\delta$:*

$$\|\hat{\theta} - \theta_0\| = \frac{O\left(s^{-\frac{1}{2\alpha d}} + \sqrt{\frac{s\log(\frac{n}{s\delta})}{n}} + \chi^2_{n,\delta}\right)}{\sigma - O(\chi_{n,\delta})} \tag{8}$$

For asymptotic normality we will restrict our framework to the case of parametric nuisance functions, i.e. $h(X,W) = g(W;\nu(X))$ for some known function g and to a particular type of nuisance estimators that recover the true parameter $\nu_0(x)$. Albeit we note that the parameter $\nu(X)$ can be an arbitrary non-parametric function of X and can also be high-dimensional. We will further assume that the moments also have a smooth co-variance structure in X, i.e. if we let

$$V = \psi(Z;\theta_0(x), g(W;\nu_0(x)))$$

then $\mathrm{Var}(V \mid X = x')$ is Lipschitz in x' for any $x' \in [0,1]^d$.

Theorem 4.4 (Asymptotic Normality)**.** *Suppose that $h_0(X,W)$ takes a locally parametric form $g(W;\nu_0(X))$, for some known function $g(\cdot;\nu)$ that is $O(1)$-Lipschitz in ν w.r.t. the ℓ_r norm for some $r \geq 1$ and the nuisance estimate is of the form $\hat{h}(X,W) = g(W;\hat{\nu}(x))$ and satisfies:*

$$\mathbb{E}\left[\|\hat{\nu}(x) - \nu_0(x)\|_r^4\right]^{1/4} \leq \chi_{n,4} = o\left((s/n)^{1/4}\right)$$

Suppose that s is chosen such that: $s^{-1/(2\alpha d)} = o((s/n)^{1/2-\varepsilon})$, for any $\varepsilon > 0$, and $s = o(n)$. Moreover, $\mathrm{Var}(V \mid X = x')$ is Lipschitz in x' for any $x' \in [0,1]^d$. Then for any coefficient $\beta \in \mathbb{R}^p$, with $\|\beta\| \leq 1$, assuming $\mathrm{Var}(\beta^\intercal M^{-1} V | X = x') > 0$ for any $x' \in [0,1]^d$, there exists a sequence $\sigma_n = \Theta(\sqrt{\mathrm{polylog}(n/s)s/n})$, such that:

$$\sigma_n^{-1}\left\langle \beta, \hat{\theta} - \theta_0 \right\rangle \to_d \mathcal{N}(0,1) \tag{9}$$

Given the result in Theorem 4.4, we can follow the same approach of *Bootstrap of Little Bags* by (Athey et al., 2017; Sexton & Laake, 2009) to build valid confidence intervals.

5. Nuisance Estimation: Forest Lasso

Next, we study the nuisance estimation problem in the first stage and provide a general nuisance estimation method that leverages locally sparse parameterization of the nuisance function, permitting low error rates even for high-dimensional problems. Consider the case when the nuisance

function h takes the form $h(x, w) = g(w; \nu(x))$ for some known functional form g, for some known function g but unknown function $\nu : \mathcal{X} \to \mathbb{R}^{d_\nu}$, with d_ν potentially growing with n. Moreover, the parameter $\nu_0(x)$ of the true nuisance function h_0 is identified as the minimizer of some local loss, as defined in Equation (6).

We consider the following estimation process: given a set of observations D_1, we run the same tree learner in Section 3 over B random subsamples (without replacement) to compute ORF weights a_i for each observation i over D_1. Then we apply a local ℓ_1 penalized M-estimation:

$$\hat{\nu}(x) = \arg\min_{\nu \in \mathcal{V}} \sum_{i=1}^{n} a_i \, \ell(Z_i; \nu) + \lambda \|\nu\|_1 \qquad (10)$$

To provide formal guarantees for this method we will need to make the following assumptions.

