A Additional Simulations

In Table 4, we expand on the simulation results given in Table 2, and present results for data generated according to (28) all while varying s and n. As expected, we observe a bias-variance trade-off: the causal forest is more affected by bias when s is small relative to n, and by variance when s is larger relative to n. Reassuringly, we observe that our confidence intervals obtain close-to-nominal coverage when the mean-squared error matches the average variance estimate $\hat{\sigma}^2(X)$ generated by the infinitesimal jackknife, corroborating the hypothesis that failures in coverage mostly arise when the causal forest is bias- instead of variance-dominated.

Finally, all our experiments relied on settings with strong, low-dimensional structure that forests could pick up on to improve over k-NN matching. This intuition is formally supported by, e.g., the theory developed by Biau [2012]. Here, we consider how forests perform when the signal is spread out over a larger number of features, and so forests have less upside over other methods. We find that—as expected—they do not improve much over baselines. Specifically, we generate data with a treatment effect function

$$\tau(x) = \frac{4}{q} \sum_{j=1}^{q} \left(\frac{1}{1 + e^{-12(x_j - 0.5)}} - \frac{1}{2} \right), \tag{30}$$

where we vary both the number of signal dimensions q and ambient dimensions d. As seen in Table 5, forests vastly improve over k-NN in terms of mean-squared error when q is much smaller than d, but that this advantage decreases when d and q are comparable; and actually do worse than k-NN when we have a dense signal with $d = q = 6$. When the signal is dense, all surveyed methods have bad coverage except for 10-NN which, as always, simply has very wide intervals.

B Is Honesty Necessary for Consistency?

Our honesty assumption is the largest divergence between our framework and main-stream applications of random forests. Following Breiman [2001a], almost all practical implementations of random forests are not honest. Moreover, there has been a stream of recent work providing theoretical guarantees for

adaptive random forests: Scornet et al. [2015] establish risk consistency under assumptions on the data-generating function, i.e., they show that the test-set error of forests converges asymptotically to the Bayes risk, Mentch and Hooker [2016] provide results about uncertainty quantification, and Wager and Walther [2015] find that adaptive trees with growing leaves are in general consistent under fairly weak conditions and provide bounds on the decay rate of their bias.

However, if we want pointwise centered asymptotic Gaussianity results, then honesty appears to be necessary. Consider the following simple example, where there is no treatment heterogeneity—and in fact X is independent of W and Y. We are in a randomized controlled trial, where $X \sim \text{Uniform}([0, 1]^p)$ with $p = 10$ and $W \sim \text{Bernoulli}(0.5)$. The distribution of Y is $Y_i = 2W_iA_i + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}\left(0, 0.1^2\right)$ and $A_i \sim \text{Bernoulli}(0.05)$. Thus, the treatment effect is $\tau(x) = 0.1$ for all $x \in [0, 1]^p$. Our goal is to estimate the treatment effect $\tau(x_0)$ at $x_0 = (0, 0, \ldots, 0)$.

Results from running both honest and adaptive forests are shown in Figure 3. We see that honest forests are unbiased regardless of n, and their mean-squared error decreases with sample size, as expected. Adaptive forests, in contrast, perform remarkably badly. They have bias that far exceeds the intrinsic sampling variation; and, furthermore, this bias *increases* with n. What is happening here is that CART trees aggressively seek to separate outliers ("$Y_i \approx 1$") from the rest of the data ("$Y_i \approx 0$") and, in doing so, end up over-representing outliers in the corners of the feature space. As n increases, it appears that adaptive forests have more opportunities to push outliers into corners of features space and so the bias worsens. This phenomenon is not restricted to causal forests; an earlier technical report [Wager, 2014] observed the same phenomenon in the context of plain regression forests. Honest trees do not have this problem, as we do not know where the $\mathcal{I}$-sample outliers will be when placing splits using only the $\mathcal{J}$-sample. Thus, it appears that adaptive CART forests are pointwise biased in corners of x-space.

Finally, we note that this bias phenomenon does not contradict existing consistency results in the literature. Wager and Walther [2015] prove that this bias phenomenon discussed above can be averted if we use a minimum leaf-size k that grows with n (in contrast, Figure 3 uses $k = 1$). However, their bounds on the bias decays slower than the sampling variance of random forests, and so their results cannot be used to get centered confidence intervals.

Meanwhile, Scornet et al. [2015] prove that forests are risk-consistent at an average test point, and, in fact, the test set error of adaptive forests does decay in the setting of Figure 3 as the sample size n grows (although honest forests still maintain a lower test set error). The reason test set error can go to zero despite the bias phenomenon in Figure 3 is that, when n gets large, almost all test points will be far enough from corners that they will not be affected by the phenomenon from Figure 3.

B.1 Adaptive versus Honest Predictive Error

The discussion above implies that the theorems proved in this paper are not valid for adaptive forests. That being said, it still remains interesting to ask whether our use of honest forest hurts us in terms of mean-squared error at a random test point, as in the formalism considered by, e.g., Scornet et al. [2015]. In this setting, Denil et al. [2014] showed that honesty can hurt the performance of regression forests on some classic datasets from the UCI repository; however, in a causal inference setting, we might be concerned that the risk of overfitting with adaptive forests is higher since our signals of interest are often quite weak.

We compare the performance of honest and adaptive forests in the setting of Table 2, with $d = 8$. Here, if we simply run adaptive forests out-of-the-box with the usual minimum leaf size parameter $k = 1$, they do extremely badly; in fact, they do worse than 50 nearest neighbors. However, if we are willing to increase the minimum leaf size, their performance improves.

Figure 4 depicts the root-mean-squared error for both adaptive and honest forests over a wide range of choices for the minimum leaf size parameter k. We see that, at their best, both methods do comparably. However, honest forests attain good performance over a wide range of choices for k, including our default choice $k = 1$, whereas adaptive forests are extremely sensitive to choosing a good value of k. We also note that the optimum $k = 64$ for adaptive forests is quite far from standard choices advocated in practice; such as $k = 5$ recommended by Breiman [2001a] for regression forests, $k = 7$ in the `cforest` function used by Strobl et al. [2007], or $k = 10$ recommended by Meinshausen [2006] for quantile regression. Thus, it appears that accurately tuning adaptive forests in this setting may present a challenge and, overall, a practitioner may prefer honest forests even based on their mean-squared error characteristics alone.

C Proofs

Notation. Throughout the appendix we use the following notation to describe asymptotic scalings: $f(s) \sim g(s)$ means that $\lim_{s\to\infty} f(s)/g(s) = 1$, $f(s) \gtrsim g(s)$ means that $\liminf_{s\to\infty} f(s)/g(s) \geq 1$ and $f(s) \lesssim g(s)$ is analogous, $f(s) = \mathcal{O}(g(s))$ means that $f(s) \lesssim C\, g(s)$ for some $C > 0$, $f(s) = \Omega(g(s))$ means that $f(s) \gtrsim c\, g(s)$ for some $c > 0$, and finally $f(s) = o(g(s))$ means that $\limsup_{s\to\infty} f(s)/g(s) = 0$.

Proof of Theorem 1. Given the statements of Theorem 8 and Theorem 9, it only remains to show that (15) holds with $\mathbb{E}\left[\hat{\mu}_n(x)\right]$ replaced with $\mu(x)$. To do so, it suffices to show that $\left|\mathbb{E}\left[\hat{\mu}_n\left(x\right)\right] - \mu(x)\right| / \sigma_n\left(x\right) \to 0$; the rest follows from Slutsky's lemma. Now, recall that by Theorem 3,

$$\left|\mathbb{E}\left[\hat{\mu}_n\left(x\right)\right] - \mu(x)\right| = \mathcal{O}\left(n^{-\frac{\beta}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\pi^{-1}d\,\log\left(\alpha^{-1}\right)}}\right).$$

Meanwhile, from Theorem 5 and the proof of Theorem 8, we see that

$$\sigma_n^2(x) \gtrsim C_{f,\,d}\,\frac{s}{n}\,\frac{\mathrm{Var}\left[T\right]}{\log(s)^d}.$$

By honesty of T, $\mathrm{Var}\left[T\right] \gtrsim \mathrm{Var}\left[Y \,\middle|\, X = x\right] / \left|\{i : X_i \in L(x)\}\right| \geq \mathrm{Var}\left[Y \,\middle|\, X = x\right] / (2k)$, and so

$$\sigma_n^2(x) \gtrsim \frac{C_{f,\,d}}{2k}\,\frac{s}{n}\,\frac{\mathrm{Var}\left[Y \,\middle|\, X = x\right]}{\log(s)^d} = \Omega\left(n^{\beta-1-\varepsilon}\right)$$

for any $\varepsilon > 0$. It follows that

$$\frac{\left|\mathbb{E}\left[\hat{\mu}_n\left(x\right)\right] - \mu(x)\right|}{\sigma_n\left(x\right)} = \mathcal{O}\left(n^{\frac{1}{2}\left(1+\varepsilon-\beta\left(1+\frac{\log\left((1-\alpha)^{-1}\right)}{\pi^{-1}d\,\log\left(\alpha^{-1}\right)}\right)\right)}\right).$$

