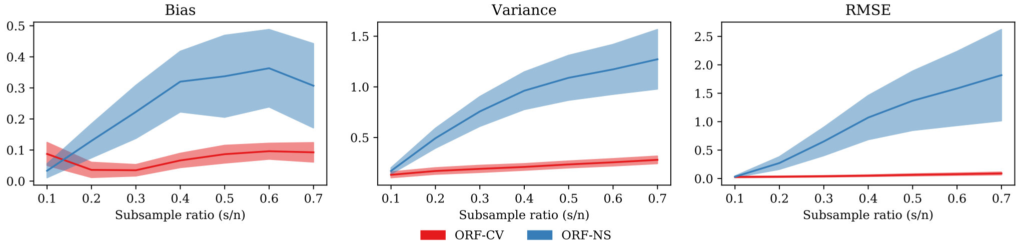

Figure 5: Bias, variance and RMSE versus subsample ratio used for training individual trees. The solid lines represent the means and the filled regions depict the standard deviation for the different metrics across test points, averaged over 100 Monte Carlo experiments.

A. Two-forest ORF v.s. One-forest ORF

A natural question about ORF is whether it is necessary to have two separate forests for the two-stage estimation. We investigate this question by implementing a variant of ORF without sample splitting (ORF-NS)—it builds only one random forest over the entire dataset, and perform the two-stage estimation using the same set of importance weights. We empirically compare ORF-CV with ORF-NS. In Figure 5, we note that the bias, variance and RMSE of the ORF-NS increase drastically with the subsample ratio (s/n), whereas the same metrics are almost constant for the ORF-CV. This phenomenon is consistent with the theory, since larger subsamples induce a higher probability of collision between independently drawn samples, and the "spill-over" can incur large bias and error.

B. Formal Guarantee for Real-Valued Treatments

Corollary B.1 (Accuracy for real-valued treatments)**.** *Suppose that $\beta_0(X)$ and each coorindate $u_0^j(X), \gamma_0^j(X)$ are Lipschitz in X and have ℓ_1 norms bounded by $O(1)$ for any X. Assume that distribution of X admits a density that is bounded away from zero and infinity. For any feature X, the conditional covariance matrices satisfy $\mathbb{E}\left[\eta\eta^\intercal \mid X\right] \succeq \Omega(1)I_p$, $\mathbb{E}\left[WW^\intercal \mid X\right] \succeq \Omega(1)I_{d_\nu}$ and $\mathbb{E}\left[\varphi_2(W)\varphi_2(W)^\intercal \mid X\right] \succeq \Omega(1)I_{d_\nu^2+d_\nu}$, where $\varphi_2(W)$ denotes the degree-2 polynomial feature vector of W. Then with probability $1-\delta$, ORF returns an estimator $\hat\theta$ such that*

$$\|\hat\theta-\theta_0\| \le O\left(n^{\frac{-1}{2+2\alpha d}}\sqrt{\log(nd_\nu/\delta)}\right)$$

as long as the sparsity $k \le O\left(n^{\frac{1}{8+8\alpha d}}\right)$ and the subsampling rate of ORF $s = \Theta(n^{\alpha d/(1+\alpha d)})$. Moreover, for any $b \in \mathbb{R}^p$ with $\|b\| \le 1$, there exists a sequence $\sigma_n = \Theta(\sqrt{\text{polylog}(n)n^{-1/(1+\alpha d)}})$ such that

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

as long as the sparsity $k = o\left(n^{\frac{1}{8+8\alpha d}}\right)$ and the subsampling rate of ORF $s = \Theta(n^{\varepsilon+\alpha d/(1+\alpha d)})$ for any $\varepsilon > 0$.

C. Uniform Convergence of Lipschitz U-Processes

Lemma C.1 (Stochastic Equicontinuity for U-statistics via Bracketing)**.** *Consider a parameter space Θ that is a bounded subset of $\mathbb{R}^p$, with $\text{diam}(\Theta) = \sup_{\theta,\theta'\in\Theta}\|\theta-\theta'\|_2 \le R$. Consider the U-statistic over n samples of order s:*

$$\mathbb{G}_{s,n}f(\cdot;\theta) = \binom{n}{s}^{-1}\sum_{1\le i_1\le\ldots\le i_s\le n} f(z_{i_1},\ldots,z_{i_s};\theta) \tag{15}$$

where $f(\cdot;\theta) : \mathcal{Z}^s \to \mathbb{R}$ is a known symmetric function in its first s sarguments and L-Lipschitz in θ. Suppose that $\sup_{\theta\in\Theta}\sqrt{\mathbb{E}\left[f(Z_1,\ldots,Z_s;\theta)^2\right]} \le \eta$ and $\sup_{\theta\in\Theta,Z_1,\ldots,Z_s\in\mathcal{Z}^s} f(Z_1,\ldots,Z_s;\theta) \le G$. Then w.p. $1-\delta$:

$$\sup_{\theta\in\Theta}\left|\mathbb{G}_{s,n}f(\cdot;\theta) - \mathbb{E}[f(Z_{1:s};\theta)]\right| = O\left(\eta\sqrt{\frac{s\left(\log(n/s)+\log(1/\delta)\right)}{n}} + (G+L\,R)\frac{s(\log(n/s)+\log(1/\delta))}{n}\right) \tag{16}$$

Proof of Lemma C.1. Note that for any fixed $\theta \in \Theta$, $\Psi_s(\theta, Z_{1:n})$ is a U-statistic of order s. Therefore by the Bernestein inequality for U-statistics (see e.g. Theorem 2 of (Peel et al., 2010)), for any fixed $\theta \in \Theta$, w.p. $1-\delta$

$$|\mathbb{G}_{s,n} f(\cdot;\theta) - \mathbb{E}[f(Z_{1:s};\theta)]| \leq \eta\sqrt{\frac{2\log(1/\delta)}{n/s}} + G\,\frac{2\log(1/\delta)}{3(n/s)}$$

Since $\text{diam}(\Theta) \leq R$, we can find a finite space Θ_ε of size R/ε, such that for any $\theta \in \Theta$, there exists $\theta_\varepsilon \in \Theta_\varepsilon$ with $\|\theta - \theta_\varepsilon\| \leq \varepsilon$. Moreover, since f is L-Lipschitz with respect to θ:

$$|\mathbb{G}_{s,n} f(\cdot;\theta) - \mathbb{E}[f(Z_{1:s};\theta)]| \leq |\mathbb{G}_{s,n} f(\cdot;\theta_\varepsilon) - \mathbb{E}[f(Z_{1:s}\theta_\varepsilon)]| + 2L\|\theta - \theta_\varepsilon\|$$

Thus we have that:

$$\sup_{\theta\in\Theta} |\mathbb{G}_{s,n} f(\cdot;\theta) - \mathbb{E}[f(Z_{1:s};\theta)]| \leq \sup_{\theta\in\Theta_\varepsilon} |\mathbb{G}_{s,n} f(\cdot;\theta_\varepsilon) - \mathbb{E}[f(Z_{1:s}\theta_\varepsilon)]| + 2L\varepsilon$$

Taking a union bound over $\theta \in \Theta_\varepsilon$, we have that w.p. $1-\delta$:

$$\sup_{\theta\in\Theta_\varepsilon} |\mathbb{G}_{s,n} f(\cdot;\theta_\varepsilon) - \mathbb{E}[f(Z_{1:s}\theta_\varepsilon)]| \leq \eta\sqrt{\frac{2\log(R/(\varepsilon\,\delta))}{n/s}} + G\,\frac{2\log(R/(\varepsilon\,\delta))}{3(n/s)}$$

Choosing $\varepsilon = \frac{s\,R}{n}$ and applying the last two inequalities, yields the desired result. □

D. Estimation Error and Asymptotic Normality

Since throughout the section we will fix the target vector x, we will drop it from the notation when possible, e.g. we will let $\theta_0 = \theta_0(x)$ and $\hat{\theta} = \hat{\theta}(x)$. We begin by introducing some quantities that will be useful throughout our theoretical analysis. First we denote with ω the random variable that corresponds to the internal randomness of the tree-splitting algorithm. Moreover, when the tree splitting algorithm is run with target x, an input dataset of $\{Z_i\}_{i=1}^s$ and internal randomness ω, we denote with $\alpha_i\left(\{Z_i\}_{i=1}^s, \omega\right)$ the weight that it assigns to the sample with index i. Finally, for each sub-sample $b = 1 \ldots B$ we denote with S_b the index of the samples chosen and ω_b the internal randomness that was drawn.

We then consider the weighted empirical score, weighted by the sub-sampled ORF weights:

$$\Psi(\theta, h) = \sum_{i=1}^n a_i\,\psi(Z_i;\theta, h(W_i)) = \frac{1}{B}\sum_{b=1}^B \sum_{i\in S_b} \alpha_i\left(\{Z_i\}_{i\in S_b}, \omega_b\right)\,\psi(Z_i;\theta, h(W_i)) \tag{17}$$

We will also be considering the complete multi-dimensional U-statistic, where we average over all sub-samples of size s:

$$\Psi_0(\theta, h) = \binom{n}{s}^{-1} \sum_{1\leq i_1\leq\ldots\leq i_s\leq n} \mathbb{E}_\omega\left[\sum_{t=1}^s \alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^s, \omega\right)\,\psi(Z_{i_t};\theta, h(W_{i_t}))\right]. \tag{18}$$

and we denote with:

$$f(Z_{i_1},\ldots,Z_{i_s};\,\theta, h) = \mathbb{E}_\omega\left[\sum_{t=1}^s \alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^s, \omega\right)\,\psi(Z_{i_t};\theta, h(W_{i_t}))\right] \tag{19}$$

First, we will bound the estimation error as a sum of $m(x;\theta,\hat{h})$ and second order terms. The proof follows from the Taylor expansion of the moment function and the mean-value theorem.

Lemma D.1. *Under Assumption 4.1, for any nuisance estimate $\hat{h}$ and for the ORF estimate $\hat{\theta}$ estimated with plug-in nuisance estimate $\hat{h}$:*

$$\hat{\theta} - \theta_0 = M^{-1}\left(m(x;\hat{\theta},\hat{h}) - \Psi(\hat{\theta},\hat{h})\right) + \xi$$

where ξ satisfies $\|\xi\| = \mathcal{O}\left(\mathbb{E}\left[\|\hat{h}(W) - h_0(W)\|^2 \mid x\right] + \|\hat{\theta} - \theta_0\|^2\right)$.

Proof outline of main theorems. We will now give a rough outline of the proof of our main results. In doing so we will also present some core technical Lemmas that we will use in the formal proofs of these theorems in the subsequent corresponding subsections.

Lemma D.1 gives rise to the following core quantity:

$$\Lambda(\theta, h) = m(x; \theta, h) - \Psi(\theta, h) \tag{20}$$

Suppose that our first stage estimation rate guarantees a local root-mean-squared-error (RMSE) of χ_n, i.e.:

$$\mathcal{E}(h) = \sqrt{\mathbb{E}\left[\|\hat{h}(W) - h_0(W)\|^2 \mid x\right]} \leq \chi_n \tag{21}$$

Then we have that:

$$\hat{\theta} - \theta_0 = M^{-1}\,\Lambda(\hat{\theta}, \hat{h}) + O(\chi_n^2 + \|\hat{\theta} - \theta_0\|^2)$$

Thus to understand the estimation error of $\hat{\theta}$ and its asymptotic distribution, we need to analyze the concentration of $\Lambda(\hat{\theta}, \hat{h})$ around zero and its asymptotic distribution. Subsequently, invoking consistency of $\hat{\theta}$ and conditions on a sufficiently fast nuisance estimation rate χ_n, we would be able to show that the remainder terms are asymptotically negligible.

Before delving into our two main results on mean absolute error (MAE) and asymptotic normality we explore a bit more the term $\Lambda(\theta, h)$ and decompose it into three main quantities, that we will control each one separately via different arguments.

Lemma D.2 (Error Decomposition)**.** *For any θ, h, let $\mu_0(\theta, h) = \mathbb{E}\left[\Psi_0(\theta, h)\right]$. Then:*

$$\Lambda(\theta, h) = \underbrace{m(x; \theta, h) - \mu_0(\theta, h)}_{\Gamma(\theta,h) \text{ = kernel error}} + \underbrace{\mu_0(\theta, h) - \Psi_0(\theta, h)}_{\Delta(\theta,h) \text{ = sampling error}} + \underbrace{\Psi_0(\theta, h) - \Psi(\theta, h)}_{E(\theta,h) \text{ = subsampling error}}\ . \tag{22}$$

When arguing about the MAE of our estimator, the decomposition presented in Lemma D.2 is sufficient to give us the final result by arguing about concentration of each of the terms. However, for asymptotic normality we need to further refine the decomposition into terms that when scaled appropriately converge to zero in probability and terms that converge to a normal random variable. In particular, we need to further refine the sampling error term $\Delta(\theta, h)$ as follows:

$$\Delta(\theta, h) = \underbrace{\Delta(\theta_0, \tilde{h}_0)}_{\text{asymptotically normal term}} + \underbrace{\Delta(\theta, h) - \Delta(\theta_0, \tilde{h}_0)}_{F(\theta,h) \text{ = stochastic equicontinuity term}} \tag{23}$$

for some appropriately defined fixed function $\tilde{h}_0$. Consider for instance the case where θ is a scalar. If we manage to show that there exists a scaling σ_n, such that $\sigma_n^{-1}\Delta(\theta_0, \tilde{h}_0) \rightarrow_d \mathcal{N}(0, 1)$, and all other terms Γ, E, F and χ_n^2 converge to zero in probability when scaled by σ_n^{-1}, then we will be able to conclude by Slutzky's theorem that: $\sigma_n^{-1}M\left(\hat{\theta} - \theta_0\right) \rightarrow_d \mathcal{N}(0, 1)$ and establish the desired asymptotic normality result.

Since controlling the convergence rate to zero of the terms Γ, Δ, E would be useful in both results, we provide here three technical lemmas that control these rates.

Lemma D.3 (Kernel Error)**.** *If the ORF weights when trained on a random sample $\{Z_i\}_{i=1}^s$, satisfy that:*

$$\mathbb{E}\left[\sup\{\|X_i - x\| : a_i(\{Z_i\}_{i=1}^s, \omega) > 0\}\right] \leq \varepsilon(s) \tag{24}$$

where expectation is over the randomness of the samples and the internal randomness ω of the ORF algorithm. Then

$$\sup_{\theta,h} \|\Gamma(\theta, h)\| = \sqrt{p}\, L\, \varepsilon(s) \tag{25}$$

Lemma D.4 (Sampling Error)**.** *Under Assumption 4.1, conditional on any nuisance estimate $\hat{h}$ from the first stage, with probability $1 - \delta$:*

$$\sup_{\theta} \|\Delta(\theta, \hat{h})\| = O\left(\sqrt{\frac{s\left(\log(n/s) + \log(1/\delta)\right)}{n}}\right) \tag{26}$$

Proof. Since $\Delta(\theta, \hat{h})$ is a U-statistic as it can be written as: $\mathbb{G}_{s,n} f(\cdot;\theta,\hat{h}) - \mathbb{E}\left[f(Z_{1:s};\theta,\hat{h})\right]$. Moreover, under Assumption 4.1, the function $f(\cdot;\theta,\hat{h})$ satisfies the conditions of Lemma C.1 with $\eta = G = \psi_{\max} = O(1)$. Moreover, $f(\cdot;\theta,\hat{h})$ is L-Lipschitz for $L = O(1)$, since it is a convex combination of $O(1)$-Lipschitz functions. Finally, $\text{diam}(\Theta) = O(1)$. Thus applying Lemma C.1, we get, the lemma. □

Lemma D.5 (Subsampling Error)**.** *If the ORF weights are built on B randomly drawn sub-samples with replacement, then*

$$\sup_{\theta,h} \|E(\theta,\hat{h})\| = O\left(\frac{\log(B) + \log(1/\delta)}{\sqrt{B}}\right) \tag{27}$$

D.1. Consistency of ORF Estimate

Theorem D.6 (Consistency)**.** *Assume that the nuisance estimate satisfies:*

$$\mathbb{E}\left[\mathcal{E}(\hat{h})\right] = o(1) \tag{28}$$

and that $B \geq n/s$, $s = o(n)$ and $s \to \infty$ as $n \to \infty$. Then the ORF estimate $\hat{\theta}$ satisfies:

$$\|\hat{\theta} - \theta_0(x)\| = o_p(1)$$

Moreover, for any constant integer $q \geq 1$:

$$\left(\mathbb{E}[\|\hat{\theta} - \theta_0\|^{2q}]\right)^{1/q} = o\left(\left(\mathbb{E}\left[\|\hat{\theta} - \theta_0\|^q\right]\right)^{1/q}\right) \tag{29}$$

Proof. By the definition of $\hat{\theta}$, we have that: $\Psi(\hat{\theta},\hat{h}) = 0$. Thus we have that:

$$\left\|m(x;\hat{\theta},\hat{h})\right\| = \left\|m(x;\hat{\theta},\hat{h}) - \Psi(\hat{\theta},\hat{h}\right\| = \left\|\Lambda(\hat{\theta},\hat{h})\right\|$$

By Lemmas 3.1, D.2, D.3, D.4 and D.5, we have that with probability $1 - 2\delta$:

$$\left\|\Lambda(\hat{\theta},\hat{h})\right\| = O\left(s^{-1/(2\alpha d)} + \sqrt{\frac{s\left(\log(n/s) + \log(1/\delta)\right)}{n}} + \sqrt{\frac{\log(B) + \log(1/\delta)}{B}}\right)$$

Integrating this tail bound we get that:

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

Thus if $B \geq n/s$, $s = o(n)$ and $s \to \infty$ then all terms converge to zero as $n \to \infty$.

Since $\psi(x;\theta,h(w))$ is L-Lipschitz in $h(w)$ for some constant L:

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

Moreover, by our consistency guarantee on $\hat{h}$:

$$\mathbb{E}\left[\|m(x;\hat{\theta},h_0) - m(x;\hat{\theta},\hat{h})\|\right] \leq L\mathbb{E}\left[\mathcal{E}(\hat{h})\right] = o(1)$$

Thus we conclude that:

$$\mathbb{E}[\|m(x;\hat{\theta},h_0)\|] = o(1)$$

which implies that $\|m(x;\hat{\theta},h_0)\| = o_p(1)$.