Assumption 5.1 (Assumptions for nuisance estimation)**.** *The target parameter and data distribution satisfy:*

- *For any $x \in \mathcal{X}$, $\nu(x)$ is k-sparse with support $S(x)$.*
- *$\nu(x)$ is a $O(1)$-Lipschitz in x and the function $\nabla_\nu L(x; \nu) = \mathbb{E}\left[\nabla_\nu \ell(Z; \nu) \mid X = x\right]$ is $O(1)$-Lipschitz in x for any ν, with respect to the ℓ_2 norm.*
- *The data distribution satisfies the conditional restricted eigenvalue condition: for all $\nu \in \mathcal{V}$ and for all $z \in \mathcal{Z}$, for some matrix $\mathcal{H}(z)$ that depends only on the data: $\nabla_{\nu\nu} \ell(z; \nu) \succeq \mathcal{H}(z) \succeq 0$, and for all x and for all $\nu \in C(S(x); 3) \equiv \{\nu \in \mathbb{R}^d : \|\nu_{S(x)^c}\|_1 \leq 3\|\nu_{S(x)}\|_1\}$:*

$$\nu^T \, \mathbb{E}\left[\mathcal{H}(Z) \mid X = x\right] \nu \geq \gamma \|\nu\|_2^2 \qquad (11)$$

Under Assumption 5.1 we show that the local penalized estimator achieves the following parameter recovery guarantee.

Theorem 5.2. *With probability $1 - \delta$:*

$$\|\hat{\nu}(x) - \nu_0(x)\|_1 \leq \frac{2\lambda k}{\gamma - 32k\sqrt{s \ln(d_\nu/\delta)/n}}$$

as long as $\lambda \geq \Theta\left(s^{-1/(2\alpha d)} + \sqrt{\frac{s \ln(d_\nu/\delta)}{n}}\right)$.

Example 5.3 (Forest Lasso)**.** *For locally sparse linear regression, $Z_i = (x_i, y_i, W_i)$ and $\ell(Z_i; \nu) = (y_i - \langle \nu, W_i \rangle)^2$. This means, $\nabla_{\nu\nu} \ell(Z_i; \nu) = W_i W_i^T = \mathcal{H}(Z_i)$. Hence, the conditional restricted eigenvalue condition is simply a conditional covariance condition: $\mathbb{E}\left[WW^\intercal \mid x\right] \succeq \gamma I$.*

Example 5.4 (Forest Logistic Lasso)**.** *For locally sparse logistic regression, $Z_i = (x_i, y_i, W_i)$, $y_i \in \{0, 1\}$ and $\ell(Z_i; \nu) = y_i \ln\left(\mathcal{L}(\langle \nu, W_i \rangle)\right) + (1 - y_i) \ln\left(1 - \mathcal{L}(\langle \nu, W_i \rangle)\right)$, where $\mathcal{L}(t) = 1/(1+e^{-t})$ is the logistic function. In this case, $\nabla_{\nu\nu} \ell(Z_i; \nu) = \mathcal{L}(\langle \nu, W_i \rangle)(1 - \mathcal{L}(\langle \nu, W_i \rangle)) W_i W_i^\intercal \succeq \rho W_i W_i^\intercal = \mathcal{H}(Z_i)$ (assuming the index $\langle \nu, w \rangle$ is bounded in some finite range). Hence, our conditional restricted eigenvalue condition is the same conditional covariance condition: $\rho \, \mathbb{E}\left[WW^T \mid x\right] \succeq \rho\gamma I$.*

6. Heterogeneous Treatment Effects

Now we apply ORF to the problem of estimating *heterogeneous treatment effects*. We will consider the following *extension* of the *partially linear regression (PLR)* model due to Robinson (1988). [5] We have $2n$ i.i.d. observations $D = \{Z_i = (T_i, Y_i, W_i, X_i)\}_{i=1}^{2n}$ such that for each i, T_i represents the treatment applied that can be either real-valued (in $\mathbb{R}^p$) or discrete (taking values in $\{0, e_1, \ldots, e_p\}$, where each e_j denotes the standard basis in $\mathbb{R}^p$), $Y_i \in \mathbb{R}$ represents the outcome, $W_i \in [-1, 1]^{d_\nu}$ represents potential confounding variables (controls), and $X_i \in \mathcal{X} = [0, 1]^d$ is the feature vector that captures the heterogeneity. The set of parameters are related via the following equations:

$$Y = \langle \mu_0(X, W), T \rangle + f_0(X, W) + \varepsilon, \qquad (12)$$

$$T = g_0(X, W) + \eta, \qquad (13)$$

where η, ε are bounded unobserved noises such that $\mathbb{E}\left[\varepsilon \mid W, X, T\right] = 0$ and $\mathbb{E}\left[\eta \mid X, W, \varepsilon\right] = 0$. In the main equation (12), $\mu_0 \colon \mathbb{R}^d \times \mathbb{R}^{d_\nu} \to [-1, 1]^p$ represents the treatment effect function. Our goal is to estimate *conditional average treatment effect* (CATE) $\theta_0(x)$ conditioned on target feature x:

$$\theta_0(x) = \mathbb{E}\left[\mu_0(X, W) \mid X = x\right]. \qquad (14)$$

The confounding equation (13) determines the relationship between treatments variable T and the feature X and confounder W. To create an orthogonal moment for identifying $\theta_0(x)$, we follow the classical *residualization* approach similar to (Chernozhukov et al., 2017). First, observe that

$$Y - \mathbb{E}\left[Y \mid X, W\right] = \left\langle \mu_0(X, W), T - \mathbb{E}\left[T \mid X, W\right] \right\rangle + \varepsilon$$

Let us define the function $q_0(X, W) = \mathbb{E}\left[Y \mid X, W\right]$, and consider the residuals $\tilde{Y} = Y - q_0(X, W)$ and $\tilde{T} = T - g_0(X, W) = \eta$. Then we can simplify the equation as $\tilde{Y} = \mu_0(X, W) \cdot \tilde{T} + \varepsilon$. As long as η is independent of $\mu_0(X, W)$ conditioned on X (e.g. η is independent of W or $\mu_0(X, W)$ does not depend on W), we also have $\mathbb{E}\left[\mu_0(X, W) \mid X, \eta\right] = \mathbb{E}\left[\mu_0(X, W) \mid X\right] = \theta(X)$. Since $\mathbb{E}\left[\varepsilon \mid X, \eta\right] = \mathbb{E}\left[\mathbb{E}\left[\varepsilon \mid X, W, T\right] \mid X, \eta\right] = 0$, we have

$$\mathbb{E}\left[\tilde{Y} \mid X, \tilde{T}\right] = \mathbb{E}\left[\mu_0(X, W) \mid X\right] \cdot \tilde{T} = \theta(X) \cdot \tilde{T}.$$

This relationship suggests that we can obtain an estimate of $\theta(x)$ by regressing $\tilde{Y}$ on $\tilde{T}$ locally around $X = x$. We can thus define the *orthogonalized* score function: for any observation $Z = (T, Y, W, x)$, any parameter $\theta \in \mathbb{R}^p$, any estimates q and g for functions q_0 and g_0, the score $\psi(Z; \theta, h(X, W))$ is:

$$\{Y - q(X, W) - \theta\left(T - g(X, W)\right)\}\left(T - g(X, W)\right),$$

[5] The standard PLR model (Robinson, 1988) considers solely the case of constant treatment effects, $Y = \langle \theta_0, T \rangle + f_0(X, W) + \varepsilon$, and the goal is the estimation of the parameter θ_0.

where $h(X,W) = (q(X,W), g(X,W))$. In the appendix, we show that this moment condition satisfies local orthogonality, and it identifies $\theta_0(x)$ as long as as the noise η is independent of $\mu_0(X,W)$ conditioned on X and the expected matrix $\mathbb{E}\left[\eta\eta^\intercal \mid X = x\right]$ is invertible. Even though the approach applies generically, to obtain formal guarantees on the nuisance estimates via our Forest Lasso, we will restrict their functional form.