The right-hand-side bound converges to 0 for some small enough $\varepsilon > 0$ provided that

$$\beta > \left(1 + \frac{\log\left((1-\alpha)^{-1}\right)}{\pi^{-1}d\,\log\left(\alpha^{-1}\right)}\right)^{-1} = 1 - \left(1 + \frac{d}{\pi}\,\frac{\log\left(\alpha^{-1}\right)}{\log\left((1-\alpha)^{-1}\right)}\right)^{-1} = \beta_{\min}.$$

□

C.1 Bounding the Bias of Regression Trees

Proof of Lemma 2. Let $c(x)$ be the number of splits leading to the leaf $L(x)$, and let $c_j(x)$ be the number of these splits along the j-th coordinate. By regularity, we know that $s\alpha^{c(x)} \leq 2k-1$, and so $c(x) \geq \log(s/(2k-1))/\log(\alpha^{-1})$. Thus, because the tree is a random split tree, $c_j(x)$ we have the following stochastic lower bound for $c_j(x)$:

$$c_j(x) \overset{d}{\geq} \text{Binom}\left(\frac{\log\left(s/(2k-1)\right)}{\log\left(\alpha^{-1}\right)}; \frac{\pi}{d}\right). \tag{31}$$

By Chernoff's inequality, it follows that

$$\begin{aligned}\mathbb{P}\left[c_j\left(x\right) \leq \frac{\pi}{d}\frac{\log\left(s/(2k-1)\right)}{\log\left(\alpha^{-1}\right)}\left(1-\eta\right)\right] &\leq \exp\left[-\frac{\eta^2}{2}\frac{\log\left(s/(2k-1)\right)}{\pi^{-1}d\log\left(\alpha^{-1}\right)}\right] \\ &= \left(\frac{s}{2k-1}\right)^{-\frac{\eta^2}{2}\frac{1}{\pi^{-1}d\log(\alpha^{-1})}}.\end{aligned} \tag{32}$$

Meanwhile, again by regularity, we might expect that $\text{diam}_j\left(L\left(x\right)\right)$ should less than $(1-\alpha)^{c_j(x)}$, at least for large n. This condition would hold directly if the regularity condition from Definition 4 were framed in terms of Lebesgue measure instead of the number of training examples in the leaf; our task is to show that it still holds approximately in our current setup.

Using the methods developed in Wager and Walther [2015], and in particular their Lemma 12, we can verify that, with high probability and simultaneously for all but the last $\mathcal{O}\left(\log\log n\right)$ parent nodes above $L(x)$, the number of training examples inside the node divided by n is within a factor $1+o(1)$ of the Lebesgue measure of the node. From this, we conclude that, for large enough s, with probability greater than $1-1/s$

$$\text{diam}_j\left(L\left(x\right)\right) \leq \left(1-\alpha+o(1)\right)^{(1+o(1))c_j(x)},$$

or, more prosaically, that

$$\text{diam}_j\left(L\left(x\right)\right) \leq \left(1-\alpha\right)^{0.991\, c_j(x)}.$$

Combining this results with the above Chernoff bound yields the desired inequality. Here, replacing 0.991 with 0.99 in the bound lets us ignore the $1/s$ asymptotic failure probability of the concentration result used above.

Finally, we note that with double-sample trees, all the "s" terms in the above argument need to be replaced by "$s/2$"; this additional factor 2, however, does not affect the final result. □

Proof of Theorem 3. We begin with two observations. First, by honesty,

$$\mathbb{E}\left[T\left(x;\, Z\right)\right] - \mathbb{E}\left[Y \,\middle|\, X = x\right] = \mathbb{E}\left[\mathbb{E}\left[Y \,\middle|\, X \in L\left(x\right)\right] - \mathbb{E}\left[Y \,\middle|\, X = x\right]\right].$$

Second, by Lipschitz continuity of the conditional mean function,

$$\left|\mathbb{E}\left[Y \,\middle|\, X \in L\left(x\right)\right] - \mathbb{E}\left[Y \,\middle|\, X = x\right]\right| \leq C\,\text{diam}\left(L\left(x\right)\right),$$

where C is the Lipschitz constant. Thus, in order to bound the bias under both Lipschitz and honesty assumptions, it suffices to bound the average diameter of the leaf $L(x)$.

To do so, we start by plugging in $\eta = \sqrt{\log((1-\alpha)^{-1}}$ in the bound from Lemma 2. Thanks our assumption that $\alpha \leq 0.2$, we see that $\eta \leq 0.48$ and so $0.99 \cdot (1-\eta) \geq 0.51$; thus, a union bound gives us that, for large enough s,

$$\mathbb{P}\left[\text{diam}\left(L(x)\right) \geq \sqrt{d}\left(\frac{s}{2k-1}\right)^{-0.51\,\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\,\frac{\pi}{d}}\right] \leq d\left(\frac{s}{2k-1}\right)^{-\frac{1}{2}\,\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\,\frac{\pi}{d}}.$$

The Lipschitz assumption lets us bound the bias on the event that that $\text{diam}(L(x))$ satisfies this bound.

Thus, for large s, we find that

$$\left|\mathbb{E}\left[T\left(x;\, Z\right)\right] - \mathbb{E}\left[Y \,\middle|\, X = x\right]\right| \lesssim d \left(\frac{s}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}} \times \left(\sup_{x\in[0,\,1]^d}\left\{\mathbb{E}\left[Y \,\middle|\, X = x\right]\right\} - \inf_{x\in[0,\,1]^d}\left\{\mathbb{E}\left[Y \,\middle|\, X = x\right]\right\}\right),$$

where $\sup_x \mathbb{E}\left[Y \,\middle|\, X = x\right] - \inf_x \mathbb{E}\left[Y \,\middle|\, X = x\right] = \mathcal{O}\left(1\right)$ because the conditional mean function is Lipschitz continuous. Finally, since a forest is just an average of trees, the above result also holds for $\hat{\mu}(x)$. □

C.2 Bounding the Incrementality of Regression Trees

Proof of Lemma 4. First, we focus on the case where f is constant, i.e., the features X_i have a uniform distribution over $[0,\, 1]^d$. To begin, recall that the S_i denote selection weights

$$T\left(x;\, Z\right) = \sum_{i=1}^{s} S_i Y_i \quad \text{where} \quad S_i = \begin{cases} \left|\{i : X_i \in L(x;\, Z)\}\right|^{-1} & \text{if } X_i \in L(x;\, Z), \\ 0 & \text{else,} \end{cases}$$

where $L(x;\, Z)$ denotes the leaf containing x. We also define the quantities

$$P_i = 1\left(\{X_i \text{ is a } k\text{-PNN of } x\}\right).$$

Because T is a k-PNN predictor, $P_i = 0$ implies that $S_i = 0$, and, moreover, $\left|\{i : X_i \in L(x;\, Z)\}\right| \geq k$; thus, we can verify that

$$\mathbb{E}\left[S_1 \,\middle|\, Z_1\right] \leq \frac{1}{k}\,\mathbb{E}\left[P_1 \,\middle|\, Z_1\right].$$

The bulk of the proof involves showing that

$$\mathbb{P}\left[\mathbb{E}\left[P_1 \,\middle|\, Z_1\right] \geq \frac{1}{s^2}\right] \lesssim k\,\frac{2^{d+1}\log\left(s\right)^d}{(d-1)!}\,\frac{1}{s}; \tag{33}$$

by the above argument, this immediately implies that

$$\mathbb{P}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right] \geq \frac{1}{k\,s^2}\right] \lesssim k\,\frac{2^{d+1}\log(s)^d}{(d-1)!}\,\frac{1}{s}. \tag{34}$$

Now, by exchangeability of the indices i, we know that

$$\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right]\right] = \mathbb{E}\left[S_1\right] = \frac{1}{s}\mathbb{E}\left[\sum_{i=1}^{s} S_i\right] = \frac{1}{s},$$

moreover, we can verify that

$$\mathbb{P}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right] \geq \frac{1}{ks^2}\right]\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right] \,\middle|\, \mathbb{E}\left[S_1 \,\middle|\, Z_1\right] \geq \frac{1}{ks^2}\right] \sim \frac{1}{s}.$$

By Jensen's inequality, we then see that

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right]^2\right] &\geq \mathbb{P}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right] \geq \frac{1}{ks^2}\right]\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right] \,\middle|\, \mathbb{E}\left[S_1 \,\middle|\, Z_1\right] \geq \frac{1}{ks^2}\right]^2 \\ &\sim \frac{s^{-2}}{\mathbb{P}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right] \geq \frac{1}{ks^2}\right]}\end{aligned}$$

which, paired with (34), implies that

$$\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right]^2\right] \gtrsim \frac{(d-1)!}{2^{d+1}\log(s)^d}\,\frac{1}{k\,s}.$$

This is equivalent to (20) because $\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right]\right]^2 = 1/s^2$ is negligibly small.