By our first assumption, for any ε, there exists a δ, such that: $\Pr[\|\hat{\theta}-\theta_0(x)\| \geq \varepsilon] \leq \Pr[\|m(x;\hat{\theta},h_0)\| \geq \delta]$. Since $\|m(x;\hat{\theta},h_0)\| = o_p(1)$, the probability on the right-hand-side converges to 0 and hence also the left hand side. Hence, $\|\hat{\theta}-\theta_0(x)\| = o_p(1)$.

We now prove the second part of the theorem which is a consequence of consistency. By consistency of $\hat{\theta}$, we have that for any ε and δ, there exists $n^*(\varepsilon,\delta)$ such that for all $n \geq n(\varepsilon,\delta)$:

$$\Pr\left[\|\hat{\theta}-\theta_0\| \geq \varepsilon\right] \leq \delta$$

Thus for any $n \geq n^*(\varepsilon,\delta)$:

$$\mathbb{E}[\|\hat{\theta}-\theta_0\|^{2q}] \leq \varepsilon^q \mathbb{E}\left[\|\hat{\theta}-\theta_0\|^q\right] + \delta \mathbb{E}\left[|\hat{\theta}-\theta_0\|^{2q}\right]$$

Choosing $\varepsilon = (4C)^{-q}$ and $\delta = (4C)^{-1}(\text{diam}(\Theta))^{-q}$ yields that:

$$\mathbb{E}[\|\hat{\theta}-\theta_0\|^{2q}] \leq \frac{1}{2C}\mathbb{E}\left[\|\hat{\theta}-\theta_0\|^q\right]$$

Thus for any constant C and for $n \geq n^*((4C)^{-q},(4C)^{-1}(\text{diam}(\Theta))^{-q}) = O(1)$, we get that:

$$\left(\mathbb{E}[\|\hat{\theta}-\theta_0\|^{2q}]\right)^{1/q} = \frac{1}{(2C)^{1/q}}\left(\mathbb{E}\left[\|\hat{\theta}-\theta_0\|^q\right]\right)^{1/q}$$

which concludes the claim that $\left(\mathbb{E}[\|\hat{\theta}-\theta_0\|^{2q}]\right)^{1/q} = o\left(\left(\mathbb{E}\left[\|\hat{\theta}-\theta_0\|^q\right]\right)^{1/q}\right)$. □

D.2. Proof of Theorem 4.2: Mean L^q Estimation Error

Proof. Applying Lemma D.1 and the triangle inequality for the L^q norm, we have that:

$$\left(\mathbb{E}\left[\|\hat{\theta}-\theta_0\|^q\right]\right)^{1/q} = O\left(\left(\mathbb{E}\left[\|\Lambda(\hat{\theta},\hat{h})\|^q\right]\right)^{1/q} + \left(\mathcal{E}(\hat{h})^{2q}\right)^{1/q} + \left(\mathbb{E}\left[\|\hat{\theta}-\theta_0\|^{2q}\right]\right)^{1/q}\right)$$

By assumption $\left(\mathcal{E}(\hat{h})^{2q}\right)^{1/q} \leq \chi^2_{n,2q}$. By the consistency Theorem D.6:

$$\left(\mathbb{E}\left[\|\hat{\theta}-\theta_0\|^{2q}\right]\right)^{1/q} = o\left(\left(\mathbb{E}\left[\|\hat{\theta}-\theta_0\|^q\right]\right)^{1/q}\right).$$

and therefore this term can be ignored for n larger than some constant. Moreover, by Lemma D.2:

$$\left(\mathbb{E}\left[\|\Lambda(\hat{\theta},\hat{h})\|^q\right]\right)^{1/q} = O\left(\left(\mathbb{E}\left[\|\Gamma(\hat{\theta},\hat{h})\|^q\right]\right)^{1/q} + \left(\mathbb{E}\left[\|\Delta(\hat{\theta},\hat{h})\|^q\right]\right)^{1/q} + \left(\mathbb{E}\left[\|E(\hat{\theta},\hat{h})\|^q\right]\right)^{1/q}\right)$$

By Lemma D.3 and Lemma 3.1 we have:

$$\left(\mathbb{E}\left[\|\Gamma(\hat{\theta},\hat{h})\|^q\right]\right)^{1/q} = O\left(\varepsilon(s)\right) = O\left(s^{-1/(2\alpha d)}\right)$$

for a constant $\alpha = \frac{\log(\rho^{-1})}{\pi\log((1-\rho)^{-1})}$. Moreover, by integrating the exponential tail bound provided by the high probability statements in Lemmas D.4 and D.5, we have that for any constant q:

$$\left(\mathbb{E}\left[\|\Delta(\hat{\theta},\hat{h})\|^q\right]\right)^{1/q} = O\left(\sqrt{\frac{s\log(n/s)}{n}}\right)$$

$$\left(\mathbb{E}\left[\|E(\hat{\theta},\hat{h})\|^q\right]\right)^{1/q} = O\left(\sqrt{\frac{\log(B)}{B}}\right)$$

For $B > n/s$, the second term is negligible compared to the first and can be ignored. Combining all the above inequalities:

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

□

D.3. Proof of Theorem 4.3: Finite Sample High Probability Error Bound for Gradients of Convex Losses

Proof. We condition on the event that $\mathcal{E}(\hat{h}) \leq \chi_{n,\delta}$, which occurs with probability $1-\delta$. Since the Jacobian of $m(x;\theta,h_0)$ has eigenvalues lower bounded by σ and each entry of the Jacobian of ψ is L-Lispchitz with respect to the nuisance for some constant L, we have that for every vector $\nu \in \mathbb{R}^p$ (with p the dimension of θ_0):

$$\begin{aligned}
\frac{\nu^T \nabla_\theta m(x;\theta,\hat{h})\nu}{\|\nu\|^2} &\geq \frac{\nu^T \nabla_\theta m(x;\theta,h_0)\nu}{\|\nu\|^2} + \frac{\nu^T \nabla_\theta \left(m(x;\theta,\hat{h}) - m(x;\theta,h_0)\right)\nu}{\|\nu\|^2} \\
&\geq \sigma - L \cdot \mathbb{E}\left[\|\hat{h}(W) - h_0(W)\| \mid x\right] \frac{\|\nu\|_1^2}{\|\nu\|^2} \\
&\geq \sigma - L\,\chi_{n,\delta}\,p = \sigma - O(\chi_{n,\delta})
\end{aligned}$$

Where in the last inequality we also used Holder's inequality to upper bound the L^1 norm by the L^2 norm of the first stage error. Thus the expected loss function $L(\theta) = \mathbb{E}\left[\ell(Z;\theta,\hat{h}(W) \mid x\right]$ is $\hat{\sigma} = \sigma - O(\chi_{n,\delta})$ strongly convex, since $\nabla_\theta m(x;\theta,\hat{h})$ is the Hessian of $L(\theta)$. We then have:

$$L(\hat{\theta}) - L(\theta_0) \geq \nabla_\theta L(\theta_0)'(\hat{\theta}-\theta_0) + \frac{\hat{\sigma}}{2}\|\hat{\theta}-\theta_0\|^2 = m(x;\theta_0,\hat{h})'(\hat{\theta}-\theta_0) + \frac{\hat{\sigma}}{2}\|\hat{\theta}-\theta_0\|^2$$

Moreover, by convexity of $L(\theta)$, we have:

$$L(\theta_0) - L(\hat{\theta}) \geq \nabla_\theta L(\hat{\theta})'(\theta_0 - \hat{\theta}) = m(x;\hat{\theta},\hat{h})'(\theta_0-\hat{\theta})$$

Combining the above we get:

$$\frac{\hat{\sigma}}{2}\|\hat{\theta}-\theta_0\|^2 \leq (m(x;\hat{\theta},\hat{h}) - m(x;\theta_0,\hat{h}))'(\hat{\theta}-\theta_0) \leq \|m(x;\hat{\theta},\hat{h}) - m(x;\theta_0,\hat{h})\|\,\|\hat{\theta}-\theta_0\|$$

Dividing over by $\|\hat{\theta}-\theta_0\|$, we get:

$$\|\hat{\theta}-\theta_0\| \leq \frac{2}{\hat{\sigma}}\left(\|m(x;\hat{\theta},\hat{h})\| + \|m(x;\theta_0,\hat{h})\|\right)$$

The term $\|m(x;\hat{\theta},\hat{h})\|$ is upper bounded by $\|\Lambda(\hat{\theta},\hat{h})\|$ (since $\Psi(\hat{\theta},\hat{h})=0$). Hence, by Lemmas 3.1, D.2, D.3, D.4 and D.5 and our assumptions on the choice of s, B, we have that with probability $1-2\delta$:

$$\left\|\Lambda(\hat{\theta},\hat{h})\right\| = O\left(s^{-1/(2\alpha d)} + \sqrt{\frac{s\,(\log(n/s) + \log(1/\delta))}{n}}\right)$$

Subsequently, using a second order Taylor expansion around h_0 and orthogonality argument almost identical to the proof of Lemma D.1, we can show that the second term $\|m(x;\theta_0,\hat{h})\|$ is it upper bounded by $O(\chi_{n,\delta}^2)$. More formally, since $m(x;\theta_0,h_0)=0$ and the moment is locally orthogonal, invoking a second order Taylor expansion:

$$\begin{aligned}
m_j(x;\theta_0,\hat{h}) =&\ m_j(x;\theta_0,h_0) + D_{\psi_j}[\hat{h}-h_0 \mid x] + \underbrace{\frac{1}{2}\mathbb{E}\left[(\hat{h}(W)-h_0(W))^\intercal \nabla_h^2 \psi_j(Z;\theta_0,\tilde{h}^{(j)}(W))(\hat{h}(W)-h_0(W)) \mid x\right]}_{\rho_j} \\
=&\ \rho_j
\end{aligned}$$

for some function $\tilde{h}_j$ implied by the mean value theorem. Since the moment is smooth, we have: $\|\rho\| = O\left(\mathbb{E}\left[\left\|\hat{h}(W)-h_0(W)\right\|^2 \mid x\right]\right) = O\left(\chi_{n,\delta}^2\right)$. Thus $\left\|m_j(x;\theta_0,\hat{h})\right\| = O\left(\chi_{n,\delta}^2\right)$. Combining all the latter inequalities yields the result. □

D.4. Proof of Theorem 4.4: Asymptotic Normality

Proof. We want to show asymptotic normality of any fixed projection $\left\langle \beta, \hat{\theta} \right\rangle$ with $\|\beta\| \leq 1$. First consider the random variable $V = \left\langle \beta, M^{-1}\Delta(\theta_0, \tilde{h}_0) \right\rangle$, where $\tilde{h}_0(X, W) = g(W; \nu_0(x))$, i.e. the nuisance function $\tilde{h}_0$ ignores the input X and uses the parameter $\nu_0(x)$ for the target point x. Asymptotic normality of V follows by identical arguments as in (Wager & Athey, 2015) or (Mentch & Hooker, 2016), since this term is equivalent to the estimate of a random forest in a regression setting, where we want to estimate $\mathbb{E}\left[Y \mid X = x\right]$ and where the observation of sample i is:

$$Y_i = \left\langle \beta, M^{-1}\left(m(X_i; \theta_0, \tilde{h}_0) - \psi(Z_i; \theta_0, \tilde{h}_0(X_i, W_i))\right) \right\rangle \tag{30}$$

By Theorem 1 of (Wager & Athey, 2015) and the fact that our forest satisfies Specification 1 and under our set of assumptions, we have, that there exists a sequence σ_n, such that:

$$\sigma_n^{-1} V \to \mathcal{N}(0, 1) \tag{31}$$

for $\sigma_n = \Theta\left(\sqrt{\text{polylog}(n/s)^{-1}\, s/n}\right)$. More formally, we check that each requirement of Theorem 1 of (Wager & Athey, 2015) is satisfed:

(i) We assume that the distribution of X admits a density that is bounded away from zero and infinity,

(ii) $\mathbb{E}[Y|X = x^*] = 0$ and hence is continuous in x^* for any x^*,

(iii) The variance of the Y conditional on $X = x^*$ for some x^* is:

$$\text{Var}(Y|X = x^*) = \mathbb{E}\left[\left\langle \beta, M^{-1}\,\psi(Z; \theta_0, \tilde{h}_0(X, W)) \right\rangle^2 \mid X = x^*\right] - \mathbb{E}\left[\left\langle \beta, M^{-1}\,\psi(Z; \theta_0, \tilde{h}_0(X, W)) \mid X = x^* \right\rangle\right]^2$$

The second term is $O(1)$-Lipschitz in x^* by Lipschitzness of $m(x^*; \theta_0, \tilde{h}_0) = \mathbb{E}\left[\psi(Z; \theta_0, \tilde{h}_0(X, W)) \mid X = x^*\right]$. For simplicity of notation consider the random variable $V = \psi(Z; \theta_0, \tilde{h}_0(X, W))$. Then the first part is equal to some linear combination of the covariance terms:

$$Q(x^*) \triangleq \mathbb{E}\left[\beta^\intercal M^{-1} V V^T (M^{-1})^\intercal \beta \mid X = x^*\right] = \beta^\intercal M^{-1}\mathbb{E}\left[V V^T \mid X = x^*\right](M^{-1})^\intercal \beta =$$

Thus by Lipschitzness of the covariance matrix of ψ, we have that: $\|\mathbb{E}\left[VV^\intercal \mid X = x^*\right] - \mathbb{E}\left[VV^\intercal \mid X = \tilde{x}\right]\|_F \leq L\|x^* - \tilde{x}\|$ and therefore by the Cauchy-Schwarz inequality and the lower bound $\sigma > 0$ on the eigenvalues of M:

$$|Q(x^*) - Q(\tilde{x})| \leq L\|x^* - \tilde{x}\| \|\beta^\intercal M^{-1}\|^2 \leq \frac{L}{\sigma^2}\|x^* - \tilde{x}\|$$

Thus $\text{Var}(Y|X = x^*)$ is $O(1)$-Lipschitz continuous in x^* and hence also $\mathbb{E}[Y^2|X = x^*]$ is $O(1)$-Lipschitz continuous.

(iv) The fact that $\mathbb{E}[|Y - \mathbb{E}[Y|X = x]|^{2+\delta}|X = x] \leq H$ for some constant δ, H follows by our assumption on the boundedness of ψ and the lower bound on the eigenvalues of M,

(v) The fact that $\text{Var}[Y|X = x'] > 0$ follows from the fact that $\text{Var}\left(\beta^\intercal M^{-1}\psi(Z; \theta_0, \tilde{h}_0(X, W)) \mid X = x'\right) > 0$,

(vi) The fact that tree is honest, α-balanced with $\alpha \leq 0.2$ and symmetric follows by Specification 1,

(vii) From our assumption on s that $s^{-1/(2\alpha d)} = o((s/n)^{1/2-\varepsilon})$, it follows that $s = \Theta(n^\beta)$ for some $\beta \in \left(1 - \frac{1}{1+\alpha d}, 1\right]$.

Since, by Lemmas D.1, D.2 and Equation (23):

$$\left\|\left\langle \beta, \hat{\theta} - \theta_0 \right\rangle - V\right\| = O\left(\|\Gamma(\hat{\theta}, \hat{h})\| + \|\Delta(\hat{\theta}, \hat{h}) - \Delta(\theta_0, \tilde{h}_0)\| + \|\mathcal{E}(\hat{h})\|^2 + \|\hat{\theta} - \theta_0\|^2\right)$$

it suffices to show that:

$$\sigma_n^{-1}\mathbb{E}\left[\|\Gamma(\hat{\theta}, \hat{h})\| + \|\Delta(\hat{\theta}, \hat{h}) - \Delta(\theta_0, \tilde{h}_0)\| + \|\mathcal{E}(\hat{h})\|^2 + \|\hat{\theta} - \theta_0\|^2\right] \to 0$$

as then by Slutzky's theorem we have that $\sigma_n^{-1}\left\langle \beta, \hat{\theta} - \theta_0 \right\rangle \to_d \mathcal{N}(0,1)$. The first term is of order $O(s^{-1/(2\alpha d)})$, hence by our assumption on the choice of s, it is $o(\sigma_n)$. The third term is $O(\chi_{n,2}^2) = O(\chi_{n,4}^2)$, which by assumption is also $o(\sigma_n)$. The final term, by applying our L^q estimation error result for $q = 2$ and the assumption on our choice of s, we get that it is of order $O\left(\frac{s\log(n/s)}{n}\right) = o(\sigma_n)$.