Real-valued treatments. Suppose f_0 and each coordinate j of g_0 and μ_0 are given by high-dimensional linear functions: $f_0(X,W) = \langle W, \beta_0(X)\rangle$, $\mu_0^j(X,W) = \langle W, u_0^j(X)\rangle$, $g_0^j(X,W) = \langle W, \gamma_0^j(X)\rangle$, where $\beta_0(X), \gamma_0^j(X), u_0^j(X)$ are k-sparse vectors in $\mathbb{R}^{d_\nu}$. Consequently, $q_0(X,W)$ can be written as a k^2-sparse linear function over degree-2 polynomial features $\phi_2(W)$ of W. Then as long as γ_0, β_0 and μ_0 are Lipschitz in X and the confounders W satisfy $\mathbb{E}\left[\phi_2(W)\phi_2(W)^\intercal \mid X\right] \succeq \Omega(1)I$, then we can use Forest Lasso to estimate both $g_0(x,w)$ and $q_0(x,w)$. Hence, we can apply the ORF algorithm to get estimation error rates and asymptotic normality results for $\hat{\theta}$. (see Appendix B for formal statement).

Discrete treatments. We now describe how our theory can be applied to discrete treatments. Suppose f_0 and each coordinate j of g_0 are of the form: $f_0(X,W) = \langle W, \beta_0(X)\rangle$ and $g_0^j(X,W) = \mathcal{L}(\langle W, \gamma_0^j(X)\rangle)$, where $\mathcal{L}(t) = 1/(1+e^{-t})$ is the logistic function. Note in this case η is not independent of W since $\mathrm{Var}(\eta_j) = g_0^j(X,W)(1 - g_0^j(X,W))$. To maintain the conditional independence between $\mu_0(X,W)$ and η conditioned on X, we focus on the setting where μ_0 is only a function of X, i.e. $\mu(X,W) = \theta(X)$ for all W, X. In this setting we can estimate g_0 by running a forest logistic lasso for each treatment j. Then we can estimate $q_0(x,W)$ as follows: For each $t \in \{e_1, \ldots, e_p\}$ estimate the expected counter-factual outcome function: $m_0^t(x,W) = \mu_0^t(x,W) + f_0(x,W)$, by running a forest lasso between Y and X, W only among the subset of samples that received treatment t. Similarly, estimate $f_0(x,W)$ by running a forest lasso between Y and X, W only among the subset of samples that received treatment $t = 0$. Then observe that $q_0(x,W)$ can be written as a function of f_0, g_0^t and m_0^t. Thus we can combine these estimates to get an estimate of q_0. Hence, we can obtain a guarantee similar to that of Corollary B.1 (see appendix).

Doubly robust moment for discrete treatments. In the setting where μ also depends on W and treatments are discrete, we can formulate an alternative orthogonal moment that identifies the CATE even when η is correlated with $\mu(X,W)$. This moment is based on first constructing unbiased estimates of the counterfactual outcome $m_0^t(X,W) = \mu_0^t(X,W) + f_0(X,W)$ for every observation X, W and for any potential treatment t, i.e. even for $t \neq T$. The latter is done by invoking the doubly robust formula (Robins & Rotnitzky, 1995; Cassel et al., 1976; Kang et al., 2007):

$$Y^{(t)} = m_0^t(X,W) + \frac{(Y - m_0^t(X,W))\mathbf{1}\{T=t\}}{g_0^t(X,W)}$$

with the convention that $g_0^0(X,W) = 1 - \sum_{t\neq 0} g_0^t(X,W)$ and $m_0^0(X,W) = f_0(X,W)$. Then we can identify the parameter $\theta^t(x)$ using the moment: $\mathbb{E}[Y^{(t)} - Y^{(0)} | X = x] = \theta_t(x)$. One can easily show that this moment satisfies the Neyman orthogonality condition with respect to the nuisance functions m and g (see appendix). In fact this property is essentially implied by the fact that the estimates $Y^{(t)}$ satisfy the double robustness property, since double robustness is a stronger condition than orthogonality. We will again consider $\mu_0^j(X,W) = \langle W, u_0^j(X)\rangle$. Then using similar reasoning as in the previous paragraph, we see that with a combination of forest logistic lasso for g_0^t and forest lasso for m_0^t, we can estimate these nuisance functions at a sufficiently fast rate for our ORF estimator (based on this doubly robust moment) to be asymptotically normal, assuming they have locally sparse linear or logistic parameterizations.