We now return to establishing (33). Recall that X_1, ..., X_s are independently and uniformly distributed over $[0,\,1]^d$, and that we are trying to find points that are k-PNNs of a prediction point x. For now, suppose that $x = 0$. We know that X_1 is a k-PNN of 0 if and only if there are at most $2k-2$ other points X_i such that $X_{ij} \leq X_{1j}$ for all $j = 1$, ..., d (because the X have a continuous density, there will

almost surely be no ties). Thus,

$$\begin{aligned}\mathbb{E}_{x=0}\left[P_1 \,\middle|\, Z_1\right] &= \mathbb{P}\left[\text{Binomial}\left(s-1;\prod_{j=1}^{d} X_{1j}\right) \leq 2k-2\right] \\ &\leq \binom{s-1}{2k-2}\left(1-\prod_{j=1}^{d} X_{1j}\right)^{s-2k+1} \\ &\leq s^{2k-2}\left(1-\prod_{j=1}^{d} X_{1j}\right)^{s-2k+1},\end{aligned} \tag{35}$$

where the second inequality can be understood as a union bound over all sets of $s-2k+1$ Bernoulli realizations that could be simultaneously 0. We can check that $X_{1j} \stackrel{d}{=} e^{-E_j}$ where E_j is a standard exponential random variable, and so

$$\mathbb{E}_{x=0}\left[P_1 \,\middle|\, Z_1\right] \stackrel{d}{\leq} s^{2k-2}\left(1-\exp\left[-\sum_{j=1}^{d} E_j\right]\right)^{s-2k+1},$$

where $A \stackrel{d}{\leq} B$ means that $B-A \geq 0$ almost surely. Thus,

$$\begin{aligned}&\mathbb{P}_{x=0}\left[\mathbb{E}\left[P_1 \,\middle|\, Z_1\right] \geq \frac{1}{s^2}\right] \\ &\quad\leq \mathbb{P}\left[s^{(2k-2)}\left(1-\exp\left[-\sum_{j=1}^{d} E_j\right]\right)^{s-2k+1} \geq \frac{1}{s^2}\right] \\ &\quad= \mathbb{P}\left[\exp\left[-\sum_{j=1}^{d} E_j\right] \leq 1-\left(\frac{1}{s^{2k}}\right)^{\frac{1}{s-2k+1}}\right] \\ &\quad= \mathbb{P}\left[\sum_{j=1}^{d} E_j \geq -\log\left(1-\exp\left[-2k\,\frac{\log(s)}{s-2k+1}\right]\right)\right].\end{aligned}$$

Notice that this quantity goes to zero as s gets large. The sum of d standard exponential random variables has a gamma distribution with shape d and scale 1, and

$$\mathbb{P}\left[\sum_{j=1}^{d} E_j \geq c\right] = \frac{\Gamma(d,\, c)}{(d-1)!},$$

where Γ is the upper incomplete gamma function. It is well known that

$$\lim_{c\to\infty} \frac{\Gamma(d,\, c)}{c^{d-1}\, e^{-c}} = 1,$$

and so

$$\begin{aligned}
&\mathbb{P}_{x=0}\left[\mathbb{E}\left[P_1 \,\middle|\, Z_1\right] \geq \frac{1}{s^2}\right] \\
&\qquad \lesssim \frac{\left(-\log\left(1-\exp\left[-2k\,\frac{\log(s)}{s-2k+1}\right]\right)\right)^{d-1}\left(1-\exp\left[-2k\,\frac{\log(s)}{s-2k+1}\right]\right)}{(d-1)!}.
\end{aligned}$$

We can check that

$$1-\exp\left[-2k\,\frac{\log(s)}{s-2k+1}\right] \sim 2k\,\frac{\log(s)}{s},$$

letting us simplify the above expression to

$$\mathbb{P}_{x=0}\left[\mathbb{E}\left[P_1 \,\middle|\, Z_1\right] \geq \frac{1}{s^2}\right] \lesssim \frac{2k}{(d-1)!}\,\frac{\log(s)^d}{s}. \tag{36}$$

We thus have obtained a tight expression for our quantity of interested for a prediction point at $x = 0$.

In the case $x \neq 0$, the ambient space around x can be divided into 2^d quadrants. In order to check whether X_i is a PNN, we only need to consider other points in the same quadrant, as no point in a different quadrant can prevent X_i from being a PNN. Now, index the quadrants by $l = 1, ..., 2^d$, and let v_l be the volume of the l-th quadrant. By applying (36) on the l-th quadrant alone, we see that the probability of $\mathbb{E}\left[P_1 \,\middle|\, Z_1\right] \geq \frac{1}{s^2}$ given that X_1 is in the l-the quadrant is asymptotically bounded on the order of

$$\frac{k+1}{(d-1)!}\,\frac{\log(s)^d}{v_k\, s}.$$

Summing over all quadrants, we find that

$$\begin{aligned}\mathbb{P}_{x=0}\left[\mathbb{E}\left[P_1 \,\middle|\, Z_1\right] \geq \frac{1}{s^2}\right] &\lesssim \sum_{\{k:v_k>0\}} v_k \, \frac{2k}{(d-1)!} \, \frac{\log(s)^d}{v_k \, s} \\ &= |\{k : v_k > 0\}| \, \frac{2k}{(d-1)!} \, \frac{\log(s)^d}{s} \\ &\leq k \, \frac{2^{d+1}}{(d-1)!} \, \frac{\log(s)^d}{s},\end{aligned}$$

thus establishing (33). Finally, to generalize to bounded densities f, we note that if $f(x) \leq C$ for all $x \in [0,\,1]^d$, then

$$\mathbb{E}_{x=0}\left[P_1 \,\middle|\, Z_1\right] \leq \mathbb{P}\left[\text{Binomial}\left(s-1;\, C\prod_{j=1}^{d} X_{1j}\right) \leq 2k-2\right],$$

and the previous argument goes though. □

Proof of Theorem 5. Our main task is to show that if T is a regular tree and $\text{Var}\left[Y \,\middle|\, X = x\right] > 0$, then

$$\text{Var}\left[\mathbb{E}\left[T\left(x;\, Z\right) \,\middle|\, Z_1\right]\right] \gtrsim \text{Var}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right]\right] \text{Var}\left[Y \,\middle|\, X = x\right]. \tag{37}$$

Given this result, Lemma 4 then implies that

$$\text{Var}\left[\mathbb{E}\left[T\left(x;\, Z\right) \,\middle|\, Z_1\right]\right] \gtrsim \frac{1}{k} \, \frac{\nu(s)}{s} \, \text{Var}\left[Y \,\middle|\, X = x\right].$$

Moreover, by Theorem 3, we know that

$$\begin{aligned}&\mathbb{E}\left[Y_i \,\middle|\, X_i \in L\left(x;\, Z\right)\right] \to_p \mathbb{E}\left[Y \,\middle|\, X = x\right], \text{ and} \\ &\mathbb{E}\left[Y_i^2 \,\middle|\, X_i \in L\left(x;\, Z\right)\right] \to_p \mathbb{E}\left[Y^2 \,\middle|\, X = x\right],\end{aligned}$$

and so

$$k \, \text{Var}\left[T\left(x;\, Z\right)\right] \leq |\{i : X_i \in L\left(x;\, Z\right)\}| \cdot \text{Var}\left[T\left(x;\, Z\right)\right] \to_p \text{Var}\left[Y \,\middle|\, X = x\right],$$

because k remains fixed while the leaf size gets smaller. Thus, we conclude that

$$\frac{\mathrm{Var}\left[\mathring{T}\left(x;\, Z\right)\right]}{\mathrm{Var}\left[T\left(x;\, Z\right)\right]} \gtrsim k\, \frac{s\ \mathrm{Var}\left[\mathbb{E}\left[T\left(x;\, Z\right) \,\middle|\, Z_1\right]\right]}{\mathrm{Var}\left[Y \,\middle|\, X = x\right]} \gtrsim \nu\left(s\right),$$

as claimed.