Thus it remains to bound the second term. For that we will invoke the stochastic equicontinuity Lemma C.1. Observe that each coordinate j of the term corresponds to the deviation from its mean of a U statistic with respect to the class of functions:

$$\gamma_j(\cdot;\theta,\hat{h}) = f_j(\cdot;\theta,\hat{h}) - f_j(\cdot;\theta_0,h_0) \tag{32}$$

Observe that by Lipschitzness of ψ with respect to θ and the output of h and the locally parametric form of h, we have that:

$$\begin{aligned}
|\gamma_j(Z_{1:s};\theta,h)| = &\left|\mathbb{E}_\omega\left[\sum_{t=1}^{s}\alpha_t\left(\{Z_t\}_{t=1}^s,\omega\right)\left(\psi_j(Z_t;\theta,\hat{h}(W_t)) - \psi_j(Z_t;\theta,\tilde{h}_0(W_t))\right)\right]\right| \\
\leq\ &\mathbb{E}_\omega\left[\sum_{t=1}^{s}\alpha_t\left(\{Z_t\}_{t=1}^s,\omega\right)\left|\psi_j(Z_t;\theta,\hat{h}(W_t)) - \psi_j(Z_t;\theta_0,\tilde{h}_0(W_t))\right|\right] \\
\leq\ &L\,\mathbb{E}_\omega\left[\sum_{t=1}^{s}\alpha_t\left(\{Z_t\}_{t=1}^s,\omega\right)\left(\|\theta-\theta_0\| + \|g(W_t;\nu) - g(W_t;\nu_0(x))\|\right)\right] \\
\leq\ &L\,\mathbb{E}_\omega\left[\sum_{t=1}^{s}\alpha_t\left(\{Z_t\}_{t=1}^s,\omega\right)\left(\|\theta-\theta_0\| + L\,\|\nu - \nu_0(x)\|\right)\right] \\
=\ &L\left(\|\theta-\theta_0\| + L\,\|\nu-\nu_0(x)\|\right)
\end{aligned}$$

Thus by Jensen's inequality and the triangle inequality:

$$\sqrt{\mathbb{E}\left[|\gamma_j(Z_{1:s};\theta,h)|^2\right]} \leq L\|\theta-\theta_0\| + L^2\,\|\nu-\nu_0(x)\|$$

Thus:

$$\sup_{\theta:\|\theta-\theta_0\|\leq\eta,\|\nu-\nu_0(x)\|\leq\gamma}\sqrt{\mathbb{E}\left[|\gamma_j(Z_{1:s};\theta,g(\cdot;\nu))|^2\right]} = O(\eta+\gamma)$$

By our L^q error result and Markov's inequality, we have that with probability $1-\delta$: $\|\hat{\theta}-\theta_0\| \leq \eta = O(\sigma_n/\delta)$. Similarly, by our assumption on the nuisance error $\left(\mathbb{E}\left[\|\hat{\nu}-\nu_0(x)\|^4\right]\right)^{1/4} \leq \chi_{n,4}$ and Markov's inequality we have that with probability $1-\delta$: $\|\hat{\nu}-\nu_0(x)\| \leq O(\chi_{n,4}/\delta)$. Thus applying Lemma C.1, we have that conditional on the event that $\|\hat{\nu}-\nu_0(x)\| \leq O(\chi_{n,4}/\delta)$, w.p. $1-\delta$:

$$\begin{aligned}
\sup_{\theta:\|\theta-\theta_0\|\leq\sigma_n/\delta}\sqrt{\mathbb{E}\left[|\gamma_j(Z_{1:s};\theta,\hat{h})|^2\right]} &= O\left((\sigma_n/\delta + \chi_{n,4}/\delta)\sqrt{\frac{s(\log(n/s)+\log(1/\delta))}{n}} + \frac{s(\log(n/s)+\log(1/\delta))}{n}\right) \\
&= O\left(\sigma_n^2\,\mathrm{polylog}(n/s)/\delta + \chi_{n,4}\sigma_n\,\mathrm{polylog}(n/s)/\delta + \frac{s(\log(n/s)+\log(1/\delta))}{n}\right) \\
&= O(\sigma_n^{3/2}\,\mathrm{polylog}(n/s)/\delta)
\end{aligned}$$

where we used the fact that $\chi_{n,4}^2 = o(\sigma_n)$, $\sqrt{\log(1/\delta)} \leq 1/\delta$ and that $\sigma_n = \Theta\left(\sqrt{\mathrm{polylog}(n/s)^{-1}\,s/n}\right)$. By a union bound we have that w.p. $1-3\delta$:

$$\|\Delta(\hat{\theta},\hat{h}) - \Delta(\theta_0,\tilde{h}_0)\| = O(\sigma_n^{3/2}\,\mathrm{polylog}(n/s)/\delta)$$

Integrating this tail bound and using the boundedness of the score we get:

$$\mathbb{E}\left[\|\Delta(\hat{\theta},\hat{h}) - \Delta(\theta_0,\tilde{h}_0)\|\right] = O(\sigma_n^{3/2}\,\mathrm{polylog}(n/s)\log(1/\sigma_n)) = o(\sigma_n) \tag{33}$$

This completes the proof of the theorem. □

D.5. Omitted Proofs of Technical Lemmas

Proof of Lemma D.1. Fix a conditioning vector x. By performing a second order Taylor expansion of each coordinate $j \in [p]$ of the expected score function m_j around the true parameters $\theta_0 = \theta_0(x)$ and h_0 and applying the multi-dimensional mean-value theorem, we can write that for any $\theta \in \Theta$:

$$\begin{aligned} m_j(x; \theta, \hat{h}) = {} & m_j(x; \theta_0, h_0) + \nabla_\theta m_j(x; \theta_0, h_0)'(\theta - \theta_0) + D_{\psi_j}[\hat{h} - h_0 \mid x] \\ & \underbrace{+ \frac{1}{2}\,\mathbb{E}\left[(\theta - \theta_0, \hat{h}(W) - h_0(W))^\intercal \nabla^2_{\theta,h}\psi_j(Z; \tilde{\theta}^{(j)}, \tilde{h}^{(j)}(W))(\theta - \theta_0, \hat{h}(W) - h_0(W)) \mid x\right]}_{\rho_j} \end{aligned}$$

where each $\tilde{\theta}^{(j)}$ is some convex combination of θ and θ_0 and each $\tilde{h}^{(j)}(W)$ is some convex combination of $\hat{h}(W)$ and $h_0(W)$. Note that $m(x; \theta_0, h_0) = 0$ by definition and $D_{\psi_j}[\hat{h} - h_0 \mid x] = 0$ by local orthogonality. Let ρ denote the vector of second order terms. We can thus write the above set of equations in matrix form as:

$$M(\theta - \theta_0) = m(x; \theta, \hat{h}) - \rho$$

where we remind that $M = \nabla_\theta m(x; \theta_0, h_0)$ is the Jacobian of the moment vector. Since by our assumptions M is invertible and has eigenvalues bounded away from zero by a constant, we can write:

$$(\theta - \theta_0) = M^{-1}\, m(x; \theta, \hat{h}) - M^{-1}\,\rho$$

Letting $\xi = -M^{-1}\,\rho$, we have that by the boundedness of the eigenvalues of M^{-1}:

$$\|\xi\| = O(\|\rho\|)$$

By our bounded eigenvalue Hessian assumption on $\mathbb{E}\left[\nabla^2_{\theta,h}\psi_j(Z; \tilde{\theta}^{(j)}, \tilde{h}^{(j)}(W)) \mid x, W\right]$, we know that:

$$\|\rho\|_\infty = O\left(\mathbb{E}\left[\|\hat{h}(W) - h_0(W)\|^2 \mid x\right] + \|\theta - \theta_0\|^2\right)$$

Combining the above two equations and using the fact that $\|\rho\| \leq \sqrt{p}\|\rho\|_\infty$, yields that for any $\theta \in \Theta$:

$$\theta - \theta_0 = M^{-1}\,\left(m(x; \theta, \hat{h}) - \Psi(\theta, \hat{h})\right) + \xi$$

Evaluating the latter at $\theta = \hat{\theta}$ and also observing that by the definition of $\hat{\theta}$, $\Psi(\hat{\theta}, \hat{h}) = 0$ yields the result. □

Proof of Lemma D.3. First we argue that by invoking the honesty of the ORF weights we can re-write $\mu_0(\theta, h)$ as:

$$\mu_0(\theta, h) = \mathbb{E}\left[\binom{n}{s}^{-1} \sum_{1 \leq i_1 \leq \ldots \leq i_s \leq n} \mathbb{E}_\omega\left[\sum_{t=1}^{s} \alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^s, \omega\right)\, m(X_{i_t}; \theta, h)\right]\right] \tag{34}$$

To prove this claim, it suffices to show that for any subset of s indices:

$$\mathbb{E}\left[\alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^s, \omega\right)\, \psi(Z_{i_t}; \theta, h)\right] = \mathbb{E}\left[\alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^s, \omega\right)\, m(X_{i_t}; \theta, h)\right] \tag{35}$$

By honesty of the ORF weights, we know that either $i_t \in S^1$, in which case $\alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^s, \omega\right) = 0$, or otherwise $i_t \in S^2$ and then $\alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^s, \omega\right)$ is independent of Z_{i_t}, conditional on $X_{i_t}, Z_{-i_t}, \omega$. Thus in any case $\alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^s, \omega\right)$ is independent of Z_{i_t}, conditional on $X_{i_t}, Z_{-i_t}, \omega$. Moreover since Z_{i_t} is independent of Z_{-i_t}, ω conditional on X_{i_t}:

$$\mathbb{E}\left[\psi(Z_{i_t}; \theta, h) \mid X_{i_t}, Z_{-i_t}, \omega\right] = \mathbb{E}\left[m(X_{i_t}; \theta, h) \mid X_{i_t}, Z_{-i_t}, \omega\right]$$

By the law of iterated expectation and the independence properties claimed above, we can write:

$$\begin{aligned} \mathbb{E}\left[\alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^s, \omega\right)\, \psi(Z_{i_t}; \theta, h)\right] = & = \mathbb{E}\left[\mathbb{E}\left[\alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^s, \omega\right) \mid X_{i_t}, Z_{-i_t}, \omega\right]\, \mathbb{E}\left[\psi(Z_{i_t}; \theta, h) \mid X_{i_t}, Z_{-i_t}, \omega\right]\right] \\ & = \mathbb{E}\left[\mathbb{E}\left[\alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^s, \omega\right) \mid X_{i_t}, Z_{-i_t}, \omega\right]\, \mathbb{E}\left[m(X_{i_t}; \theta, h) \mid X_{i_t}, Z_{-i_t}, \omega\right]\right] \\ & = \mathbb{E}\left[\alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^s, \omega\right)\, m(X_{i_t}; \theta, h)\right] \end{aligned}$$

Finally, by a repeated application of the triangle inequality and the lipschitz property of the conditional moments, we have:

$$\begin{aligned}
\|\Gamma(\theta, h)\| &\leq \binom{n}{s}^{-1} \sum_{1\leq i_1 \leq \ldots \leq i_s \leq n} \mathbb{E}\left[\sum_{t=1}^{s} \alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^{s}, \omega\right) \|m(x;\theta,h) - m(X_{i_t};\theta,h)\|\right] \\
&\leq \sqrt{p}\, L \binom{n}{s}^{-1} \sum_{1\leq i_1 \leq \ldots \leq i_s \leq n} \mathbb{E}\left[\sum_{t=1}^{s} \alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^{s}, \omega\right) \|X_{i_t} - x\|\right] \\
&\leq \sqrt{p}\, L \binom{n}{s}^{-1} \sum_{1\leq i_1 \leq \ldots \leq i_s \leq n} \mathbb{E}\left[\sup\{\|X_{i_t} - x\| : \alpha_{i_t}\left(\{Z_{i_t}\}_{t=1}^{s}, \omega\right) > 0\}\right] \\
&\leq \sqrt{p}\, L\, \varepsilon(s)
\end{aligned}$$

□

Proof of Lemma D.5. We prove that the concentration holds conditional on the samples $Z_{1:n}$ and $\hat{h}$, the result then follows. Let

$$\tilde{f}(S_b, \omega_b; \theta, h) = \sum_{i \in S_b} \alpha_i\left(S_b, \omega_b\right) \psi(Z_i; \theta, h(W_i)).$$

Observe that conditional on $Z_{1:n}$ and $\hat{h}$, the random variables $\tilde{f}(S_1, \omega_1; \theta, h), \ldots, \tilde{f}(S_B, \omega_B; \theta, h)$ are conditionally independent and identically distributed (where the randomness is over the choice of the set S_b and the internal algorithm randomness ω_b). Then observe that we can write $\Psi(\theta, h) = \frac{1}{B}\sum_{b=1}^{B} \tilde{f}(S_b, \omega_b; \theta, h)$. Thus conditional on $Z_{1:n}$ and $\hat{h}$, $\Psi(\theta, \hat{h})$ is an average of B independent and identically distributed random variables. Moreover, since S_b is drawn uniformly at random among all sub-samples of $[n]$ of size s and since the randomness of the algorithm is drawn identically and independently on each sampled tree:

$$\mathbb{E}\left[\tilde{f}(S_b, \omega_b; \theta, \hat{h}) \mid Z_{1:n}\right] = \binom{n}{s}^{-1} \sum_{1\leq i_1 \leq \ldots \leq i_s \leq n} f(\{Z_{i_t}\}_{t=1}^{s}; \theta, \hat{h}) = \Psi_0(\theta, h)$$

Finally, observe that under Assumption 4.1, $|\tilde{f}(S_b, \omega_b; \theta, \hat{h})| \leq \psi_{\max} = O(1)$ a.s.. Thus by a Chernoff bound, we have that for any fixed $\theta \in \Theta$, w.p. $1-\delta$:

$$\|\Psi(\theta, \hat{h}) - \Psi_0(\theta, \hat{h})\| \leq O\left(\sqrt{\frac{\log(1/\delta)}{B}}\right)$$

Since Θ has constant diameter, we can construct an ε-cover of Θ of size $O(1/\varepsilon)$. By Lipschitzness of ψ with respect to θ and following similar arguments as in the proof of Lemma C.1, we can also get a uniform concentration:

$$\|\Psi(\theta, \hat{h}) - \Psi_0(\theta, \hat{h})\| \leq O\left(\sqrt{\frac{\log(B) + \log(1/\delta)}{B}}\right)$$

□

E. Omitted Proofs from Section 5

Proof of Theorem 5.2. By convexity of the loss ℓ and the fact that $\hat{\nu}(x)$ is the minimizer of the weighted penalized loss, we have:

$$\begin{aligned}
\lambda\left(\|\nu_0(x)\|_1 - \|\hat{\nu}(x)\|_1\right) &\geq \sum_{i=1}^{n} a_i(x)\, \ell(Z_i; \hat{\nu}(x)) - \sum_{i=1}^{n} a_i(x)\, \ell(Z_i; \nu_0(x)) && \text{(optimality of } \hat{\nu}(x)\text{)} \\
&\geq \sum_{i=1}^{n} a_i(x) \left\langle \nabla_\nu \ell(z_i; \nu_0(x)), \hat{\nu}(x) - \nu_0(x)\right\rangle && \text{(convexity of } \ell\text{)} \\
&\geq -\left\|\sum_i a_i(x) \nabla_\nu \ell(z_i; \nu_0(x))\right\|_\infty \|\hat{\nu}(x) - \nu_0(x)\|_1 && \text{(Cauchy-Schwarz)} \\
&\geq -\frac{\lambda}{2} \|\hat{\nu}(x) - \nu_0(x)\|_1 && \text{(assumption on } \lambda\text{)}
\end{aligned}$$

If we let $\rho(x) = \hat{\nu}(x) - \nu_0(x)$, then observe that by the definition of the support S of $\nu_0(x)$ and the triangle inequality, we have:

$$\begin{aligned}
\|\nu_0(x)\|_1 - \|\hat{\nu}(x)\|_1 &= \|\nu_0(x)_S\|_1 + \|\nu_0(x)_{S^c}\|_1 - \|\hat{\nu}(x)_S\|_1 - \|\hat{\nu}(x)_{S^c}\|_1 && \text{(separability of } \ell_1 \text{ norm)}\\
&= \|\nu_0(x)_S\|_1 - \|\hat{\nu}(x)_S\|_1 - \|\hat{\nu}(x)_{S^c}\|_1 && \text{(definition of support)}\\
&= \|\nu_0(x)_S\|_1 - \|\hat{\nu}(x)_S\|_1 - \|\hat{\nu}(x)_{S^c} - \nu_0(x)_{S^c}\|_1 && \text{(definition of support)}\\
&= \|\nu_0(x)_S - \hat{\nu}(x)_S\|_1 - \|\hat{\nu}(x)_{S^c} - \nu_0(x)_{S^c}\|_1 && \text{(triangle inequality)}\\
&\leq \|\rho(x)_S\|_1 - \|\rho(x)_{S^c}\|_1 && \text{(definition of } \rho(x))
\end{aligned}$$

Thus re-arranging the terms in the latter series of inequalities, we get that $\rho(x) \in C(S(x); 3)$.

We now show that the weighted empirical loss function satisfies a conditional restricted strong convexity property with constant $\hat{\gamma} = \gamma - k\sqrt{s\ln(d_\nu/\delta)/n}$ with probability $1-\delta$. This follows from observing that:

$$H = \nabla_{\nu\nu} \sum_{i=1}^{n} a_i(x)\, \ell(z_i; \nu) = \sum_{i=1}^{n} a_i \nabla_{\nu\nu} \ell(z_i; \nu) \succeq \sum_{i=1}^{n} a_i \mathcal{H}(z_i) = \frac{1}{B} \sum_b \sum_{i \in b} a_{ib}(x) \mathcal{H}(z_i)$$

Thus the Hessian is lower bounded by a matrix whose entries correspond to a Monte-Carlo approximation of the U-statistic:

$$U = \frac{1}{\binom{n}{s}} \sum_{S \subseteq [n] : |S| = s} \frac{1}{s!} \sum_{i \in S} \mathop{\mathbb{E}}_{\omega} \left[a_i(S, \omega) \mathcal{H}(z_i) \right] \tag{36}$$

where Π_s denotes the set of permutations of s elements, S_π denotes the permuted elements of S according to π and S_π^1, S_π^2 denotes the first and second half of the ordered elements of S according to π. Finally, $a_i(S, \omega)$ denotes the tree weight assigned to point i by a tree learner trained on S under random seed ω.