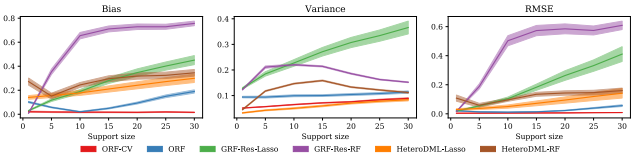

Figure 2: Bias, variance and RMSE as a function of support size for $n = 5000$, $p = 500$, $d = 1$ and a piecewise linear treatment response function. The solid lines represent the mean of the metrics over Monte Carlo experiments and test points, and the filled regions depict the standard deviation, scaled down by 3 for clarity.

7. Monte Carlo Experiments

We compare the empirical performance of ORF with other methods in the literature (and their variants).[6] The data generating process we consider is described by the following equations: $Y_i = \theta_0(x_i)\, T_i + \langle W_i, \gamma_0\rangle + \varepsilon_i$, $T_i = \langle W_i, \beta_0\rangle + \eta_i$. Moreover, x_i is drawn from the uniform distribution $U[0,1]$, W_i is drawn from $\mathcal{N}(0, I_p)$, and the noise terms $\varepsilon_i \sim U[-1,1], \eta_i \sim U[-1,1]$. The k-sparse vectors $\beta_0, \gamma_0 \in \mathbb{R}^p$ have coefficients drawn independently from $U[0,1]$. The dimension $p = 500$ and we vary the support size k over the range of $\{1, 5, 10, 15, 20, 25, 30\}$. We examine a treatment function $\theta(x)$ that is continuous and piecewise linear (detailed in Figure 3). In Appendix H we analyze other forms for $\theta(x)$.

For each fixed treatment function, we repeat 100 experiments, each of which consists of generating 5000 observa-

[6] The source code for running these experiments is available in the git repo Microsoft/EconML.

tions from the DGP, drawing the vectors β_0 and γ_0, and estimating $\hat{\theta}(x)$ at 100 test points x over a grid in $[0, 1]$. We then calculate the bias, variance and root mean squared error (RMSE) of each estimate $\hat{\theta}(x)$. Here we report summary statistics of the median and $5 - 95$ percentiles of these three quantities across test points, so as to evaluate the average performance of each method. We compare two variants of ORF with two variants of GRF (Athey et al., 2017) (see Appendix H for a third variant) and two extensions of double ML methods for heterogeneous treatment effect estimation (Chernozhukov et al., 2017).

ORF variants. (1) ORF: We implement ORF as described in Section 3, setting parameters under the guidance of our theoretical result: subsample size $s \approx (n/\log(p))^{1/(2\tau+1)}$, Lasso regularization $\lambda_\gamma, \lambda_q \approx \sqrt{\log(p)s/n}/20$ (for both tree learner and kernel estimation), number of trees $B = 100 \geq n/s$, a max tree depth of 20, and a minimum leaf size of $r = 5$. (2) ORF with LassoCV (ORF-CV): we replaced the Lasso algorithm in ORF's kernel estimation, with a cross-validated Lasso for the selection of the regularization parameter λ_γ and λ_q. ORF-CV provides a more systematic optimization over the parameters.

GRF variants. (1) GRF-Res-Lasso: We perform a naive combination of double ML and GRF by first residualizing the treatments and outcomes on both the features x and controls W, then running GRF R package by (Tibshirani et al., 2018) on the residualized treatments $\hat{T}$, residualized outcomes $\hat{Y}$, and features x. A cross-validated Lasso is used for residualization. (2) GRF-Res-RF: We combine DoubleML and GRF as above, but we use cross-validated Random Forests for calculating residuals $\hat{T}$ and $\hat{Y}$.