Now, in order to verify (37), we first recall that by Lemma 4

$$\mathrm{Var}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right]\right] = \Omega\left(\frac{1}{s\, \log\left(s\right)^d}\right). \tag{38}$$

Thus, any terms that decay faster than the right-hand-side rate can safely be ignored in establishing (37). We begin by verifying that we can take the leaf $L(x)$ containing x to have a small diameter $\mathrm{diam}(L(x))$. Define the truncated tree predictor

$$T'\left(x;\, Z\right) = T\left(x;\, Z\right)\, 1\left(\left\{\mathrm{diam}\left(L(x)\right) \leq s^{-\omega}\right\}\right), \text{ where } \omega = \frac{1}{2}\,\frac{\pi}{d}\,\frac{\log\left(\left(1-\alpha\right)^{-1}\right)}{\log\left(\alpha^{-1}\right)},$$

and define similarly the truncated selection variables $S'_i = S_i\, 1(\{\mathrm{diam}\left(L(x)\right) \leq s^{-\omega}\})$. Now, thanks to the ANOVA decomposition (3rd line), we see that

$$\begin{aligned}
&\mathrm{Var}\left[\mathbb{E}\left[T'\left(x;\, Z\right) \,\middle|\, Z_1\right] - \mathbb{E}\left[T\left(x;\, Z\right) \,\middle|\, Z_1\right]\right] \\
&\quad = \mathrm{Var}\left[\mathbb{E}\left[T\left(x;\, Z\right)\, 1\left(\left\{\mathrm{diam}\left(L(x)\right) > s^{-\omega}\right\}\right) \,\middle|\, Z_1\right]\right] \\
&\quad \leq \frac{1}{s}\,\mathrm{Var}\left[T\left(x;\, Z\right)\, 1\left(\left\{\mathrm{diam}\left(L(x)\right) > s^{-\omega}\right\}\right)\right] \\
&\quad \leq \frac{\sup_{x\in[0,\,1]^d}\left\{\mathbb{E}\left[Y^2 \,\middle|\, X = x\right]\right\}}{s}\,\mathbb{P}\left[\mathrm{diam}\left(L(x)\right) > s^{-\omega}\right],
\end{aligned}$$

where the sup term is bounded by Lipschitz-continuity of the second moment of Y. Thus, by Lemma 2, the variance of the difference between $\mathbb{E}\left[T' \,\middle|\, Z_1\right]$ and $\mathbb{E}\left[T \,\middle|\, Z_1\right]$ decays faster than the target rate (38), and so

$$\mathrm{Var}\left[\mathbb{E}\left[T\left(x;\, Z\right) \,\middle|\, Z_1\right]\right] \sim \mathrm{Var}\left[\mathbb{E}\left[T'\left(x;\, Z\right) \,\middle|\, Z_1\right]\right],$$

provided that T' satisfies (37), as we will see it does. By the same argument, we also note that

$$\text{Var}\left[\mathbb{E}\left[S_1' \,\middle|\, Z_1\right]\right] \sim \text{Var}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right]\right].$$

We can now proceed to analyze T' instead of T.

Recall that our goal is to provide a lower bound on the variance of the expectation of $T'(x;\, Z)$ conditionally on Z_1. First, an elementary decomposition shows that

$$\begin{aligned}
&\text{Var}\left[\mathbb{E}\left[T'(z;\, Z) \,\middle|\, Z_1\right]\right] \\
&\quad = \text{Var}\left[\mathbb{E}\left[T'(z;\, Z) \,\middle|\, X_1\right]\right] + \text{Var}\left[\mathbb{E}\left[T'(z;\, Z) \,\middle|\, X_1,\, Y_1\right] - \mathbb{E}\left[T'(z;\, Z) \,\middle|\, X_1\right]\right] \\
&\quad \geq \text{Var}\left[\mathbb{E}\left[T'(z;\, Z) \,\middle|\, X_1,\, Y_1\right] - \mathbb{E}\left[T'(z;\, Z) \,\middle|\, X_1\right]\right],
\end{aligned}$$

and so it suffices to provide a lower bound for the latter term. Next we note that, thanks to honesty as in Definition 2, part (a), and i.i.d. sampling,

$$\mathbb{E}\left[T'(z;\, Z) \,\middle|\, X_1,\, Y_1\right] - \mathbb{E}\left[T'(z;\, Z) \,\middle|\, X_1\right] = \mathbb{E}\left[S_1' \,\middle|\, X_1\right]\left(Y_1 - \mathbb{E}\left[Y_1 \,\middle|\, X_1\right]\right).$$

Because of honesty and our Lipschitz assumption, the above implies that

$$\begin{aligned}
&\text{Var}\left[\mathbb{E}\left[T'(z;\, Z) \,\middle|\, X_1,\, Y_1\right] - \mathbb{E}\left[T'(z;\, Z) \,\middle|\, X_1\right]\right] \\
&\quad = \text{Var}\left[\mathbb{E}\left[S_1' \,\middle|\, X_1\right](Y_1 - \mu(x))\right] + \mathcal{O}\left(\mathbb{E}\left[S_1'^2\right] s^{-2\omega}\right),
\end{aligned}$$

where we note that the error term decays as $s^{-(1+2\omega)}$, which will be prove to be negligible relative to the main term. Finally, we can verify that

$$\begin{aligned}
&\text{Var}\left[\mathbb{E}\left[S_1' \,\middle|\, X_1\right](Y_1 - \mu(x))\right] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[S_1' \,\middle|\, X_1\right]^2 \mathbb{E}\left[(Y_1 - \mu(x))^2 \,\middle|\, X_1\right]\right] - \mathbb{E}\left[\mathbb{E}\left[S_1' \,\middle|\, X_1\right]\mathbb{E}\left[Y_1 - \mu(x) \,\middle|\, X_1\right]\right]^2.
\end{aligned} \tag{39}$$

Now, because the first two conditional moments of Y given X are Lipschitz, and since $\mathbb{E}\left[S_1' \,\middle|\, X_1\right]$ is 0 for $\|X_1 - x\|_2 > s^{-\omega}$ thanks to our truncating argument, we see that

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[S_1' \,\middle|\, X_1\right]^2 \mathbb{E}\left[\left(Y_1 - \mu\left(x\right)\right)^2 \,\middle|\, X_1\right]\right] &\sim \mathbb{E}\left[\mathbb{E}\left[S_1' \,\middle|\, X_1\right]^2\right] \operatorname{Var}\left[Y \,\middle|\, X = x\right] \\ &\sim \mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, X_1\right]^2\right] \operatorname{Var}\left[Y \,\middle|\, X = x\right].\end{aligned}$$

Meanwhile, the second term in the expansion (39) is of order $1/s^2$ and thus negligible. To recap, we have shown that a version of (37) holds with T replaced by T'; and so (37) must also hold thanks to the previously established coupling result. □

Proof of Corollary 6. For a double sample tree, we start by noting that

$$\begin{aligned}\operatorname{Var}\left[\mathring{T}\right] &= s \operatorname{Var}\left[\mathbb{E}\left[T \,\middle|\, Z_1\right]\right] = s \operatorname{Var}\left[\mathbb{E}\left[1\left(\{1 \in \mathcal{I}\}\right) T \,\middle|\, Z_1\right] + \mathbb{E}\left[1\left(\{1 \notin \mathcal{I}\}\right) T \,\middle|\, Z_1\right]\right] \\ &\geq \frac{s}{2} \operatorname{Var}\left[\mathbb{E}\left[1\left(\{1 \in \mathcal{I}\}\right) T \,\middle|\, Z_1\right]\right] - s \operatorname{Var}\left[\mathbb{E}\left[1\left(\{1 \notin \mathcal{I}\}\right) T \,\middle|\, Z_1\right]\right] \\ &\sim \frac{s}{8} \operatorname{Var}\left[\mathbb{E}\left[T \,\middle|\, Z_1\right] \,\middle|\, 1 \in \mathcal{I}\right] - \frac{s}{4} \operatorname{Var}\left[\mathbb{E}\left[T \,\middle|\, Z_1\right] \,\middle|\, 1 \notin \mathcal{I}\right],\end{aligned}$$

where to verify the last line we note that $\mathbb{P}\left[1 \in \mathcal{I} \,\middle|\, Z_1\right] = \lfloor s/2 \rfloor$ regardless of Z_1. Now, an immediate application of Theorem 5 shows us that

$$\lfloor s/2 \rfloor \operatorname{Var}\left[\mathbb{E}\left[T \,\middle|\, Z_1\right] \,\middle|\, 1 \in \mathcal{I}\right] \gtrsim C_{f,\,d} / \log\left(s\right)^d \operatorname{Var}\left[T\right],$$

which corresponds to the rate we seek to establish. Meanwhile, by standard results going back to Hoeffding [1948],

$$\lceil s/2 \rceil \operatorname{Var}\left[\mathbb{E}\left[T \,\middle|\, Z_1\right] \,\middle|\, 1 \notin \mathcal{I}\right] \leq \operatorname{Var}\left[\mathbb{E}\left[T \,\middle|\, \{Z_j : j \notin \mathcal{I}\}\right] \,\middle|\, \mathcal{I}\right];$$

then, Lemma 2 and the argument used to establish Theorem 3 imply that

$$\operatorname{Var}\left[\mathbb{E}\left[T \,\middle|\, \{Z_j : j \notin \mathcal{I}\}\right] \,\middle|\, \mathcal{I}\right] = \mathcal{O}\left(s^{-\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)} \frac{\pi}{d}}\right),$$

and so the term arising under the $1 \notin \mathcal{I}$ condition is negligibly small. □