Hence, for sufficiently large B, by a U-statistic concentration inequality (Hoeffding, 1963) and a union bound, each entry will concentrate around the expected value of the U statistic to within $2\sqrt{s\ln(d_\nu/\delta)/n}$, i.e.: with probability $1-\delta$:

$$\left\| \frac{1}{B} \sum_b \sum_{i \in b} a_{ib}(x) \mathcal{H}(z_i) - \mathbb{E}\left[U\right] \right\|_\infty \leq 2\sqrt{\frac{s\ln(d_\nu/\delta)}{n}} \tag{37}$$

Moreover, observe that by the tower law of expectation and by honesty of the ORF trees we can write:

$$\mathbb{E}\left[U\right] = \mathbb{E}\left[\frac{1}{\binom{n}{s}} \sum_{S \subseteq [n] : |S| = s} \frac{1}{s!} \sum_{i \in S} a_i(S, \omega)\, \mathbb{E}\left[\mathcal{H}(z_i) \mid x_i\right] \right] \tag{38}$$

Since each $\mathbb{E}\left[\mathcal{H}(z_i) \mid x_i\right]$ satisfies the restricted eigenvalue condition with constant γ, we conclude that $\mathbb{E}\left[U\right]$ also satisfies the same condition as it is a convex combination of these conditional matrices. Thus for any vector $\rho \in C(S(x); 3)$, we have w.p. $1-\delta$:

$$\begin{aligned}
\rho^T H \rho \geq{}& \rho^T \left(\sum_{i=1}^{n} a_i(x) \mathcal{H}(z_i) \right) \rho && \text{(lower bound on Hessian)}\\
\geq{}& \rho^T\, \mathbb{E}\left[U\right] \rho - 2\sqrt{\frac{s\ln(d_\nu/\delta)}{n}} \|\rho\|_1^2 && \text{(}U\text{-statistic matrix concentration)}\\
\geq{}& \gamma \|\rho\|_2^2 - 2\sqrt{\frac{s\ln(d_\nu/\delta)}{n}} \|\rho\|_1^2 && \text{(restricted strong convexity of population)}\\
\geq{}& \left(\gamma - 32k\sqrt{\frac{s\ln(d_\nu/\delta)}{n}} \right) \|\rho\|_2^2 && (\rho \in C(S(x);3) \text{ and sparsity, imply: } \|\rho\|_1 \leq 4\sqrt{k}\|\rho\|_2)
\end{aligned}$$

Since $\rho(x) \in C(S(x); 3)$ and since the weighted empirical loss satisfies a $\hat{\gamma}$ restricted strong convexity:

$$\begin{aligned}
\sum_{i=1}^{n} a_i(x)\, \ell(z_i; \hat{\nu}(x)) - \sum_{i=1}^{n} a_i(x)\, \ell(z_i; \nu_0(x)) \geq{}& \sum_{i=1}^{n} a_i(x) \left\langle \nabla_\nu \ell(z_i; \nu_0(x)), \hat{\nu} - \nu_0(x) \right\rangle + \hat{\gamma} \|\rho(x)\|_2^2\\
\geq{}& -\frac{\lambda}{2} \|\rho(x)\|_1^2 + \hat{\gamma} \|\rho(x)\|_2^2 && \text{(assumption on } \lambda)
\end{aligned}$$

Combining with the upper bound of $\lambda\left(\|\rho(x)_{S(x)}\|_1 - \|\rho(x)_{S(x)^c}\|_1\right)$ on the difference of the two weighted empirical losses via the chain of inequalities at the beginning of the proof, we get that:

$$\hat{\gamma}\|\rho(x)\|_2^2 \ge \frac{3\lambda}{2}\|\rho(x)_{S(x)}\|_1 - \frac{\lambda}{2}\|\rho(x)_{S(x)^c}\|_1 \le \frac{3\lambda}{2}\|\rho(x)_{S(x)}\|_1 \le \frac{3\lambda\sqrt{k}}{2}\|\rho(x)_{S(x)}\|_2 \le \frac{3\lambda\sqrt{k}}{2}\|\rho(x)\|_2$$

Dividing both sides by $\|\rho(x)\|_2$ and combining with the fact that $\|\rho(x)\|_1 \le 4\sqrt{k}\|\rho(x)\|_2$ yields the first part of the theorem.

Bounding the gradient. Let $\tau = 1/(2\alpha d)$. We first upper bound the expected value of each entry of the gradient. By the shrinkage property of the ORF weights:

$$\begin{aligned}
\left|\sum_{i=1}^n \mathbb{E}\left[a_i(x)\nabla_{\nu_j}\ell(z_i;\nu_0(x)) \mid x_i\right]\right| &\le \left|\mathbb{E}\left[\nabla_{\nu_j}\ell(z;\nu_0(x)) \mid x\right]\right| + \mathbb{E}\left[\sum_{i=1}^n a_i(x)\left|\mathbb{E}\left[\nabla_{\nu_j}\ell(z_i;\nu_0(x)) \mid x_i\right] - \mathbb{E}\left[\nabla_{\nu_j}\ell(z;\nu_0(x)) \mid x\right]\right|\right] \\
&\le \left|\nabla_{\nu_j}L(\nu_0(x);x)\right| + L\,\mathbb{E}\left[\sum_{i=1}^n a_i(x)\,\|x_i - x\|\right] && \text{(Lipschitzness of } \nabla_\nu L(\nu;x)\text{)} \\
&\le \left|\nabla_{\nu_j}L(\nu_0(x);x)\right| + L\,s^{-\tau} && \text{(Kernel shrinkage)} \\
&\le L\,s^{-\tau} && \text{(First order optimality condition of } \nu_0(x)\text{)}
\end{aligned}$$

Moreover, since the quantity $\sum_i a_i(x)\nabla_{\nu_j}\ell(z_i;\nu_0(x))$ is also a Monte-Carlo approximation to an appropriately defined U-statistic (defined analogous to quantity U), for sufficiently large B, it will concentrate around its expectation to within $\sqrt{s\ln(1/\delta)/n}$, w.p. $1-\delta$. Since the absolute value of its expectation is at most $Ls^{-\tau}$, we get that the absolute value of each entry w.p. $1-\delta$ is at most $Ls^{-\tau} + \sqrt{s\ln(1/\delta)/n}$. Thus with a union bound over the p entries of the gradient, we get that uniformly, w.p. $1-\delta$ all entries have absolute values bounded within $Ls^{-\tau} + \sqrt{s\ln(d_\nu/\delta)/n}$. □

F. Omitted Proofs from Heterogeneous Treatment Effects Estimation

We now verify the moment conditions for our CATE estimation satisfies the required conditions in Assumption 4.1.

F.1. Local Orthogonality

Recall that for any observation $Z = (T, Y, W, X)$, any parameters $\theta \in \mathbb{R}^p$, nuisance estimate $\hat{h}$ parameterized by functions q, g, we first consider the following *residualized* score function for PLR is defined as:

$$\psi(Z;\theta, h(X,W)) = \{Y - q(X,W) - \langle\theta, (T - g(X,W))\rangle\}\,(T - g(X,W)), \tag{39}$$

with $h(X,W) = (q(X,W), g(X,W))$.

For discrete treatments, we also consider the following *doubly robust* score function, with each coordinate indexed by treament t defined as:

$$\psi^t(Z;\theta, h(X,W)) = m^t(X,W) + \frac{(Y - m^t(X,W))\,\mathbf{1}[T=t]}{g^t(X,W)} - m^0(X,W) - \frac{\left(Y - m^0(X,W)\right)\mathbf{1}[T=0]}{g^0(X,W)} - \theta^t \tag{40}$$

where $h(X,W) = (m(X,W), g(X,W))$.

Lemma F.1 (Local orthogonality for residualized moments)**.** *The moment condition with respect to the score function ψ defined in* (39) *satisfies conditional orthogonality.*

Proof. We establish local orthogonality via an even stronger *conditional orthogonality*:

$$\mathbb{E}\left[\nabla_h\psi(Z,\theta_0(x), h_0(X,W)) \mid W, x\right] = 0 \tag{41}$$

In the following, we will write $\nabla_h\psi$ to denote the gradient of ψ with respect to the nuisance argument. For any W, x, we can write

$$\mathbb{E}\left[\nabla_h\psi\left(Z;\theta_0(x), (q_0(x,W), g_0(x,W))\right) \mid W, x\right] = \mathbb{E}\left[(T - g_0(x,W), -Y + q_0(x,W) + 2\theta_0(x)^\intercal\,(T - g_0(x,W))) \mid W, x\right]$$

Furthermore, we have $\mathbb{E}\left[T - g_0(x, W) \mid W, x\right] = \mathbb{E}\left[\eta \mid W, x\right] = 0$ and

$$\mathbb{E}\left[-Y + q_0(x, W) + 2\theta_0(x)^\intercal \left(T - g_0(x, W)\right) \mid W, x\right] = \mathbb{E}\left[q_0(x, W) - Y + 2\theta(x)^\intercal \eta \mid W, x\right] = 0$$

where the last equality follows from that $\mathbb{E}\left[\eta \mid W, x\right] = 0$ and $\mathbb{E}\left[\langle W, q_0\rangle - Y \mid W, x\right] = 0$. □

Lemma F.2 (Local orthogonality for doubly robust moments)**.** *The moment condition with respect to the score function ψ defined in* (44) *satisfies conditional orthogonality.*

Proof. For every coordinate (or treatment) t, we have

$$\begin{aligned}
&\mathbb{E}\left[\nabla_g \psi^t\left(Z; \theta_0(x), (m_0(x, W), g_0(x, W))\right) \mid W, x\right] \\
&= \mathbb{E}\left[-\frac{(Y - m_0^t(X, W))\mathbf{1}[T = t]}{(g_0^t(x, W))^2} + \frac{(Y - m_0^0(X, W))\mathbf{1}[T = t]}{(g_0^0(x, W))^2} \mid W, x\right] \\
&= \mathbb{E}\left[-\frac{(Y - m_0^t(X, W))}{(g_0^t(x, W))^2} \mid W, x, T = t\right] \Pr[T = t \mid W, x] \\
&+ \mathbb{E}\left[\frac{(Y - m_0^0(X, W))}{(g_0^0(x, W))^2} \mid W, x, T = 0\right] \Pr[T = 0 \mid W, x] = 0
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\nabla_m \psi^t\left(Z; \theta_0(x), (m_0(x, W), g_0(x, W))\right) \mid W, x\right] \\
&= \mathbb{E}\left[\nabla_m \left(m_0^t(x, W) + \frac{\left(Y - m_0^t(x, W)\right)\mathbf{1}[T = t]}{g_0^t(x, W)} - m_0^0(x, W) - \frac{\left(Y - m_0^0(x, W)\right)\mathbf{1}[T = 0]}{g_0^0(x, W)}\right) \mid W, x\right] \\
&= \mathbb{E}\left[\nabla_m \left(m_0^t(x, W) + \frac{\left(-m_0^t(x, W)\right)\mathbf{1}[T = t]}{g_0^t(x, W)} - m_0^0(x, W) - \frac{\left(-m_0^0(x, W)\right)\mathbf{1}[T = 0]}{g_0^0(x, W)}\right) \mid W, x\right] \\
&= \nabla_m \left(\mathbb{E}\left[m_0^t(x, W) - m_0^t(x, W) - m_0^0(x, W) + m_0^0(x, W) \mid W, x\right]\right) = 0
\end{aligned}$$

This completes the proof. □

F.2. Identifiability

Lemma F.3 (Identifiability for residualized moments.)**.** *As long as $\mu(X, W)$ is independent of η conditioned on X and the matrix $\mathbb{E}\left[\eta\eta^\intercal \mid X = x\right]$ is invertible for any x, the parameter $\theta(x)$ is the unique solution to $m(x; \theta, h) = 0$.*

Proof. The moment conditions $m(x; \theta, h) = 0$ can be written as

$$\mathbb{E}\left[\left\{Y - q_0(X, W) - \theta^\intercal \left(T - g_0(X, W)\right)\right\}\left(T - g_0(X, W)\right) \mid X = x\right] = 0$$

The left hand side can re-written as

$$\begin{aligned}
\mathbb{E}\left[\left\{\langle \eta, \mu_0(X, W)\rangle + \varepsilon - \langle \theta, \eta\rangle\right\}\eta \mid X = x\right] &= \mathbb{E}\left[\left\{\langle \eta, \mu_0(X, W)\rangle - \langle \theta, \eta\rangle\right\}\eta \mid X = x\right] \\
&= \mathbb{E}\left[\langle \mu_0(X, W) - \theta, \eta\rangle\eta \mid X = x\right] \\
&= \mathbb{E}\left[\eta\eta^\intercal \mid X = x\right]\mathbb{E}\left[\mu_0(X, W) - \theta \mid X = x\right]
\end{aligned}$$

Since the conditional expected covariance matrix $\mathbb{E}\left[\eta\eta^\intercal \mid X = x\right]$ is invertible, the expression above equals to zero only if $\mathbb{E}\left[\mu_0(X, W) - \theta \mid X = x\right] = 0$. This implies that $\theta = \mathbb{E}\left[\mu_0(X, W) \mid X = x\right] = \theta_0(x)$. □

Lemma F.4 (Identifiability for doubly robust moments.)**.** *As long as $\mu(X, W)$ is independent of η conditioned on X and the matrix $\mathbb{E}\left[\eta\eta^\intercal \mid X = x\right]$ is invertible for any x, the parameter $\theta(x)$ is the unique solution to $m(x; \theta, h) = 0$.*

Proof. For each coordinate t, the moment condition can be written as

$$\mathbb{E}\left[m_0^t(X,W) + \frac{(Y - m_0^t(X,W))\,\mathbf{1}[T=t]}{g_0^t(X,W)} - m_0^0(X,W) - \frac{\left(Y - m_0^0(X,W)\right)\mathbf{1}[T=0]}{g_0^0(X,W)} - \theta^t \mid X = x\right] = 0 \quad (42)$$

Equivalently,

$$\mathbb{E}\left[m_0^t(X,W) - m_0^0(X,W) - \theta^t \mid X = x\right] = \mathop{\mathbb{E}}_{W}\left[\mathbb{E}\left[-\frac{(Y - m_0^t(X,W))\,\mathbf{1}[T=t]}{g_0^t(X,W)} + \frac{\left(Y - m_0^0(X,W)\right)\mathbf{1}[T=0]}{g_0^0(X,W)} \mid W, X = x\right]\right]$$

The inner expectation of the right hand side can be written as:

$$\mathbb{E}\left[-\frac{(Y - m_0^t(X,W))\,\mathbf{1}[T=t]}{g_0^t(X,W)} + \frac{\left(Y - m_0^0(X,W)\right)\mathbf{1}[T=0]}{g_0^0(X,W)} \mid W, X = x\right] = 0$$

This means the moment condition is equivalent to

$$\mathbb{E}\left[m_0^t(X,W) - m_0^0(X,W) \mid X = x\right] = \theta^t.$$

This completes the proof. □

F.3. Smooth Signal

Now we show that the moments $m(x;\theta,h)$ are $O(1)$-Lipschitz in x for any θ and h under standard boundedness conditions on the parameters.

First, we consider the residualized moment function is defined as

$$\psi(Z;\theta,h(X,W)) = \{Y - q(X,W) - \theta^\intercal\,(T - g(X,W))\}\,(T - g(X,W)),$$

Then for any θ and h given by functions g and q,

$$m(x;\theta,h) = \mathbb{E}\left[\{Y - q(x,W) - \theta^\intercal\,(T - g(x,W))\}\,(T - g(x,W)) \mid X = x\right]$$

Real-valued treatments In the real-valued treatment case, each coordinate j of g is given by a high-dimensional linear function: $g^j(x,W) = \langle W, \gamma^j\rangle$, where γ^j is a k-sparse vectors in $\mathbb{R}^{d_\nu}$ withℓ_1 norm bounded by a constant, and $q(x,W)$ can be written as a $\langle q', \phi_2(W)\rangle$ with q' is a k^2-sparse vector in $\mathbb{R}^{d_\nu^2}$ and $\phi_2(W)$ denotes the degree-2 polynomial feature vector of W.

$$m_j(x;\theta,h) = \mathbb{E}\left[\left\{Y - \langle q', \phi_2(W)\rangle - \theta_j\left(T - \langle\gamma^j, W\rangle\right)\right\}(T - g(x,W)) \mid X = x\right]$$

Note that as long as we restrict the space Θ and H to satisfy $\|\theta\| \leq O(1)$, $\|\gamma\|_1, \|q'\|_1 \leq 1$, we know each coordinate m_j is smooth in x.

Discrete treatments with residualized moments. In the discrete treatment case, each coordinate j of g is of the form $g^j(x,W) = \mathcal{L}(\langle W, \gamma^j\rangle)$, where $\mathcal{L}(t) = 1/(1 + e^{-t})$ is the logistic function. The estimate q consists of several components. First, consider function f of the form $f(x,W) = \langle W, \beta\rangle$ as an estimate for the outcome of the null treament. For each $t \in \{e_1, \ldots, e_p\}$, we also have an estimate $m^t(x,W)$ for the expected counter-factual outcome function $\mu^t(x) + f(x,W)$, which takes the form of $\langle b, W\rangle$. Then the estimate q is defined as:

$$q(x,W) = \sum_{t=1}^{p}(m^t(x,W) - f(x,W))g^t(x,W) + f(x,W).$$

With similar reasoning, as long as we restrict Θ and H to satisfy $|\theta| \leq O(1)$, $\|\gamma^j\|_1 \leq 1$ for all j, and $\|\beta\|_1, \|b\|_1 \leq 1$, we know each coordinate m_j is smooth in x.

Discrete treatments with doubly robust moments. Redcall that for each coordinate t, the moment function with input θ and nuisance parameters m, g is defined as

$$\mathbb{E}\left[m^t(x,W) + \frac{(Y - m^t(x,W))\,\mathbf{1}[T=t]}{g^t(x,W)} - m^0(x,W) - \frac{\big(Y - m^0(x,W)\big)\,\mathbf{1}[T=0]}{g^0(x,W)} - \theta^t \mid X = x\right] \tag{43}$$

where each $m^t(x,W)$ takes the form of $\langle b, W\rangle$ and each $g^t(x,W) = \mathcal{L}(\langle W, \gamma^t\rangle)$, with $\mathcal{L}$ denoting the logistic function. Then as long as we restrict the parameter space and H to satisfy $\|\gamma^t\|_1$ for all t, then we know that $|\langle \gamma^t, W\rangle| \leq O(1)$ and so $g^t(x,W) \geq \Omega(1)$. Furthermore, if we restrict the vector b to satisfy $\|b\|_1 \leq 1$, we know each coordinate m_j is smooth in x.

F.4. Curvature

Now we show that the jacobian $\nabla_\theta m(x; \theta_0(x), h_0)$ has minimum eigenvalues bounded away from 0.

Residualized moments. First, we consider the residualized moment function is defined as

$$\psi(Z; \theta, h(X,W)) = \{Y - q(X,W) - \theta^\intercal (T - g(X,W))\}\,(T - g(X,W)),$$

Then for any θ and h given by functions g and q,

$$m(x;\theta,h) = \mathbb{E}\left[\{Y - q(x,W) - \theta^\intercal (T - g(x,W))\}\,(T - g(x,W)) \mid X = x\right]$$

Let J be the expected Jacobian $\nabla_\theta m(x; \theta_0(x), h_0)$, and we can write

$$J_{jj'} = \mathbb{E}\left[(T_j - g_0^j(x,W))(T_{j'} - g_0^{j'}(x,W)) \mid X = x\right]$$

Then for any $v \in \mathbb{R}^p$ with unit ℓ_2 norm, we have

$$\begin{aligned} vJv^\intercal &= \mathbb{E}\left[\sum_j (T_j - g_0^j(x,W))^2 v_j^2 + 2\sum_{j,j'} (T_j - g_0^j(x,W))(T_{j'} - g_0^{j'}(x,W)) v_j v_{j'} \mid X = x\right] \\ &= \mathbb{E}\left[\left(\sum_j (T_j - g_0^j(x,W)) v_j\right)^2 \mid X = x\right] \\ &= \mathbb{E}\left[v^\intercal (\eta\eta^\intercal) v \mid X = x\right] \end{aligned}$$

Then as long as the conditional expected covariance matrix $\mathbb{E}\left[\eta\eta^\intercal \mid X = x\right]$ has minimum eigenvalue bounded away from zero, we will also have $\min_v vJv^\intercal$ bounded away from zero.