Double ML with Polynomial Heterogeneity (DML-Poly). An extension of the classic Double ML procedure for heterogeneous treatment effects introduced in (Chernozhukov et al., 2017). This method accounts for heterogeneity by creating an expanded linear base of composite treatments (cross products between treatments and features). (1) Heterogeneous Double ML using LassoCV for first-stage estimation (HeteroDML-Lasso): In this version, we use Lasso with cross-validation for calculating residuals on $x \cup W$ in the first stage. (2) Heterogeneous Double ML using random forest for first-stage estimation (HeteroDML-RF): A more flexible version that uses random forests to perform residualization on treatments and outcomes. The latter performs better when treatments and outcomes have a non-linear relationship with the joint features of (x, W).

We generated data according to the Monte Carlo process above and set the parameters to $n = 5000$ samples, $p = 500$ controls, $d = 1$ features and support size $k \in \{1, 5, 10, 15, 20, 25, 30\}$ and three types of treatment effect functions. In this section, we present the results for a piecewise linear treatment effect function.

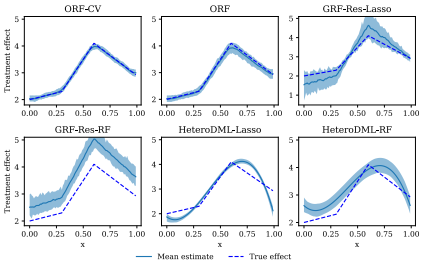

Figure 3: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $k = 15$, and $\theta(x) = (x+2)\mathbb{I}_{x\leq 0.3} + (6x+0.5)\mathbb{I}_{x>0.3 \text{ and } x\leq 0.6} + (-3x+5.9)\mathbb{I}_{x>0.6}$. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

In Figure 3, we inspect the goodness of fit for the chosen estimation methods across 100 Monte Carlo experiments. We note the limitations of two versions of the GRF-Res estimators, GRF-Res-Lasso and GRF-Res-RF, in capturing the treatment effect function well. The GRF-Res-RF estimations have a consistent bias as the Random Forest residualization cannot capture the dependency on the controls W given their high-dimensionality. The HeteroDML methods are not flexible enough to capture the complexity of the treatment effect function. The best performers are the ORF-CV, ORF, and GRF-Res-Lasso, with the latter estimator having a larger bias and variance.

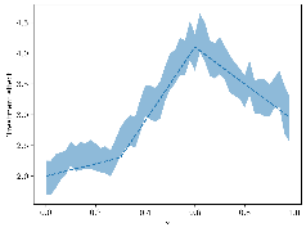

Figure 4: Sample 1%-99% confidence intervals for 1000 bootstrap iterations with parameters $n = 5000$, $p = 500$, $d = 1$, $k = 15$, and $\theta(x) = (x+2)\mathbb{I}_{x\leq 0.3} + (6x+0.5)\mathbb{I}_{x>0.3 \text{ and } x\leq 0.6} + (-3x+5.9)\mathbb{I}_{x>0.6}$. Approximately 90% of the sampled test points are contained in the interval.

We analyze these estimators as we increase the support size of W. Figures 2 illustrate the evaluation metrics across different support sizes. The ORF-CV performs very well, with consistent bias and RMSE across support sizes and treatment functions. The bias, variance and RMSE of the ORF grow with support size, but this growth is at a lower rate compared to the alternative estimators. The ORF-CV and ORF algorithms perform better than the GRF-Res methods on all metrics for this example. We observe this pattern for the other choices of support size, sample size and treatment effect function (see Appendix H). In figure 4, we provide a snapshot of the bootstrap confidence interval coverage for this example.

References

Ai, C. and Chen, X. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71 (6):1795–1843, 2003. doi: 10.1111/1468-0262.00470. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00470.

Athey, S., Tibshirani, J., and Wager, S. Generalized Random Forests. *ArXiv e-prints*, October 2017.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976.