C.3 Properties of Subsampled Incremental Base Learners

The results presented in this section rely heavily on the Efron-Stein ANOVA decomposition, summarized here for convenience. Suppose we have any symmetric function $T : \Omega^n \to \mathbb{R}$, and suppose that $Z_1, ..., Z_n$ are independent and identically distributed on Ω such that $\text{Var}\left[T(Z_1, ..., Z_n)\right] < \infty$. Then Efron and Stein [1981] show that there exist functions $T_1, ..., T_n$ such that

$$T\left(Z_1, ..., Z_n\right) = \mathbb{E}\left[T\right] + \sum_{i=1}^{n} T_1\left(Z_i\right) + \sum_{i<j} T_2\left(Z_i, Z_j\right) + ... + T_n\left(Z_1, ..., Z_n\right), \tag{40}$$

and that all $2^n - 1$ random variables on the right side of the above expression are all mean-zero and uncorrelated. It immediately follows that

$$\text{Var}\left[T\right] = \sum_{k=1}^{n} \binom{n}{k} V_k, \text{ where } V_k = \text{Var}\left[T_k\left(Z_1, ..., Z_k\right)\right]. \tag{41}$$

For our purposes, it is also useful to note that the Hájek projection $\mathring{T}$ can be written as

$$\mathring{T}\left(Z_1, ..., Z_n\right) = \mathbb{E}\left[T\right] + \sum_{i=1}^{n} T_1\left(Z_i\right), \text{ and } \text{Var}\left[\mathring{T}\right] = n\, V_1.$$

Thus, the ANOVA decomposition provides a convenient abstract framework for analyzing our quantities of interest.

Proof of Lemma 7. Applying the ANOVA decomposition to the individual trees T, we see that a random forest estimator $\hat{\mu}(x)$ of the form (12) can equivalently be written as

$$\begin{aligned}\hat{\mu}\left(x; Z_1, ..., Z_n\right) = \mathbb{E}\left[T\right] + \binom{n}{s}^{-1} \Bigg(&\binom{n-1}{s-1} \sum_{i=1}^{n} T_1\left(Z_i\right) \\ &+ \binom{n-2}{s-2} \sum_{i<j} T_2\left(Z_i, Z_j\right) + ... + \sum_{i_1<...<i_s} T_s\left(Z_{i_1}, ..., Z_{i_s}\right) \Bigg).\end{aligned}$$

The above formula holds because each training point Z_i appears in $\binom{n-1}{s-1}$ out of $\binom{n}{s}$ possible subsamples of size s, each pair (Z_i, Z_j) appears is $\binom{n-2}{s-2}$ subsets, etc.

Now, we can also show that the Hájek projection of $\hat{\mu}$ is

$$\mathring{\hat{\mu}}(x; Z_1, ..., Z_n) = \mathbb{E}[T] + \frac{s}{n}\sum_{i=1}^{n} T_1(Z_i).$$

As with all projections [Van der Vaart, 2000],

$$\mathbb{E}\left[\left(\hat{\mu}(x) - \mathring{\hat{\mu}}(x)\right)^2\right] = \mathrm{Var}\left[\hat{\mu}(x) - \mathring{\hat{\mu}}(x)\right].$$

Recall that the $T_k(\cdot)$ are all pairwise uncorrelated. Thus, using the notation $s_k = s \cdot (s-1) \cdots (s-k)$ it follows that

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\mu}(x) - \mathring{\hat{\mu}}(x)\right)^2\right] &= \sum_{k=2}^{s} \left(\frac{s_k}{n_k}\right)^2 \binom{n}{k} V_k, \\
&= \sum_{k=2}^{s} \frac{s_k}{n_k} \binom{s}{k} V_k, \\
&\leq \frac{s_2}{n_2} \sum_{k=2}^{s} \binom{s}{k} V_k, \\
&\leq \frac{s_2}{n_2} \mathrm{Var}[T],
\end{aligned}$$

where on the last line we used (41). We recover the stated result by noticing that $s_2/n_2 \leq s^2/n^2$ for all $2 \leq s \leq n$. □

Proof of Theorem 8. Using notation from the previous lemma, let $\sigma_n^2 = s^2/n\, V_1$ be the variance of $\mathring{\hat{\mu}}$. We know that

$$\sigma_n^2 = \frac{s^2}{n^2} n V_1 \leq \frac{s^2}{n^2} \mathrm{Var}[T],$$

and so $\sigma_n \to 0$ as desired. Now, by Theorem 5 or Corollary 6 combined with Lemma 7, we find that

$$\begin{aligned}
\frac{1}{\sigma_n^2}\mathbb{E}\left[\left(\hat{\mu}\left(x\right) - \mathring{\hat{\mu}}\left(x\right)\right)^2\right] &\leq \left(\frac{s}{n}\right)^2 \frac{\operatorname{Var}\left[T\right]}{\sigma_n^2} \\
&= \frac{s}{n} \operatorname{Var}\left[T\right] / \operatorname{Var}\left[\mathring{T}\right] \\
&\lesssim \frac{s}{n} \frac{\log\left(s\right)^d}{C_{f,\,d}/4} \\
&\to 0
\end{aligned}$$

by hypothesis. Thus, by Slutsky's lemma, it suffices to show that (22) holds for the Hájek projection of the random forest $\mathring{\hat{\mu}}(x)$.

By our definition of σ_n, all we need to do is check that $\mathring{\hat{\mu}}$ is asymptotically normal. One way to do so is using the Lyapunov central limit theorem [e.g., Billingsley, 2008]. Writing

$$\mathring{\hat{\mu}}(x) = \frac{s}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[T \,\middle|\, Z_i\right] - \mathbb{E}\left[T\right]\right),$$

it suffices to check that

$$\lim_{n\to\infty}\sum_{i=1}^{n}\mathbb{E}\left[\left|\mathbb{E}\left[T \,\middle|\, Z_i\right] - \mathbb{E}\left[T\right]\right|^{2+\delta}\right] \Big/ \left(\sum_{i=1}^{n}\operatorname{Var}\left[\mathbb{E}\left[T \,\middle|\, Z_i\right]\right]\right)^{1+\delta/2} = 0. \tag{42}$$

Using notation from Section 3.3.1, we write $T = \sum_{i=1}^{n} S_i Y_i$. Thanks to honesty, we can verify that for any index $i > 1$, Y_i is independent of S_i conditionally on X_i and Z_1, and so

$$\mathbb{E}\left[T \,\middle|\, Z_1\right] - \mathbb{E}\left[T\right] = \mathbb{E}\left[S_1\left(Y_1 - \mathbb{E}\left[Y_1 \,\middle|\, X_1\right]\right) \,\middle|\, Z_1\right] + \left(\mathbb{E}\left[\sum_{i=1}^{n} S_i \mathbb{E}\left[Y_i \,\middle|\, X_i\right] \,\middle|\, Z_1\right] - \mathbb{E}\left[T\right]\right).$$

Note that the two right-hand-side terms above are both mean-zero. By Jensen's inequality, we also have

that

$$2^{-(1+\delta)}\,\mathbb{E}\left[\left|\mathbb{E}\left[T \,\middle|\, Z_1\right] - \mathbb{E}\left[T\right]\right|^{2+\delta}\right] \leq \mathbb{E}\left[\left|\mathbb{E}\left[S_1\left(Y_1 - \mathbb{E}\left[Y_1 \,\middle|\, X_1\right]\right) \,\middle|\, Z_1\right]\right|^{2+\delta}\right] + \mathbb{E}\left[\left|\mathbb{E}\left[\sum_{i=1}^{n} S_i \mathbb{E}\left[Y_i \,\middle|\, X_i\right] \,\middle|\, Z_1\right] - \mathbb{E}\left[T\right]\right|^{2+\delta}\right].$$