Discrete treatments with doubly robust moments. Redcall that for each coordinate t, the moment function with input θ and nuisance parameters m, g is defined as

$$\mathbb{E}\left[m^t(x,W) + \frac{(Y - m^t(x,W))\,\mathbf{1}[T=t]}{g^t(x,W)} - m^0(x,W) - \frac{\big(Y - m^0(x,W)\big)\,\mathbf{1}[T=0]}{g^0(x,W)} - \theta^t \mid X = x\right] \tag{44}$$

Then $\nabla_\theta m(x; \theta_0(x), h_0) = -I$, which implies the minimum eigevalue is 1.

F.5. Smoothness of scores

Residualized moments. First, we consider the residualized moment function with each coordinate defined as

$$\psi_j(Z; \theta, h(X,W)) = \{Y - q(X,W) - \theta^\intercal (T - g(X,W))\}\,(T_j - g_j(X,W)),$$

Observe that for both real-valued and discrete treatments, the scales of θ, $q(X,W)$, and $g(X,W)$ are bounded by $O(1)$. Thus, the smoothness condition immediately follows.

Doubly robust moments. For every treatment t,

$$\psi_t(Z;\theta,h(X,W)) = m^t(X,W) + \frac{(Y - m^t(X,W))\,\mathbf{1}[T=t]}{g^t(X,W)} - m^0(X,W) - \frac{\big(Y - m^0(X,W)\big)\,\mathbf{1}[T=0]}{g^0(X,W)} - \theta^t$$

Recall that each $m^t(x,W)$ takes the form of $\langle b, W\rangle$ and each $g^t(x,W) = \mathcal{L}(\langle W, \gamma^t\rangle)$, with $\mathcal{L}$ denoting the logistic function. Then as long as we restrict the parameter space and H to satisfy $\|\gamma^t\|_1$ for all t, then we know that $|\langle \gamma^t, W\rangle| \le O(1)$ and so $g^t(X,W) \ge \Omega(1)$. Furthermore, if we restrict the vector b to satisfy $\|b\|_1 \le 1$, we know each $m_j(X,W) \le O(1)$. Therefore, the smoothness condition also holds.

F.6. Accuracy for discrete treatments

For both score functions, we require that each discrete treatment (including the null treatment) is assigned with constant probability.

Corollary F.5 (Accuracy for residualized scores)**.** *Suppose that $\beta_0(X)$ and each coorindate $\beta_0(X), \gamma_0^j(X)$ and $\theta(X)$ are Lipschitz in X and have ℓ_1 norms bounded by $O(1)$ for any X. Assume that distribution of X admits a density that is bounded away from zero and infinity. For any feature X, the conditional covariance matrices satisfy $\mathbb{E}\left[\eta\eta^\intercal \mid X\right] \succeq \Omega(1)$, $\mathbb{E}\left[WW^\intercal \mid X\right] \succeq \Omega(1) I_{d_\nu}$. Then with probability $1-\delta$, ORF returns an estimator $\hat\theta$ such that*

$$\|\hat\theta - \theta_0\| \le O\left(n^{\frac{-1}{2+2\alpha d}}\sqrt{\log(nd_\nu/\delta)}\right)$$

as long as the sparsity $k \le O\left(n^{\frac{1}{4+4\alpha d}}\right)$ and the subsampling rate of ORF $s = \Theta(n^{\alpha d/(1+\alpha d)})$. Moreover, for any $b \in \mathbb{R}^p$ with $\|b\| \le 1$, there exists a sequence $\sigma_n = \Theta(\sqrt{\mathrm{polylog}(n) n^{-1/(1+\alpha d)}})$ such that

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

as long as the sparsity $k = o\left(n^{\frac{1}{4+4\alpha d}}\right)$ and the subsampling rate of ORF $s = \Theta(n^{\varepsilon + \alpha d/(1+\alpha d)})$ for any $\varepsilon > 0$.

Corollary F.6 (Accuracy for doubly robust scores)**.** *Suppose that $\beta_0(X)$ and each coorindate $u_0^j(X), \gamma_0^j(X)$ are Lipschitz in X and have ℓ_1 norms bounded by $O(1)$ for any X. Assume that distribution of X admits a density that is bounded away from zero and infinity. For any feature X, the conditional covariance matrices satisfy $\mathbb{E}\left[\eta\eta^\intercal \mid X\right] \succeq \Omega(1)$, $\mathbb{E}\left[WW^\intercal \mid X\right] \succeq \Omega(1) I_{d_\nu}$. Then with probability $1-\delta$, ORF returns an estimator $\hat\theta$ such that*

$$\|\hat\theta - \theta_0\| \le O\left(n^{\frac{-1}{2+2\alpha d}}\sqrt{\log(nd_\nu/\delta)}\right)$$

as long as the sparsity $k \le O\left(n^{\frac{1}{4+4\alpha d}}\right)$ and the subsampling rate of ORF $s = \Theta(n^{\alpha d/(1+\alpha d)})$. Moreover, for any $b \in \mathbb{R}^p$ with $\|b\| \le 1$, there exists a sequence $\sigma_n = \Theta(\sqrt{\mathrm{polylog}(n) n^{-1/(1+\alpha d)}})$ such that

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

as long as the sparsity $k = o\left(n^{\frac{1}{4+4\alpha d}}\right)$ and the subsampling rate of ORF $s = \Theta(n^{\varepsilon + \alpha d/(1+\alpha d)})$ for any $\varepsilon > 0$.

G. Orange Juice Experiment

Dominick's orange juice dataset (provided by the University of Chicago Booth School of Business) contains 28,947 entries of store-level, weekly prices and sales of different brands of orange juice. The dataset also contains 15 continuous and categorical variables that encode store-level customer information such as the mean age, income, education level, etc, as well as brand information. The goal is to learn the elasticity of orange juice as a function of income (or education, etc) in the presence of high-dimensional controls.

In the experiment depicted in Figure 1, we trained the ORF using 500 trees, a minimum leaf size of 50, subsample ratio of 0.02, with Lasso models for both residualization and kernel estimation. We evaluated the resulting algorithm on 50 $\log(Income)$ points between 10.4 and 10.9. We then followed-up with 100 experiments on bootstrap samples of the original dataset to build bootstrap confidence intervals. The emerging trend in the elasticity as a function of income follows our intuition: higher income levels correspond to a more inelastic demand.

H. All Experimental Results

We present all experimental results for the parameter choices described in Section 7. We vary the support size $k \in \{1, 5, 10, 15, 20, 25, 30\}$, the dimension $d \in \{1, 2\}$ of the feature vector x and the treatment response function $\theta \in \{\text{piecewise linear, piecewise constant and piecewise polynomial}\}$. We measure the bias, variance and root mean square error (RMSE) as evaluation metrics for the different estimators we considered in Section 7. In addition, we add another version of the GRF (GRF-xW) where we run the GRF R package directly on the observations, using features and controls (x, W) jointly as the covariates. For the parameter space we consider, the ORF-CV and the ORF algorithms outperform the other estimators on all regimes.

H.1. Experimental results for one-dimensional, piecewise linear θ_0

Consider a piecewise linear function: $\theta_0(x) = (x + 2)\mathbb{I}_{x \le 0.3} + (6x + 0.5)\mathbb{I}_{x > 0.3 \text{ and } x \le 0.6} + (-3x + 5.9)\mathbb{I}_{x > 0.6}$.

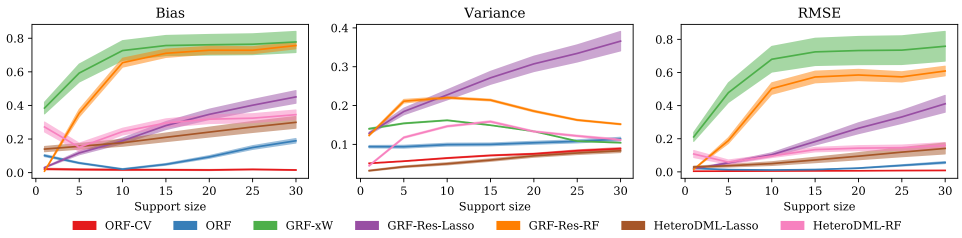

Figure 6: Bias, variance and RMSE as a function of support size for $n = 5000$, $p = 500$, $d = 1$ and θ_0. The solid lines represent the mean of the metrics across test points, averaged over the 100 experiments, and the filled regions depict the standard deviation, scaled down by a factor of 3 for clarity.

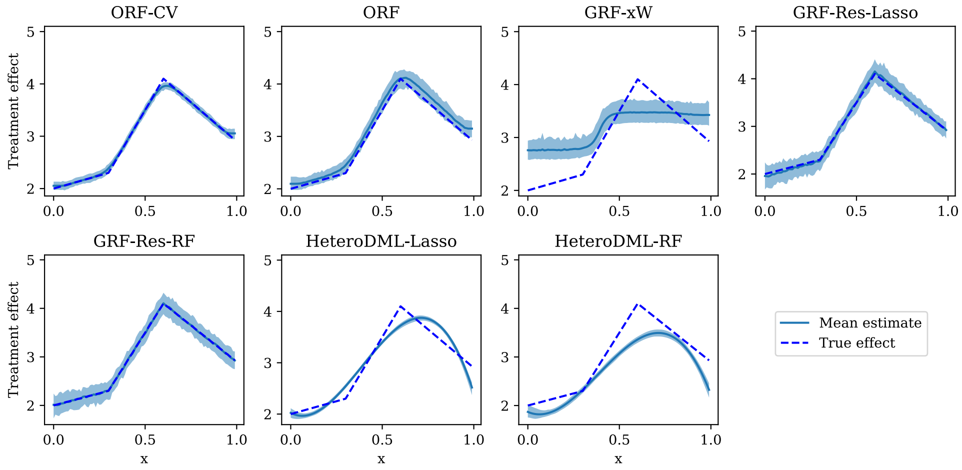

Figure 7: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k = 1}$, and a piecewise linear treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

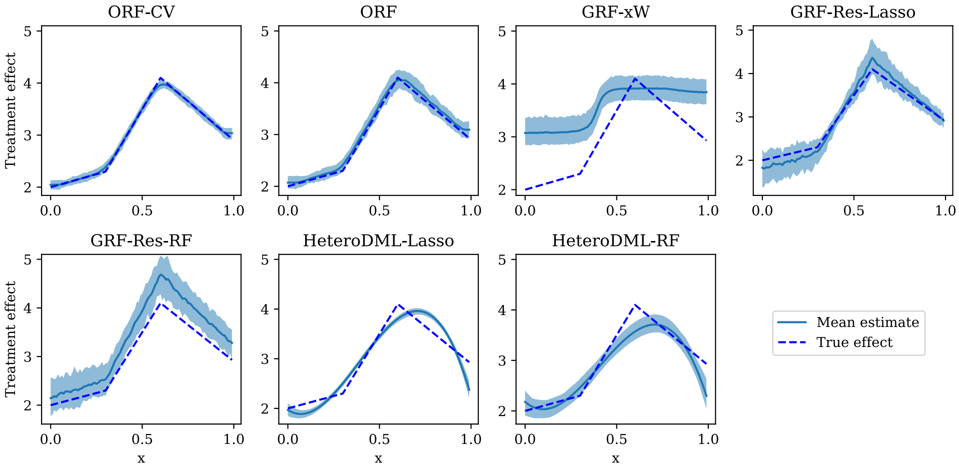

Figure 8: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{5}$, and a piecewise linear treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

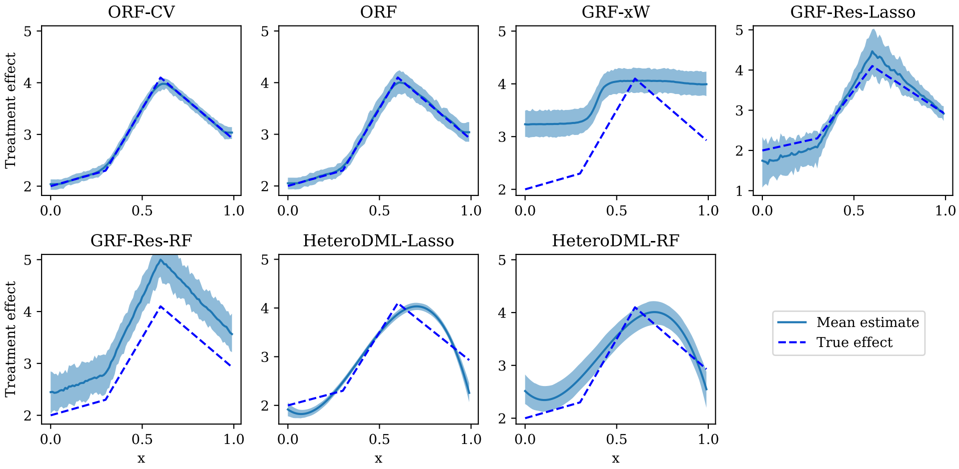

Figure 9: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{10}$, and a piecewise linear treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

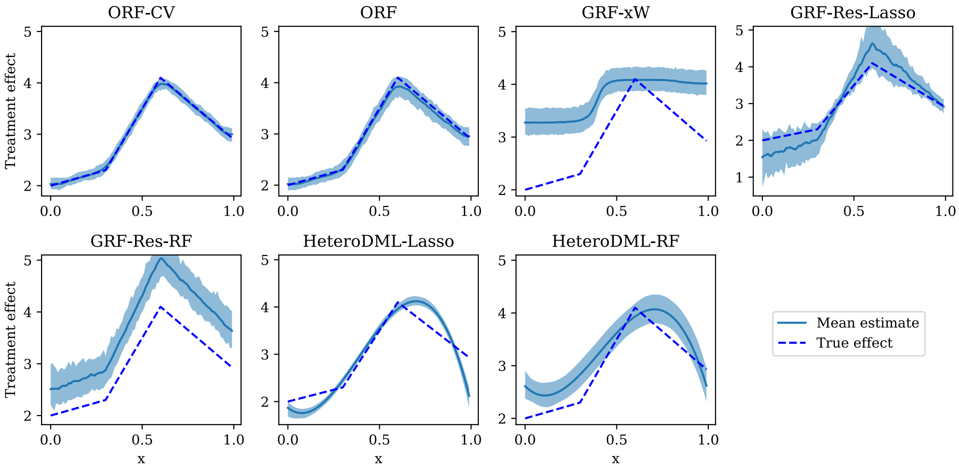

Figure 10: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{15}$, and a piecewise linear treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

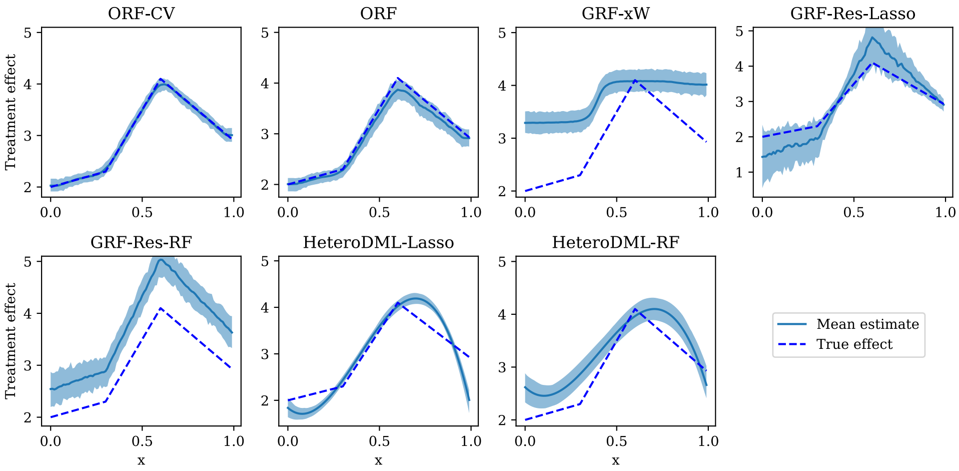

Figure 11: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{20}$, and a piecewise linear treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

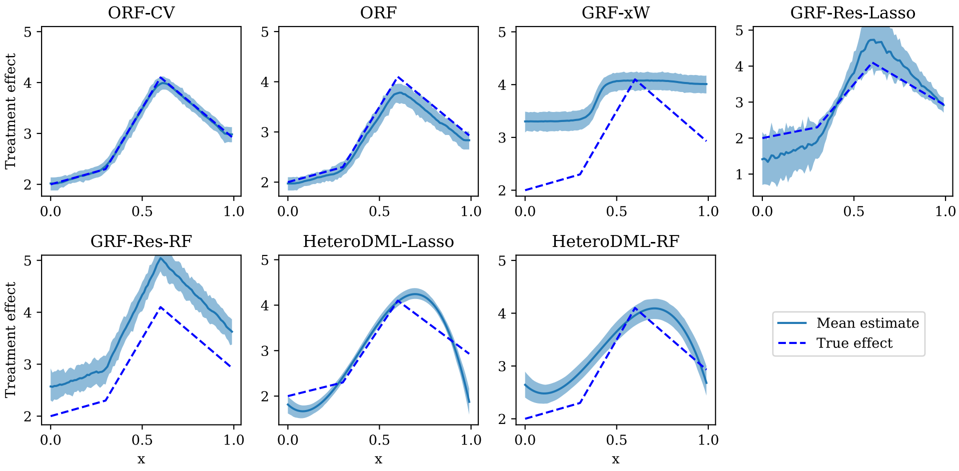

Figure 12: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{25}$, and a piecewise linear treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