Chen, X. and Pouzo, D. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, 152(1):46 – 60, 2009. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2009.02.002. URL http://www.sciencedirect.com/science/article/pii/S0304407609000529. Recent Adavances in Nonparametric and Semiparametric Econometrics: A Volume Honouring Peter M. Robinson.

Chernozhukov, V., Newey, W. K., and Santos, A. Constrained Conditional Moment Restriction Models. *arXiv e-prints*, art. arXiv:1509.06311, September 2015.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. Locally Robust Semiparametric Estimation. *arXiv e-prints*, art. arXiv:1608.00033, July 2016.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, May 2017. doi: 10.1257/aer.p20171038. URL http://www.aeaweb.org/articles?id=10.1257/aer.p20171038.

Chernozhukov, V., Goldman, M., Semenova, V., and Taddy, M. Orthogonal Machine Learning for Demand Estimation: High Dimensional Causal Inference in Dynamic Panels. *ArXiv e-prints*, December 2017.

Chernozhukov, V., Nekipelov, D., Semenova, V., and Syrgkanis, V. Plug-in Regularized Estimation of High-Dimensional Parameters in Nonlinear Semiparametric Models. *arXiv e-prints*, art. arXiv:1806.04823, June 2018.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Counterfactual Prediction with Deep Instrumental Variables Networks. *ArXiv e-prints*, December 2016.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015. ISBN 1498712169, 9781498712163.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459. URL http://www.jstor.org/stable/2282952.

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Kang, J. D., Schafer, J. L., et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Mentch, L. and Hooker, G. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1): 841–881, 2016.

Newey, W. K. 16 efficient estimation of models with conditional moment restrictions. In *Econometrics*, volume 11 of *Handbook of Statistics*, pp. 419 – 454. Elsevier, 1993. doi: https://doi.org/10.1016/S0169-7161(05)80051-3. URL http://www.sciencedirect.com/science/article/pii/S0169716105800513.

Neyman, J. C(α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 41(1/2):1–21, 1979. ISSN 0581572X. URL http://www.jstor.org/stable/25050174.

Nie, X. and Wager, S. Learning Objectives for Treatment Effect Estimation. *ArXiv e-prints*, December 2017.

Peel, T., Anthoine, S., and Ralaivola, L. Empirical bernstein inequalities for u-statistics. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pp. 1903–1911. Curran Associates, Inc., 2010.

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. *ArXiv e-prints*, July 2017.

Reiss, P. C. and Wolak, F. A. Chapter 64 structural econometric modeling: Rationales and examples from industrial organization. volume 6 of *Handbook of Econometrics*, pp. 4277 – 4415. Elsevier, 2007. doi: https://doi.org/10.1016/S1573-4412(07)06064-3. URL http://www.sciencedirect.com/science/article/pii/S1573441207060643.

Robins, J. M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912705.

Sexton, J. and Laake, P. Standard errors for bagged and random forest estimators. *Computational Statistics and Data Analysis*, 53(3):801 – 811, 2009. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2008.08.007. URL http://www.sciencedirect.com/science/article/pii/S0167947308003988. Computational Statistics within Clinical Research.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. *CoRR*, abs/1502.02362, 2015. URL http://arxiv.org/abs/1502.02362.

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudík, M., Langford, J., Jose, D., and Zitouni, I. Off-policy evaluation for slate recommendation. *CoRR*, abs/1605.04812, 2016. URL http://arxiv.org/abs/1605.04812.

Tibshirani, J., Athey, S., Wager, S., Friedberg, R., Miner, L., and Wright, M. *grf: Generalized Random Forests (Beta)*, 2018. URL https://CRAN.R-project.org/package=grf. R package version 0.10.2.

Wager, S. and Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *ArXiv e-prints*, October 2015.

Wang, Y.-X., Agarwal, A., and Dudík, M. Optimal and adaptive off-policy evaluation in contextual bandits. 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. 3589–3597, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/wang17a.html.