Now, again by honesty, $\mathbb{E}\left[S_1 \,\middle|\, Z_1\right] = \mathbb{E}\left[S_1 \,\middle|\, X_1\right]$, and so our uniform $(2+\delta)$-moment bounds on the distribution of Y_i conditional on X_i implies that

$$\begin{aligned}\mathbb{E}\left[\left|\mathbb{E}\left[S_1\left(Y_1 - \mathbb{E}\left[Y_1 \,\middle|\, X_1\right]\right) \,\middle|\, Z_1\right]\right|^{2+\delta}\right] &= \mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, X_1\right]^{2+\delta}\left|Y_1 - \mathbb{E}\left[Y_1 \,\middle|\, X_1\right]\right|^{2+\delta}\right] \\ &\leq M\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, X_1\right]^{2+\delta}\right] \leq M\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, X_1\right]^{2}\right],\end{aligned}$$

because $S_1 \leq 1$. Meanwhile, because $\mathbb{E}\left[Y \,\middle|\, X = x\right]$ is Lipschitz, we can define $u := \sup\left\{\left|\mathbb{E}\left[Y \,\middle|\, X = x\right]\right| : x \in [0,\,1]^d\right\}$, and see that

$$\begin{aligned}\mathbb{E}\left[\left|\mathbb{E}\left[\sum_{i=1}^{n} S_i \mathbb{E}\left[Y_i \,\middle|\, X_i\right] \,\middle|\, Z_1\right] - \mathbb{E}\left[T\right]\right|^{2+\delta}\right] &\leq (2u)^{\delta}\,\mathrm{Var}\left[\mathbb{E}\left[\sum_{i=1}^{n} S_i \mathbb{E}\left[Y_i \,\middle|\, X_i\right] \,\middle|\, Z_1\right]\right] \\ &\leq 2^{1+\delta}u^{2+\delta}\left(\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right]^2\right] + \mathrm{Var}\left[(n-1)\mathbb{E}\left[S_2 \,\middle|\, Z_1\right]\right]\right) \leq (2u)^{2+\delta}\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, X_1\right]^2\right].\end{aligned}$$

Thus, the condition we need to check simplifies to

$$\lim_{n\to\infty} n\,\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, X_1\right]^2\right] \,/\, \left(n\,\mathrm{Var}\left[\mathbb{E}\left[T \,\middle|\, Z_1\right]\right]\right)^{1+\delta/2} = 0.$$

Finally, as argued in the proof of Theorem 5,

$$\mathrm{Var}\left[\mathbb{E}\left[T \,\middle|\, Z_1\right]\right] \gtrsim \mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, X_1\right]^2\right]\mathrm{Var}\left[Y \,\middle|\, X = x\right].$$

Because $\text{Var}\left[Y \,\middle|\, X = x\right] > 0$ by assumption, we can use Lemma 4 to conclude our argument, noting that

$$\left(n\,\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, X_1\right]^2\right]\right)^{-\delta/2} \lesssim \left(\frac{C_{f,\,d}}{2k}\frac{n}{s\log(s)^d}\right)^{-\delta/2},$$

which goes to 0 thanks to our assumptions on the scaling of s (and the factor 2 comes from potentially using a double-sample tree). □

Proof of Theorem 9. Let F denote the distribution from which we drew Z_1, ..., Z_n. Then, the variance σ_n^2 of the Hájek projection of $\hat{\mu}(x)$ is

$$\begin{aligned}\sigma_n^2 &= \sum_{i=1}^n \left(\mathbb{E}_{Z\sim F}\left[\hat{\mu}(x) \,\middle|\, Z_i\right] - \mathbb{E}_{Z\sim F}\left[\hat{\mu}(x)\right]\right)^2 \\ &= \frac{s^2}{n^2}\sum_{i=1}^n \left(\mathbb{E}_{Z\sim F}\left[T \,\middle|\, Z_i\right] - \mathbb{E}_{Z\sim F}\left[T\right]\right)^2,\end{aligned}$$

whereas we can check that the infinitesimal jackknife as defined in (8) is equal to

$$\widehat{V}_{IJ} = \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2\frac{s^2}{n^2}\sum_{i=1}^n \left(\mathbb{E}_{Z^*\subset\widehat{F}}\left[T \,\middle|\, Z_1^* = Z_i\right] - \mathbb{E}_{Z^*\subset\widehat{F}}\left[T\right]\right)^2,$$

where $\widehat{F}$ is the empirical distribution on $\{Z_1, ..., Z_n\}$. Recall that we are sampling the Z^* from $\widehat{F}$ without replacement.

It is useful to write our expression of interest $\widehat{V}_{IJ}$ using the Hájek projection $\mathring{T}$ of T:

$$\begin{aligned}\widehat{V}_{IJ} &= \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2\frac{s^2}{n^2}\sum_{i=1}^n \left(A_i + R_i\right)^2, \text{ where} \\ A_i &= \mathbb{E}_{Z^*\subset\widehat{F}}\left[\mathring{T} \,\middle|\, Z_1^* = Z_i\right] - \mathbb{E}_{Z^*\subset\widehat{F}}\left[\mathring{T}\right] \text{ and} \\ R_i &= \mathbb{E}_{Z^*\subset\widehat{F}}\left[T - \mathring{T} \,\middle|\, Z_1^* = Z_i\right] - \mathbb{E}_{Z^*\subset\widehat{F}}\left[T - \mathring{T}\right].\end{aligned}$$

As we show in Lemma 12, the main effects A_i give us σ_n^2, in that

$$\frac{1}{\sigma_n^2}\frac{s^2}{n^2}\sum_{i=1}^n A_i^2 \to_p 1. \tag{43}$$

Meanwhile, Lemma 13 establishes that the B_i all satisfy

$$\mathbb{E}\left[R_i^2\right] \lesssim \frac{2}{n} \operatorname{Var}\left[T\left(x;, Z_1, ..., Z_s\right)\right], \tag{44}$$

and so

$$\begin{aligned}\mathbb{E}\left[\frac{s^2}{n^2}\sum_{i=1}^{n} R_i^2\right] &\lesssim \frac{2\,s^2}{n^2} \operatorname{Var}\left[T\left(x;, Z_1, ..., Z_s\right)\right] \\ &\lesssim \frac{2\,s}{n} \log\left(n\right)^d \sigma_n^2.\end{aligned}$$

Because all terms are positive and $s \log\left(n\right)^d / n$ goes to zero by hypothesis, Markov's inequality implies that

$$\frac{1}{\sigma_n^2}\frac{s^2}{n^2}\sum_{i=1}^{n} R_i^2 \to_p 0.$$

Using Cauchy-Schwarz to bound the cross terms of the form $A_i R_i$, and noting that $\lim_{n\to\infty} n(n-1)/(n-s)^2 = 1$, we can thus conclude that $\widehat{V}_{IJ}/\sigma_n^2$ converges in probability to 1. □

Lemma 12. *Under the conditions of Theorem 9, (43) holds.*

Proof. We can write

$$\begin{aligned}A_i &= \mathbb{E}_{Z^* \subset \widehat{F}}\left[\mathring{T} \mid Z_1^* = Z_i\right] - \mathbb{E}_{Z^* \subset \widehat{F}}\left[\mathring{T}\right] \\ &= \left(1 - \frac{s}{n}\right) T_1\left(Z_i\right) + \left(\frac{s-1}{n-1} - \frac{s}{n}\right)\sum_{j \neq i} T_1\left(Z_j\right),\end{aligned}$$

and so our sum of interest is asymptotically unbiased for σ_n^2:

$$\begin{aligned}\mathbb{E}\left[\frac{n-1}{n}\left(\frac{n}{n-s}\right)^2 \frac{s^2}{n^2}\sum_{i=1}^{n} A_i^2\right] &= \frac{s^2}{n}\mathbb{E}\left[T_1\left(Z\right)^2\right] \\ &= \frac{s}{n} \operatorname{Var}\left[\mathring{T}\left(Z_1, ..., Z_s\right)\right] \\ &= \sigma_n^2.\end{aligned}$$

Finally, to establish concentration, we first note that the above calculation also implies that $\sigma_n^{-2} \frac{s^2}{n^2} \sum_{i=1}^n (A_i - T_1(Z_i))^2 \to 0$. Meanwhile, following the argumentation in the proof of Theorem 8, we can apply $(2+\delta)$-moment bounds on $Y_i - \mathbb{E}\left[Y_i \,\middle|\, X_i\right]$ to verify that

$$\lim_{u \to \infty} \lim_{n \to \infty} \left(\mathbb{E}\left[T_1^2(Z_1)\right] - \mathbb{E}\left[\min\left\{u, T_1^2(Z_1)\right\}\right] \right) = 0,$$

and so we obtain can apply a truncation-based argument to derive a weak law of large numbers for triangular arrays for $\sigma_n^{-2} \frac{s^2}{n^2} \sum_{i=1}^n T_1^2(Z_i)$. □