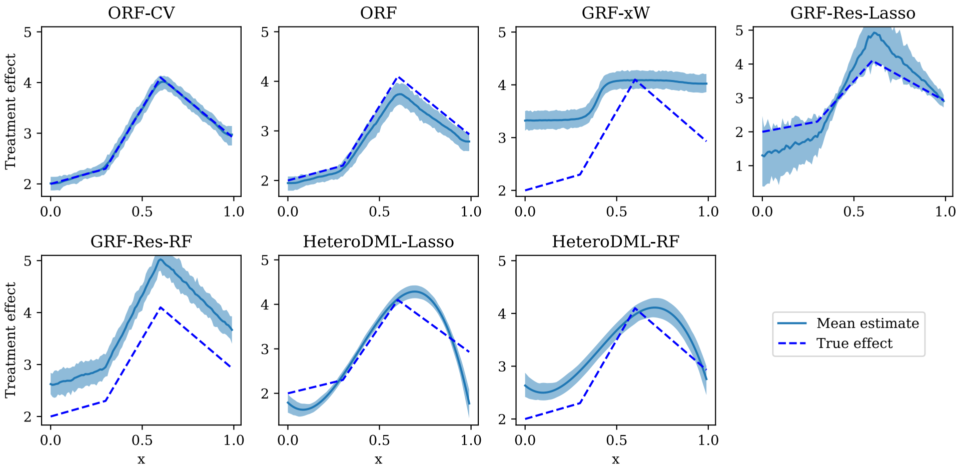

Figure 13: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{30}$, and a piecewise linear treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

H.2. Experimental results for one-dimensional, piecewise constant θ_0

We introduce the results for a piecewise constant function given by:

$$\theta_0(x) = \mathbb{I}_{x \leq 0.2} + 5\mathbb{I}_{x > 0.2 \text{ and } x \leq 0.6} + 3\mathbb{I}_{x > 0.6}$$

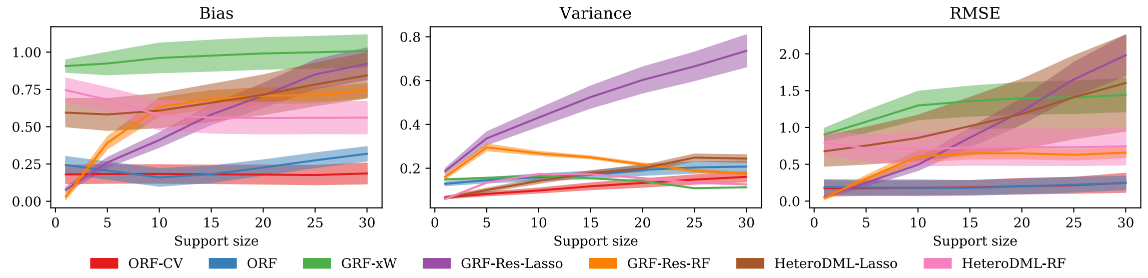

Figure 14: Bias, variance and RMSE as a function of support size for $n = 5000$, $p = 500$, $d = 1$ and a piecewise constant treatment function. The solid lines represent the mean of the metrics across test points, averaged over the Monte Carlo experiments, and the filled regions depict the standard deviation, scaled down by a factor of 3 for clarity.

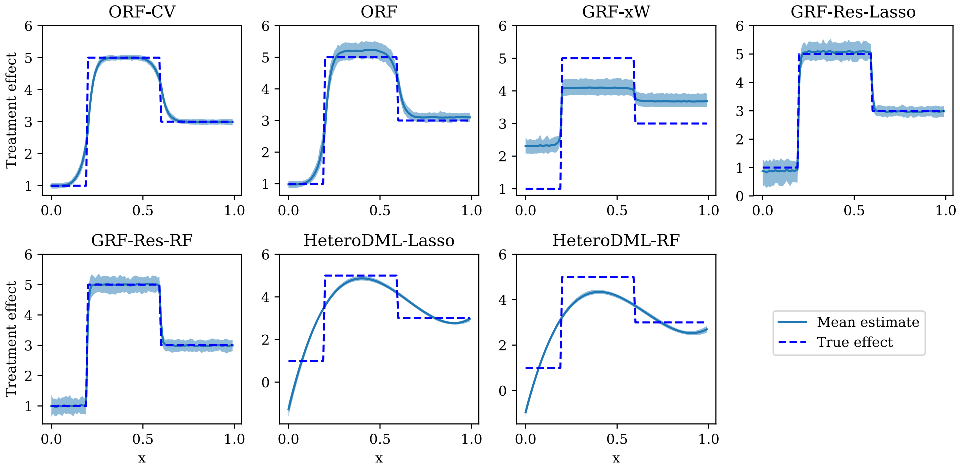

Figure 15: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k = 1}$, and a piecewise constant treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

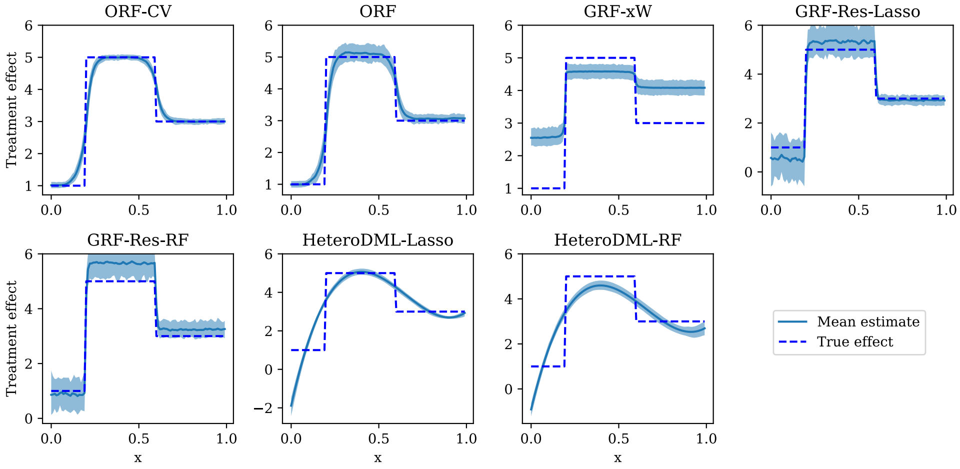

Figure 16: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k = 5}$, and a piecewise constant treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

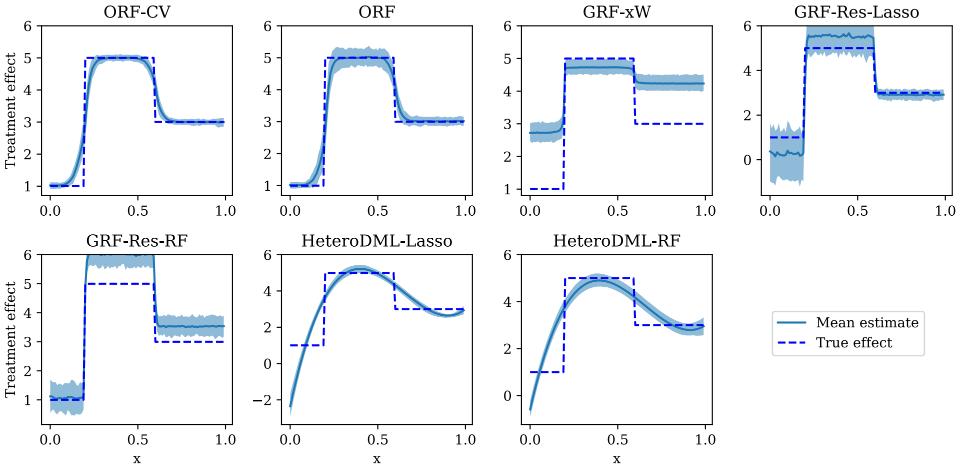

Figure 17: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{10}$, and a piecewise constant treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

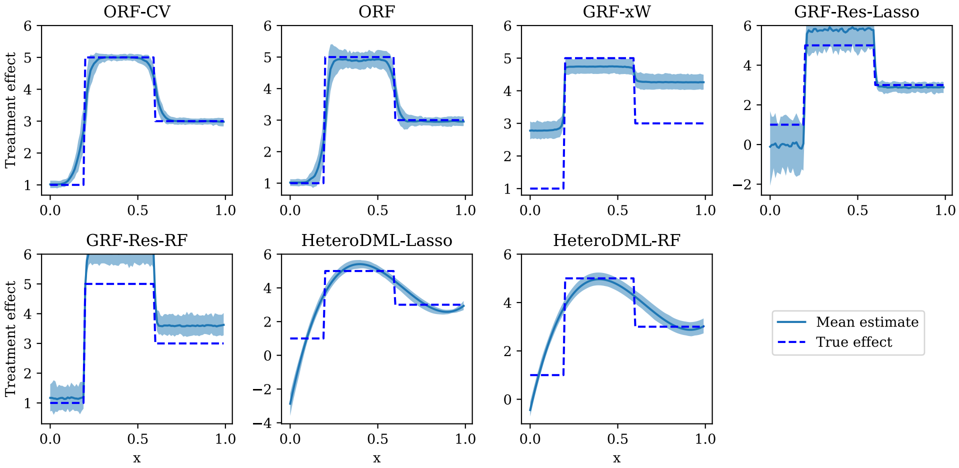

Figure 18: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{15}$, and a piecewise constant treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

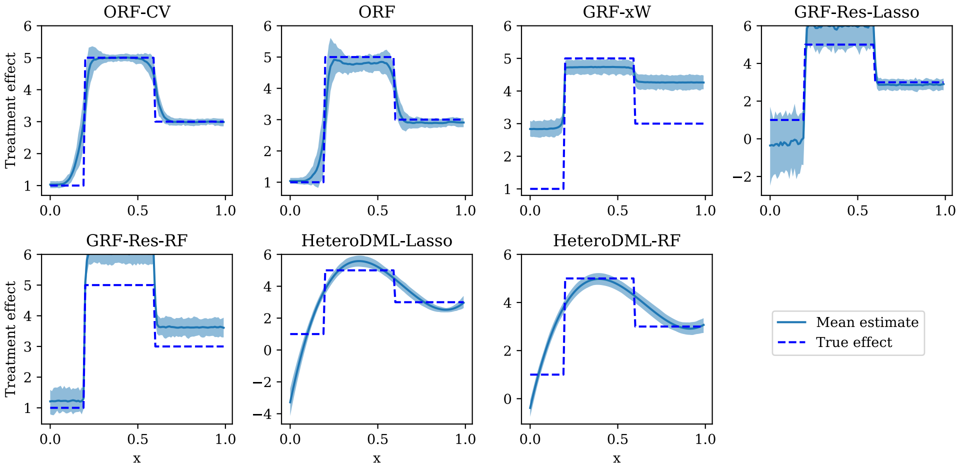

Figure 19: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{20}$, and a piecewise constant treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

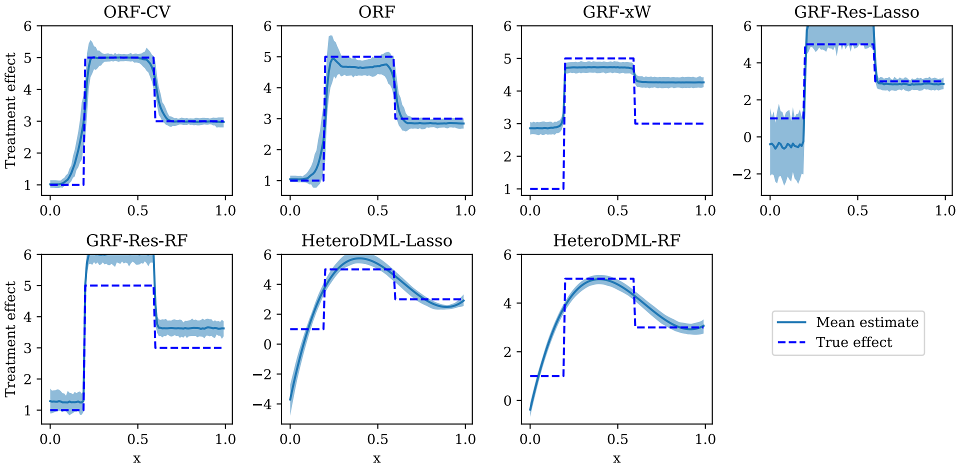

Figure 20: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{25}$, and a piecewise constant treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

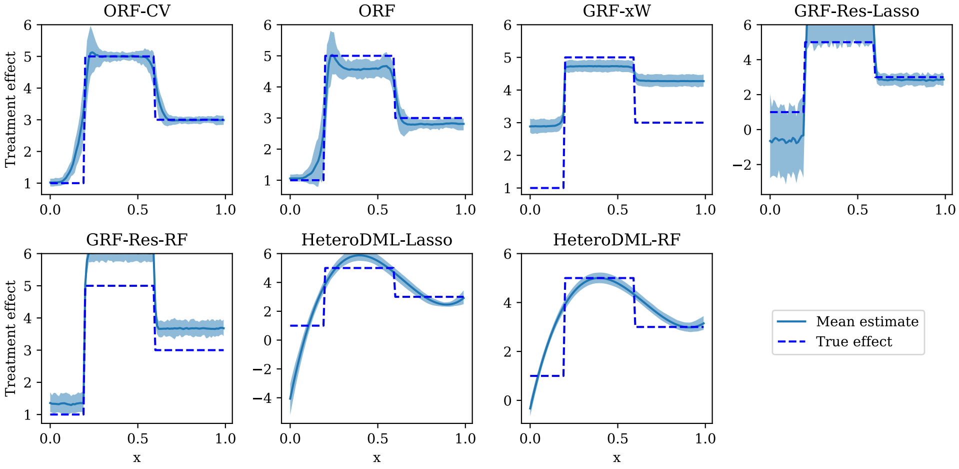

Figure 21: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{30}$, and a piecewise constant treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

H.3. Experimental results for one-dimensional, piecewise polynomial θ_0

We present the results for a piecewise polynomial function given by:

$$\theta_0(x) = 3x^2\mathbb{I}_{x\leq 0.2} + (3x^2+1)\mathbb{I}_{x>0.2 \text{ and } x\leq 0.6} + (6x+2)\mathbb{I}_{x>0.6}$$

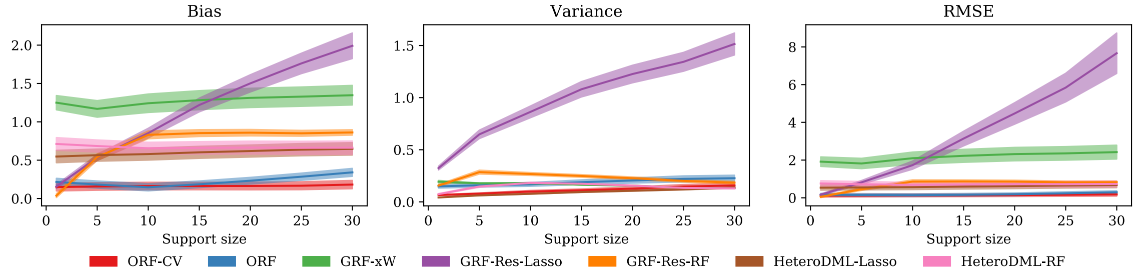

Figure 22: Bias, variance and RMSE as a function of support size for $n = 5000$, $p = 500$, $d = 1$ and a piecewise polynomial treatment function. The solid lines represent the mean of the metrics across test points, averaged over the Monte Carlo experiments, and the filled regions depict the standard deviation, scaled down by a factor of 3 for clarity.

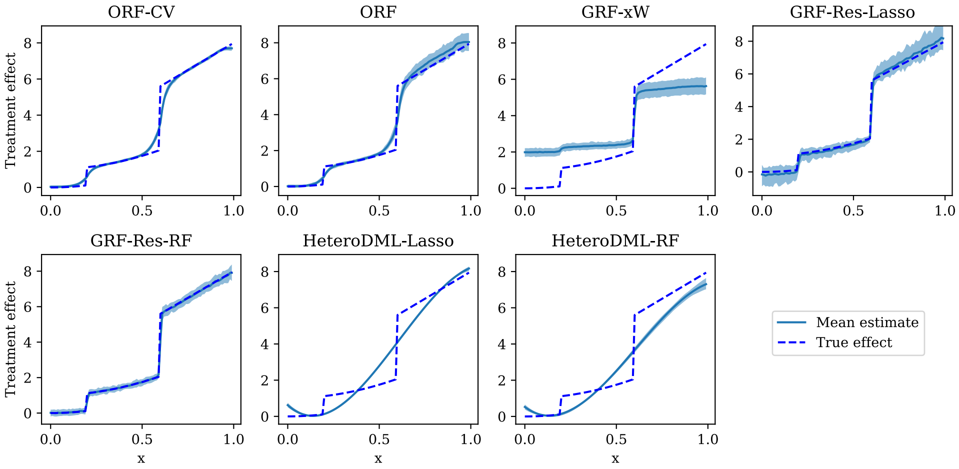

Figure 23: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{1}$, and a piecewise polynomial treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

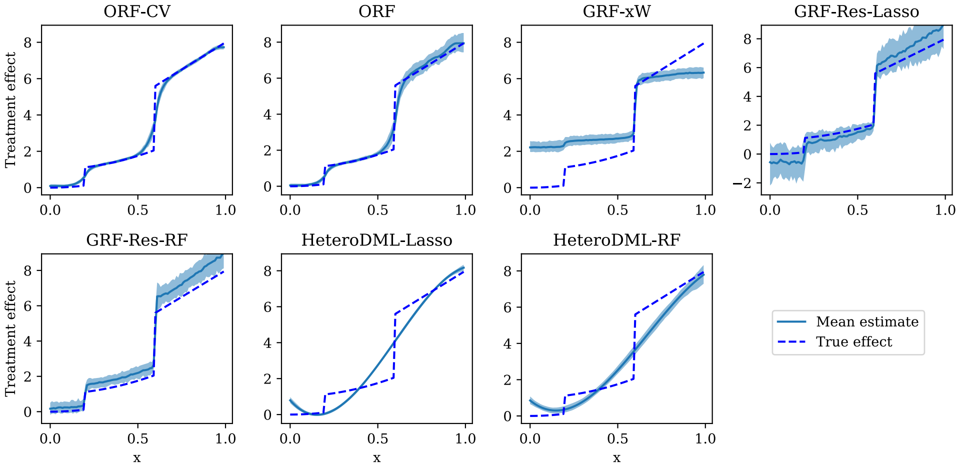

Figure 24: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{5}$, and a piecewise polynomial treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

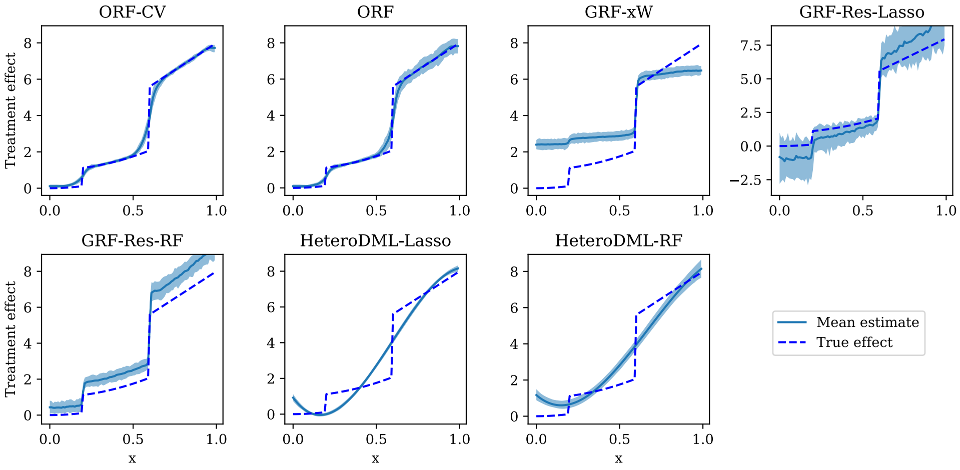

Figure 25: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{10}$, and a piecewise polynomial treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

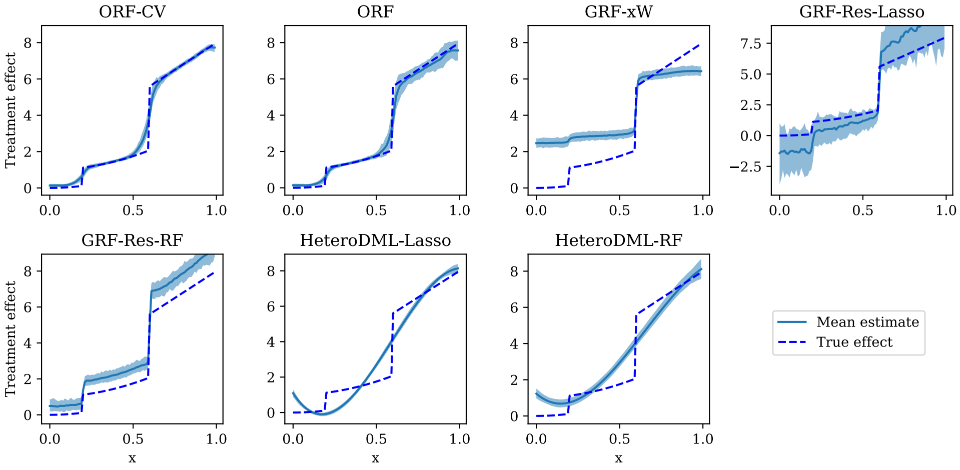

Figure 26: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{15}$, and a piecewise polynomial treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

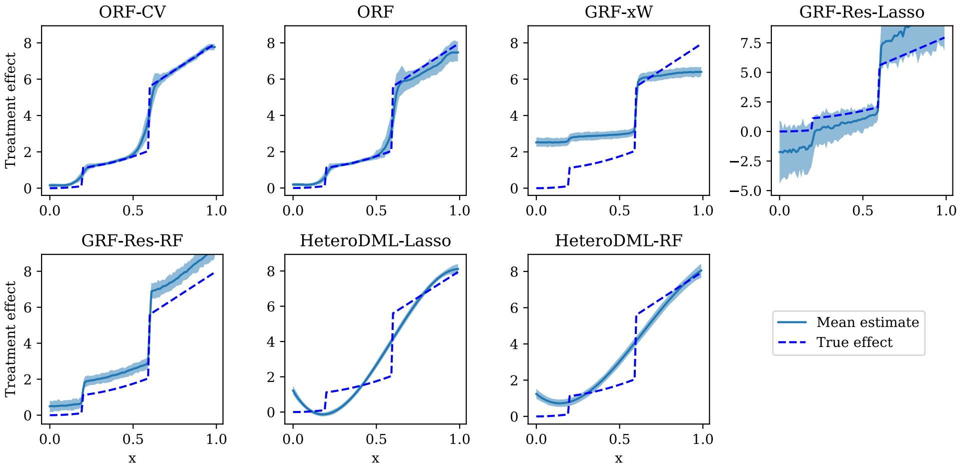

Figure 27: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{20}$, and a piecewise polynomial treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

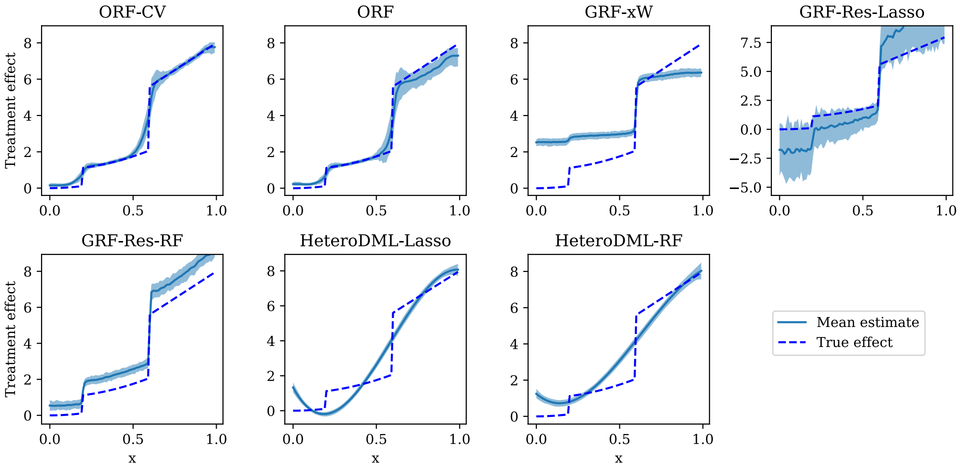

Figure 28: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{25}$, and a piecewise polynomial treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

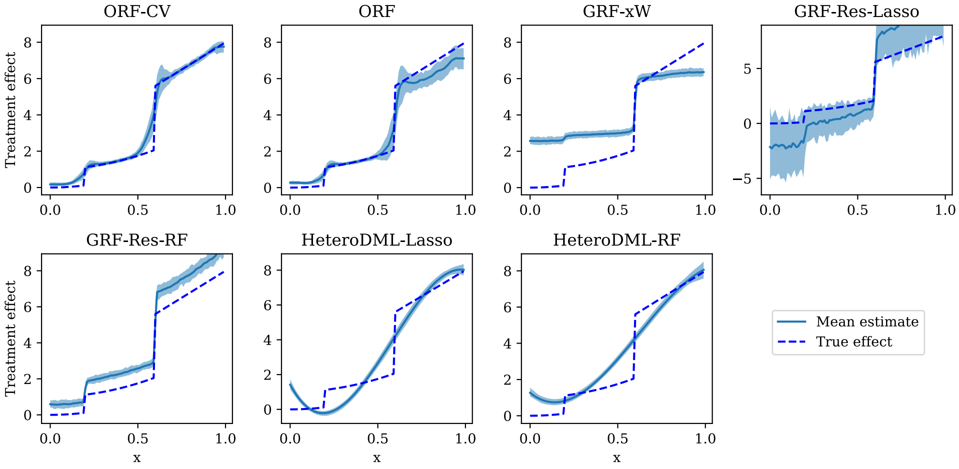

Figure 29: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k} = \mathbf{30}$, and a piecewise polynomial treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

H.4. Experimental results for larger control support

We present experimental results for a piecewise linear treatment response θ_0, with $n = 5000$ samples and large support $k \in \{50, 75, 100, 150, 200\}$. Figures 30-35 illustrate that the behavior of the ORF-CV algorithm, with parameters set in accordance our theoretical results, is consistent up until fairly large support sizes. Our method performs well with respect to the chosen evaluation metrics and outperform other estimators for larger support sizes.

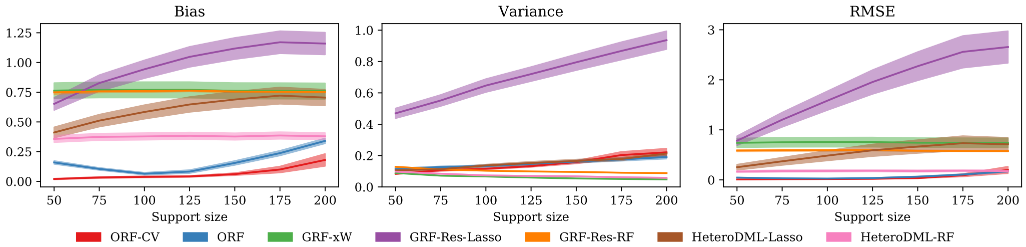

Figure 30: Bias, variance and RMSE as a function of support size. The solid lines represent the mean of the metrics across test points, averaged over the Monte Carlo experiments, and the filled regions depict the standard deviation, scaled down by a factor of 3 for clarity.

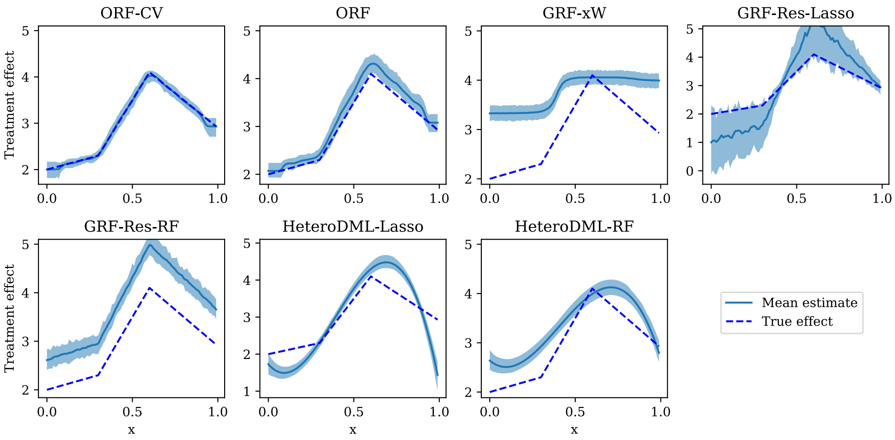

Figure 31: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k = 50}$, and a piecewise linear treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

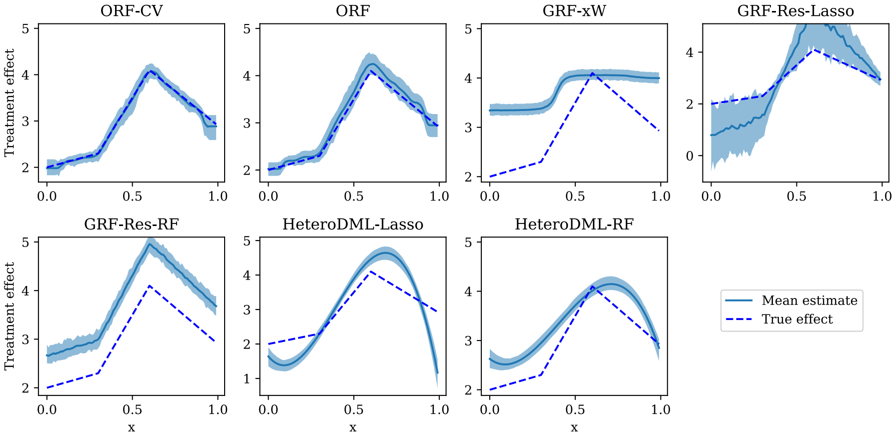

Figure 32: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k = 75}$, and a piecewise linear treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

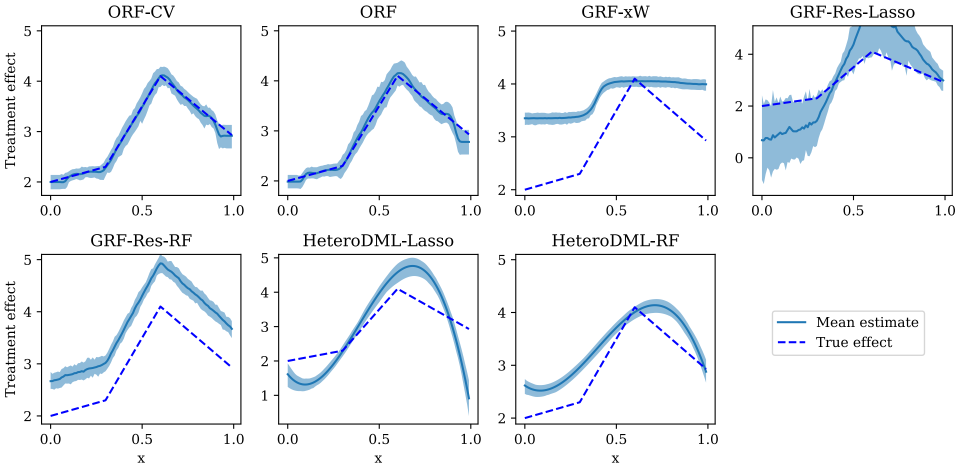

Figure 33: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k = 100}$, and a piecewise linear treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

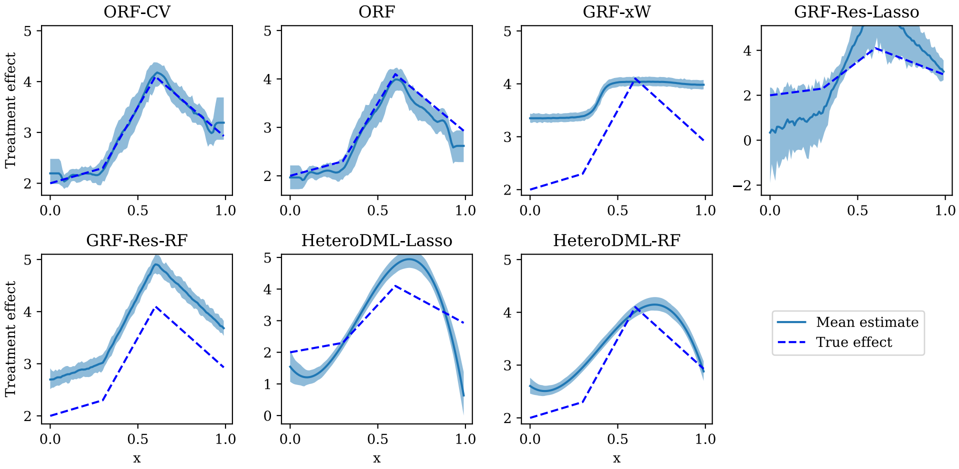

Figure 34: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k = 150}$, and a piecewise linear treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

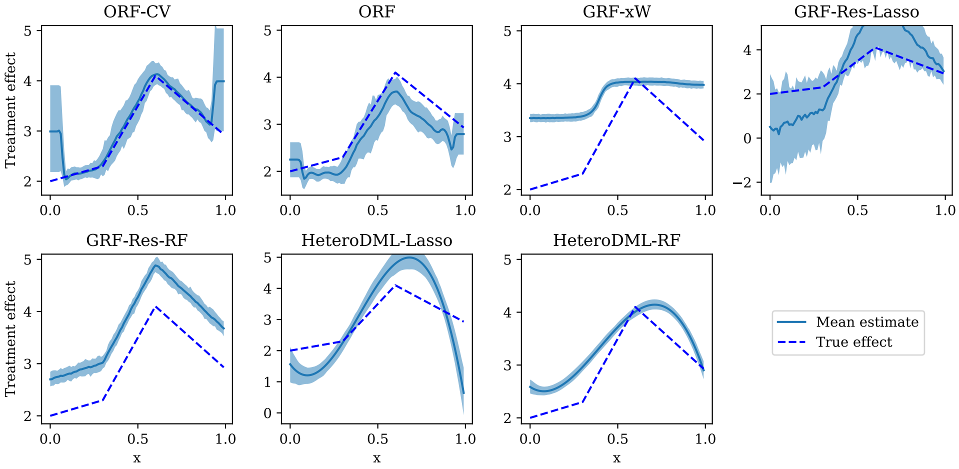

Figure 35: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 1$, $\mathbf{k = 200}$, and a piecewise linear treatment response. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

H.5. Experimental results for two-dimensional heterogeneity

We introduce experimental results for a two-dimensional x and corresponding θ_0 given by:

$$\theta_0(x_1, x_2) = \theta_{\text{piecewise linear}}(x_1)\mathbb{I}_{x_2=0} + \theta_{\text{piecewise constant}}(x_1)\mathbb{I}_{x_2=1}$$

where $x_1 \sim U[0, 1]$ and $x_2 \sim Bern(0.5)$. In Figures 36-44, we examine the overall behavior of the ORF-CV and ORF estimators, as well as the behavior across the slices $x_2 = 0$ and $x_2 = 1$. We compare the performance of the ORF-CV and ORF estimators with alternative methods for $n = 5000$ and $k \in \{1, 5, 10, 15, 20, 25, 30\}$. We conclude that the ORF-CV algorithm yields a better performance for all support sizes and evaluation metrics.

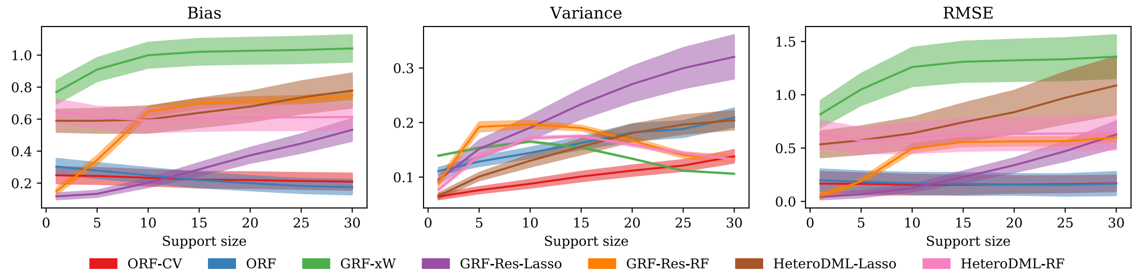

Figure 36: Overall bias, variance and RMSE as a function of support size for $n = 5000$, $p = 500$, $d = 2$. The solid lines represent the mean of the metrics across test points, averaged over the Monte Carlo experiments, and the filled regions depict the standard deviation, scaled down by a factor of 3 for clarity.