Lemma 13. *Under the conditions of Theorem 9, (44) holds.*

Proof. Without loss of generality, we establish (44) for R_1. Using the ANOVA decomposition (40), we can write our term of interest as

$$\begin{aligned}
R_1 &= \mathbb{E}_{Z^* \subset \widehat{F}}\left[T - \mathring{T} \,\middle|\, Z_1^* = Z_1\right] - \mathbb{E}_{Z^* \subset \widehat{F}}\left[T - \mathring{T}\right] \\
&= \left(\frac{s-1}{n-1} - \binom{s}{2}\Big/\binom{n}{2}\right) \sum_{i=2}^n T_2\left(Z_1, Z_i\right) \\
&\quad + \left(\binom{s-1}{2}\Big/\binom{n-1}{2} - \binom{s}{2}\Big/\binom{n}{2}\right) \sum_{2 \le i < j \le n} T_2\left(Z_i, Z_j\right) \\
&+ \left(\binom{s-1}{2}\Big/\binom{n-1}{2} - \binom{s}{3}\Big/\binom{n}{3}\right) \sum_{2 \le i < j \le n} T_3\left(Z_1, Z_i, Z_j\right) \\
&\quad + \left(\binom{s-1}{3}\Big/\binom{n-1}{3} - \binom{s}{3}\Big/\binom{n}{3}\right) \sum_{2 \le i < j < k \le n} T_3\left(Z_i, Z_j, Z_k\right) \\
&+ \ldots
\end{aligned}$$

Because all the terms in the ANOVA expansion are mean-zero and uncorrelated, we see using notation

from (41) that

$$\begin{aligned}
\mathbb{E}\left[R_1^2\right] = (n-1)\left(\frac{s-1}{n-1} - \binom{s}{2}\Big/\binom{n}{2}\right)^2 V_2 \\
+ \binom{n-1}{2}\left(\binom{s-1}{2}\Big/\binom{n-1}{2} - \binom{s}{2}\Big/\binom{n}{2}\right)^2 V_2 \\
+ \binom{n-1}{2}\left(\binom{s-1}{2}\Big/\binom{n-1}{2} - \binom{s}{3}\Big/\binom{n}{3}\right)^2 V_3 \\
+ \binom{n-1}{3}\left(\binom{s-1}{3}\Big/\binom{n-1}{3} - \binom{s}{3}\Big/\binom{n}{3}\right)^2 V_3 \\
+ \dots
\end{aligned}$$

Recall that

$$\sum_{k=1}^{s}\binom{s}{k}V_k = \operatorname{Var}\left[T\left(Z_1, ..., Z_s\right)\right].$$

The above sum is maximized when all the variance is contained in second-order terms, and $\binom{s}{2}V_2 = \operatorname{Var}[T]$. This implies that

$$\begin{aligned}
\mathbb{E}\left[R_1^2\right] &\lesssim (n-1)\left(\frac{s-1}{n-1} - \binom{s}{2}\Big/\binom{n}{2}\right)^2 \binom{s}{2}^{-1} \operatorname{Var}\left[T\left(Z_1, ..., Z_s\right)\right] \\
&\sim \frac{2}{n}\operatorname{Var}\left[T\left(Z_1, ..., Z_s\right)\right],
\end{aligned}$$

thus completing the proof. □

Proof of Proposition 10. Let N_i^* denote whether or not the i-training example was used for a subsample, as in (13). For trivial trees

$$T(x;\, \xi,\, Z_{i_1}, ...,\, Z_{i_s}) = \frac{1}{s}\sum_{j=1}^{s} Y_{i_j}$$

we can verify that for any $i = 1, ..., n$, $\mathbb{E}_*\left[\hat{\mu}^*\right]\mathbb{E}\left[N_i^*\right] = s/n\ \overline{Y}$,

$$\begin{aligned}
\mathbb{E}_*\left[\hat{\mu}^*\, N_1^*\right] &= \frac{s}{n}\left(\frac{Y_i}{s} + \frac{s-1}{s}\frac{n\overline{Y} - Y_i}{n-1}\right) = \frac{1}{n}\frac{n-s}{n-1}Y_i + \frac{s-1}{n-1}\overline{Y}, \text{ and} \\
\operatorname{Cov}_*\left[\hat{\mu}^*,\, N_i^*\right] &= \frac{1}{n}\frac{n-s}{n-1}Y_i + \left(\frac{s-1}{n-1} - \frac{s}{n}\right)\overline{Y} = \frac{1}{n-1}\frac{n-s}{n}\left(Y_i - \overline{Y}\right).
\end{aligned}$$

Thus, we find that

$$\begin{aligned}\widehat{V}_{IJ} &= \frac{n-1}{n}\left(\frac{n}{n-s}\right)^2 \sum_{i=1}^{n} \mathrm{Cov}_*\left[\hat{\mu}^*,\, N_i^*\right]^2 \\ &= \frac{1}{n\,(n-1)} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \\ &= \widehat{V}_{simple},\end{aligned}$$

as we sought to verify. □

C.4 Extension to Causal Forests

Proof of Theorem 11. Our argument mirrors the proof of Theorem 1. The main steps involve bounding the bias of causal forests with an analogue to Theorem 3 and their incrementality using an analogue to Theorem 5. In general, we find that the same arguments as used with regression forests go through, but the constants in the results get worse by a factor ε depending on the amount of overlap (6). Given these results, the subsampling-based argument from Section 3.3.2 can be reproduced almost verbatim, and the final proof of Theorem 11 is identical to that of Theorem 1 presented at the beginning of Appendix C.

Bias. Under the conditions of Lemma 2, suppose that $\mathbb{E}\left[Y^{(0)} \,\middle|\, X = x\right]$ and $\mathbb{E}\left[Y^{(1)} \,\middle|\, X = x\right]$ are Lipschitz continuous, that the trees Γ comprising the random forest are honest, and, moreover, that the overlap condition (6) holds for some $\varepsilon > 0$. These conditions also imply that $|\mathbb{E}[Y^{(0)} \,|\, X = x]|$, $|\mathbb{E}[Y^{(1)} \,|\, X = x]| \leq M$ for some constant M, for all $x \in [0,\, 1]^d$. Then, provided that $\alpha \leq 0.2$, the bias of the random forest at x is bounded by

$$\left|\mathbb{E}\left[\hat{\tau}\left(x\right)\right] - \tau\left(x\right)\right| \lesssim 2M\, d\, \left(\frac{\varepsilon\, s}{2k-1}\right)^{-\frac{1}{2}\frac{\log\left((1-\alpha)^{-1}\right)}{\log\left(\alpha^{-1}\right)}\frac{\pi}{d}}.$$

To establish this claim, we first seek with an analogue to Lemma 2, except now s in (31) is replaced by $s_{\min}$, i.e., the minimum of the number of cases (i.e., observations with $W_i = 1$) or controls (i.e., observations with $W_i = 0$) in the sample. A straight-forward computation then shows that $s_{\min}/s \gtrsim \varepsilon$,

and that a variant of (32) where we replace s with εs still holds for large s. Next, to bound the bias itself, we start by applying unconfoundedness as in (25); then, the argument of Theorem 3 goes through without modifications, provided we replace every instance of "s" with "εs".

Incrementality. Suppose that the conditions of Lemma 4 hold and that Γ is an honest k-regular causal tree in the sense of Definitions 2b and 4b. Suppose moreover that $\mathbb{E}\left[Y^{(0/1)} \,\middle|\, X = x\right]$ and $\text{Var}\left[Y^{(0/1)} \,\middle|\, X = x\right]$ are all Lipschitz continuous at x, and that $\text{Var}\left[Y \,\middle|\, X = x\right] > 0$. Suppose, finally, that the overlap condition (6) holds with $\varepsilon > 0$. Then T is $\nu(s)$-incremental at x with

$$\nu(s) = \varepsilon \, C_{f,\,d} \,/\, \log(s)^d,$$

where $C_{f,\,d}$ is the constant from Lemma 4.

To prove this claim, we again focus on the case where $f(x) = 1$, in which case we use $C_{f,\,d} = 2^{-(d+1)}\,(d-1)!$. We begin by setting up notation as in the proof of Lemma 4. We write our causal tree as $\Gamma(x;\, Z) = \sum_{i=1}^{s} S_i Y_i$, where

$$S_i = \begin{cases} \left|\{i : X_i \in L(x;\, Z)\},\, W_i = 1\right|^{-1} & \text{if } X_i \in L(x;\, Z) \text{ and } W_i = 1, \\ -\left|\{i : X_i \in L(x;\, Z)\},\, W_i = 0\right|^{-1} & \text{if } X_i \in L(x;\, Z) \text{ and } W_i = 0, \\ 0 & \text{else,} \end{cases}$$

where $L(x;\, Z)$ denotes the leaf containing x, and let

$$P_i^W = 1\left(\{X_i \text{ is a } k\text{-PNN of } x \text{ among points with treatment status } W_i\}\right).$$

Finally, in a break from Lemma 4, define $w_{\min}(x;\, Z)$ as the minority class within the leaf $L(x;\, Z)$; more formally,

$$w_{\min} = 1\left(\left\{\left|\{i : X_i \in L(x;\, Z)\},\, W_i = 1\right| \leq \left|\{i : X_i \in L(x;\, Z)\},\, W_i = 0\right|\right\}\right).$$