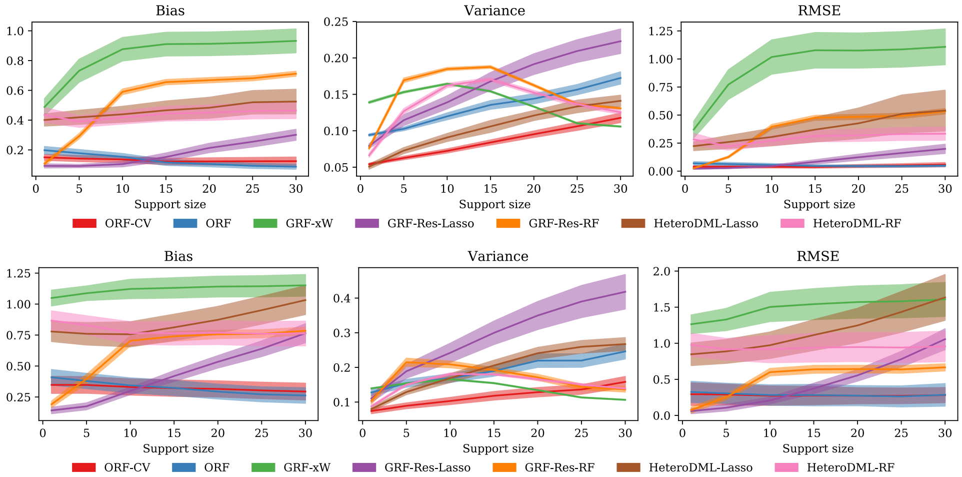

Figure 37: Bias, variance and RMSE as a function of support for $n = 5000$, $p = 500$, $d = 2$ and slices $\mathbf{x_2} = \mathbf{0}$ and $\mathbf{x_2} = \mathbf{1}$, respectively. The solid lines represent the mean of the metrics across test points, averaged over the Monte Carlo experiments, and the filled regions depict the standard deviation, scaled down by a factor of 3 for clarity.

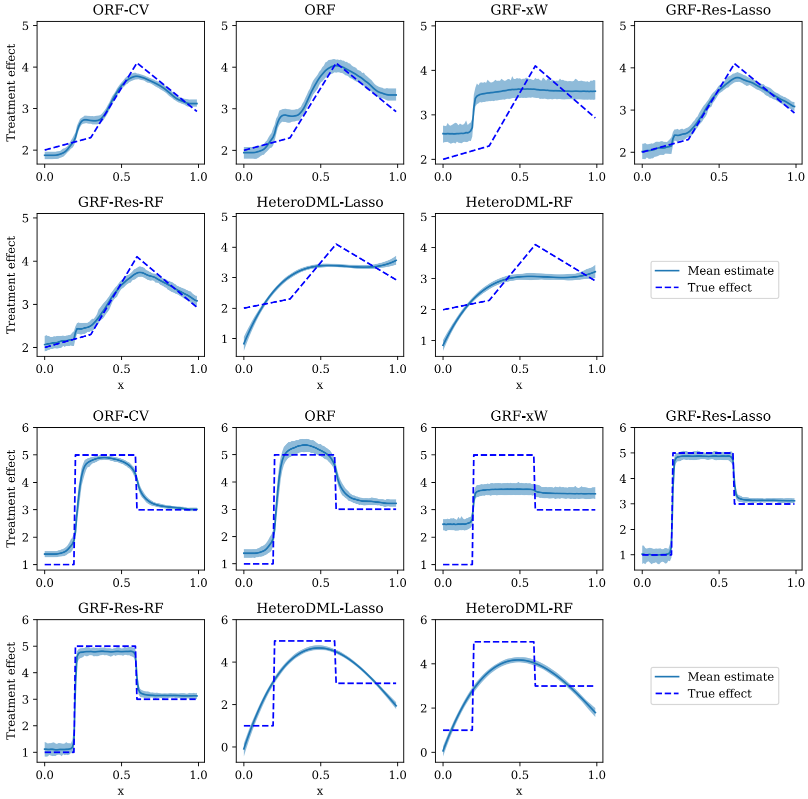

Figure 38: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 2$, $\mathbf{k} = \mathbf{1}$, and slices $\mathbf{x_2} = \mathbf{0}$ and $\mathbf{x_2} = \mathbf{1}$, respectively. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

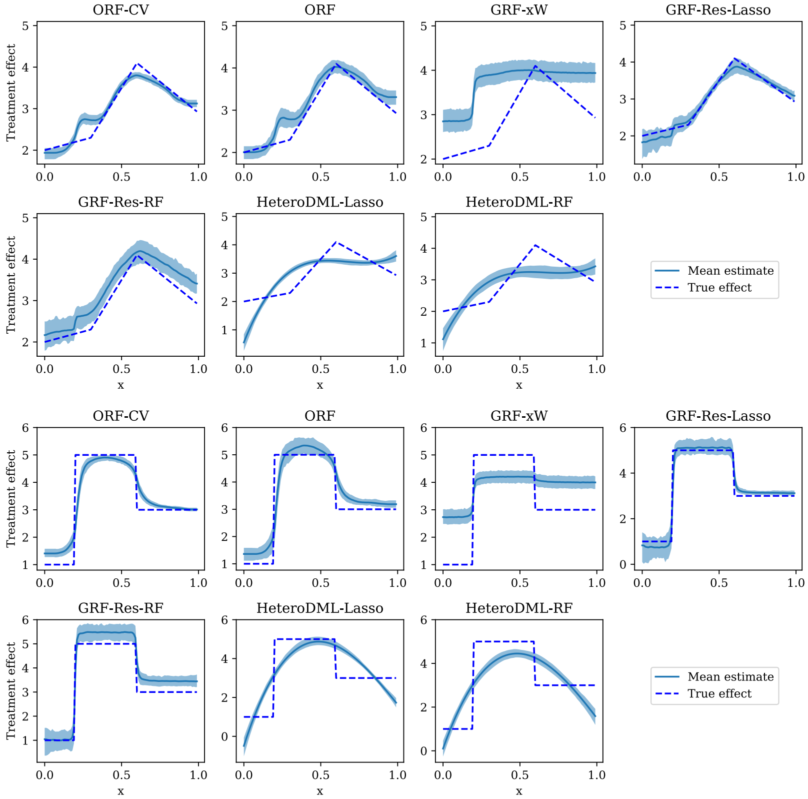

Figure 39: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 2$, $\mathbf{k} = \mathbf{5}$, and slices $\mathbf{x_2} = \mathbf{0}$ and $\mathbf{x_2} = \mathbf{1}$, respectively. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

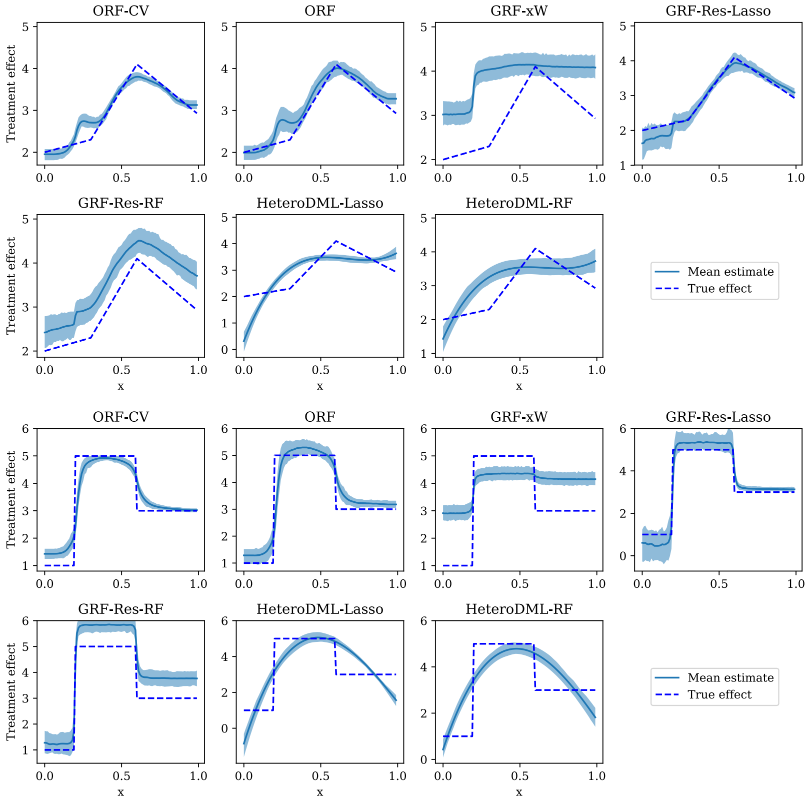

Figure 40: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 2$, $\mathbf{k} = \mathbf{10}$, and slices $\mathbf{x_2} = \mathbf{0}$ and $\mathbf{x_2} = \mathbf{1}$, respectively. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

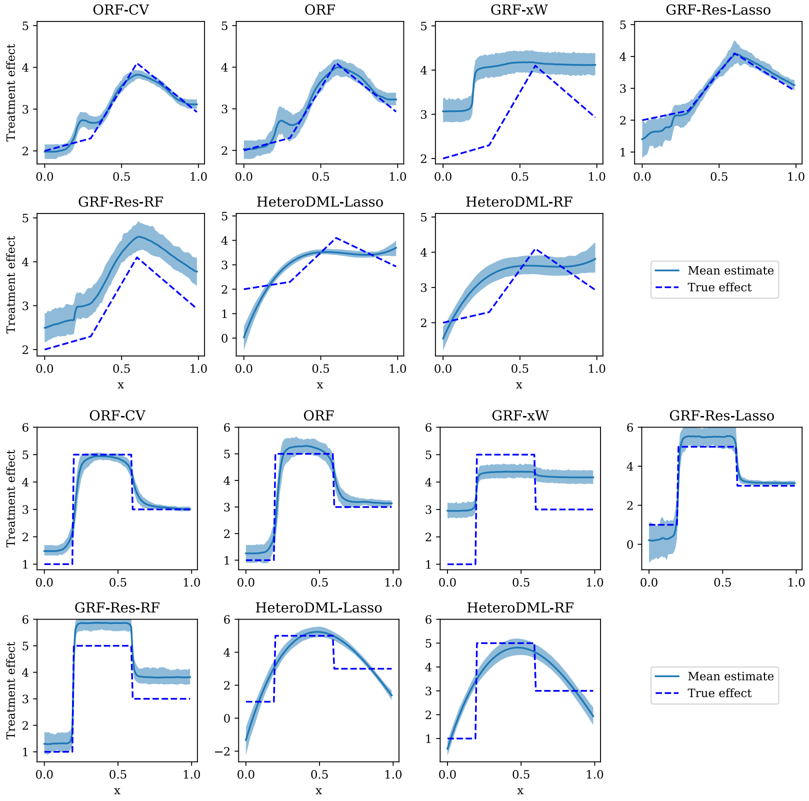

Figure 41: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 2$, $\mathbf{k} = \mathbf{15}$, and slices $\mathbf{x_2} = \mathbf{0}$ and $\mathbf{x_2} = \mathbf{1}$, respectively. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

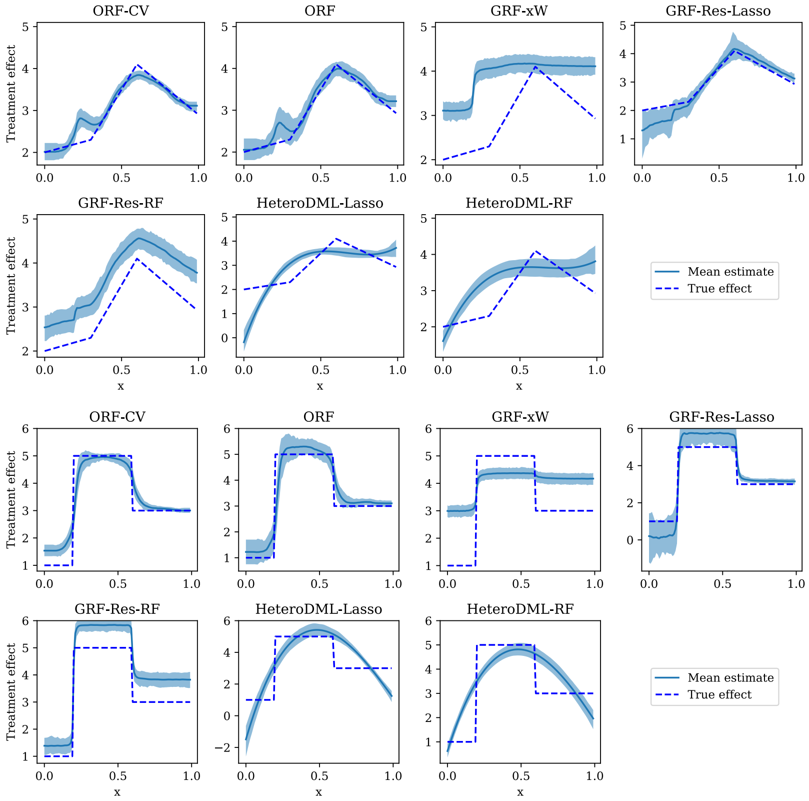

Figure 42: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 2$, $\mathbf{k = 20}$, and slices $\mathbf{x_2 = 0}$ and $\mathbf{x_2 = 1}$, respectively. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

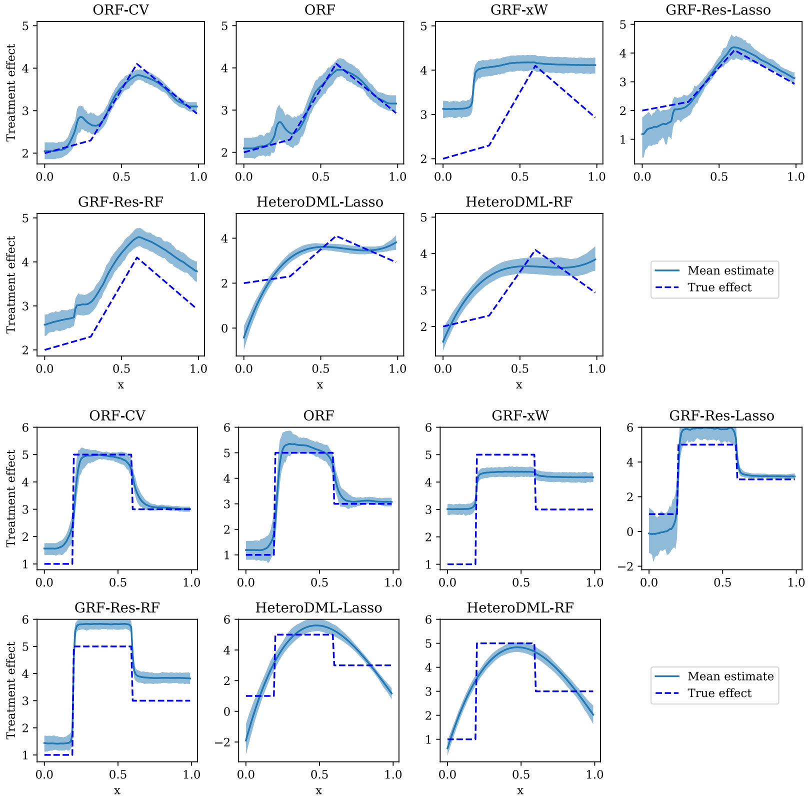

Figure 43: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 2$, $\mathbf{k = 25}$, and slices $\mathbf{x_2 = 0}$ and $\mathbf{x_2 = 1}$, respectively. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.

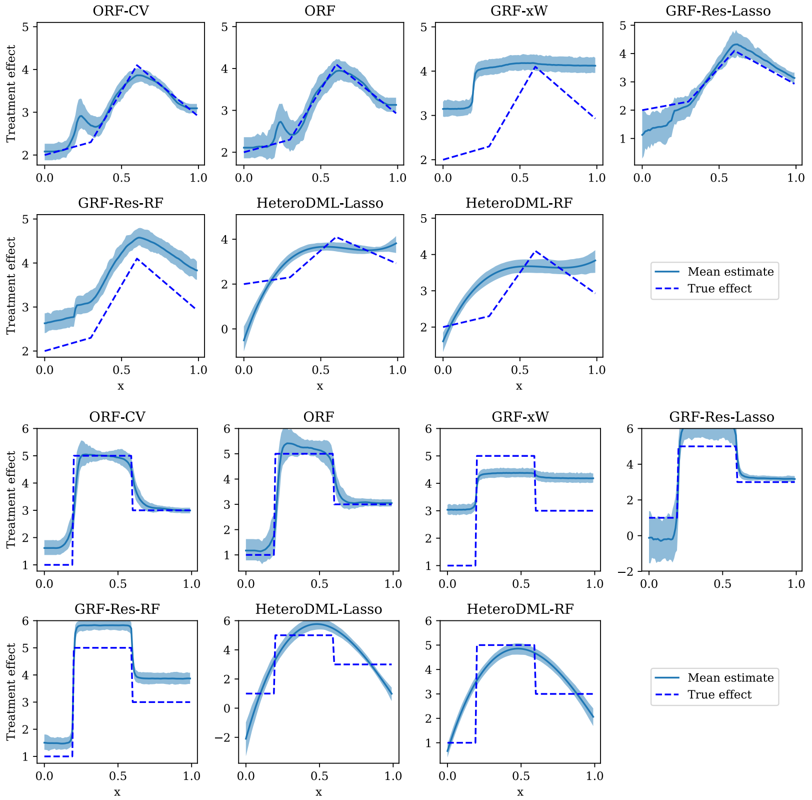

Figure 44: Treatment effect estimations for 100 Monte Carlo experiments with parameters $n = 5000$, $p = 500$, $d = 2$, $\mathbf{k = 30}$, and slices $\mathbf{x_2 = 0}$ and $\mathbf{x_2 = 1}$, respectively. The shaded regions depict the mean and the 5%-95% interval of the 100 experiments.