By regularity of Γ, we know that the leaf $L(x;\, Z)$ can contain at most $2k-1$ examples from its minority

class, and so $P_i^W = 0$ and $W = w_{\min}$ together imply that $S_i = 0$. Thus, we can verify that

$$\mathbb{E}\left[|S_1|\ 1\left(\{W_1 = w_{\min}\}\right) \,\middle|\, Z_1\right] \leq \frac{1}{k}\,\mathbb{E}\left[P_1^W \,\middle|\, Z_1\right].$$

We are now ready to use the same machinery as before. The random variables P_1^W now satisfy

$$\mathbb{P}\left[\mathbb{E}\left[P_1^W \,\middle|\, Z_1\right] \geq \frac{1}{s^2\,\mathbb{P}\left[W = W_1\right]^2}\right] \lesssim k\,\frac{2^{d+1}\log\left(s\right)^d}{(d-1)!}\,\frac{1}{s\,\mathbb{P}\left[W = W_1\right]};$$

by the above argument and ε-overlap (6), this immediately implies that

$$\mathbb{P}\left[\mathbb{E}\left[|S_1|\ 1\left(\{W_1 = w_{\min}\} \,\middle|\, Z_1\right)\right] \geq \frac{1}{k\,\varepsilon^2\,s^2}\right] \lesssim k\,\frac{2^{d+1}\log\left(s\right)^d}{(d-1)!}\,\frac{1}{\varepsilon s}.$$

By construction, we know that

$$\mathbb{E}\left[\mathbb{E}\left[|S_1|\ 1\left(\{W_1 = w_{\min}\}\right) \,\middle|\, Z_1\right]\right] = \mathbb{E}\left[|S_1|\ 1\left(\{W_1 = w_{\min}\}\right)\right] = \frac{1}{s},$$

which by the same argument as before implies that

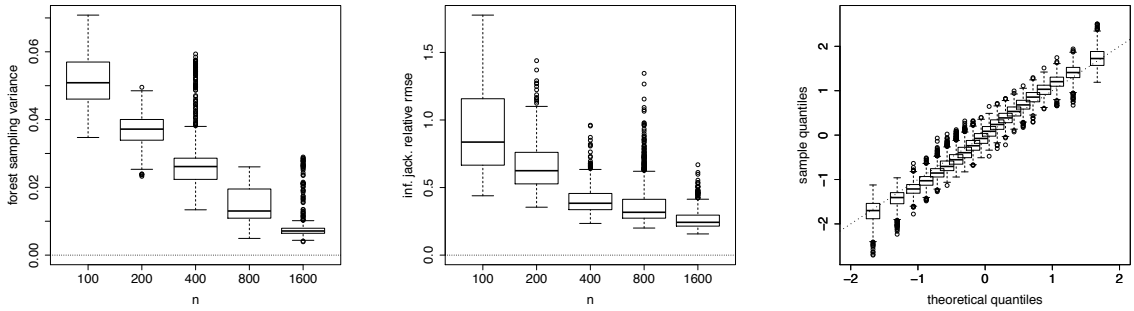
$$\mathbb{E}\left[\mathbb{E}\left[|S_1|\ 1\left(\{W_1 = w_{\min}\}\right) \,\middle|\, Z_1\right]^2\right] \gtrsim \frac{(d-1)!}{2^{d+1}\log\left(s\right)^d}\,\frac{\varepsilon}{k\,s}.$$

By monotonicity, we then conclude that

$$\mathbb{E}\left[\mathbb{E}\left[S_1 \,\middle|\, Z_1\right]^2\right] = \mathbb{E}\left[\mathbb{E}\left[|S_1| \,\middle|\, Z_1\right]^2\right] \gtrsim \frac{(d-1)!}{2^{d+1}\log\left(s\right)^d}\,\frac{\varepsilon}{k\,s}.$$

The second part of the proof follows from a straight-forward adaptation of Theorem 5. □

List of Figures



Causal forest variance | Inf. jack. relative RMSE | Prediction QQ-plot

Figure 1: Graphical diagnostics for causal forests in the setting of (27). The first two panels evaluate the sampling error of causal forests and our infinitesimal jackknife estimate of variance over 1,000 randomly draw test points, with $d = 20$. The right-most panel shows standardized Gaussian QQ-plots for predictions at the same 1000 test points, with $n = 800$ and $d = 20$. The first two panels are computed over 50 randomly drawn training sets, and the last one over 20 training sets.

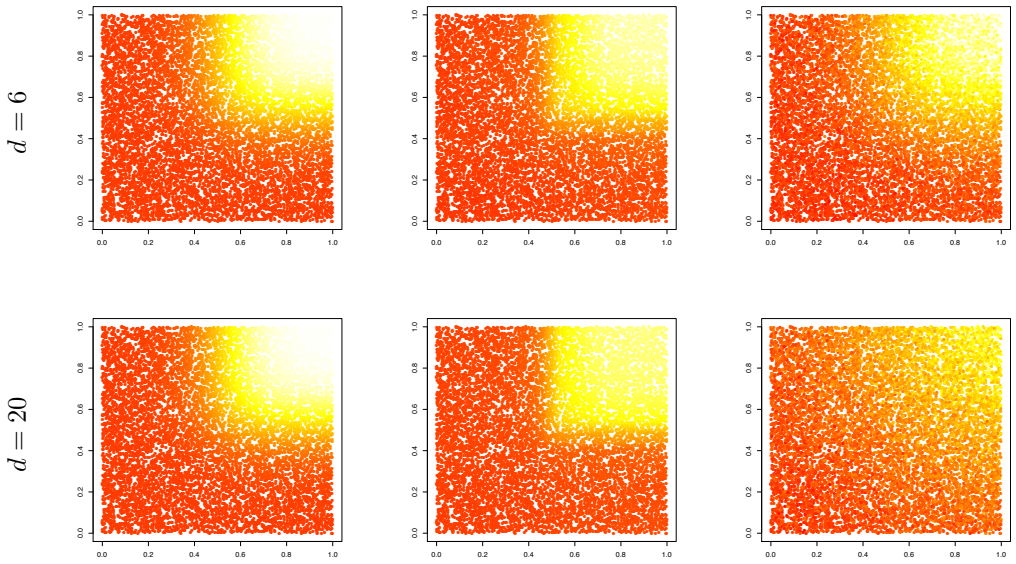


Figure 2: The true treatment effect $\tau(X_i)$ at 10,000 random test examples X_i, along with estimates $\hat{\tau}(X_i)$ produced by a causal forest and optimally-tuned k-NN, on data drawn according to (29) with $d = 6$, 20. The test points are plotted according to their first two coordinates; the treatment effect is denoted by color, from dark (low) to light (high). On this simulation instance, causal forests and k^*-NN had a mean-squared error of 0.03 and 0.13 respectively for $d = 6$, and of 0.05 and 0.62 respectively for $d = 20$. The optimal tuning choices for k-NN were $k^* = 39$ for $d = 6$, and $k^* = 24$ for $d = 20$.

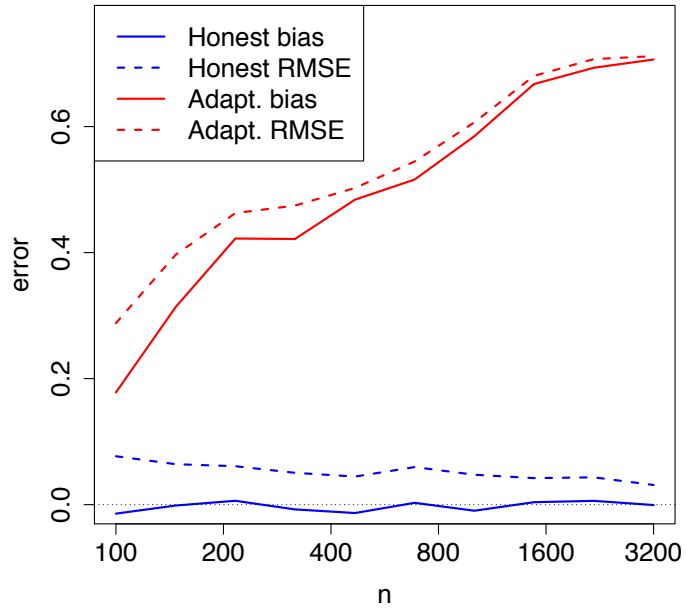


Figure 3: Comparison of the performance of honest and adaptive causal forests when predicting at $x_0 = (0, 0, \ldots, 0)$, which is a corner of the support of the features X_i. Both forests have $B = 500$ trees, and use a leaf-size of $k = 1$. We use a subsample size $s = n^{0.8}$ for adaptive forests and $s = 2\,n^{0.8}$ for honest forests. All results are averaged over 40 replications; we report both bias and root-mean-squared error (RMSE).

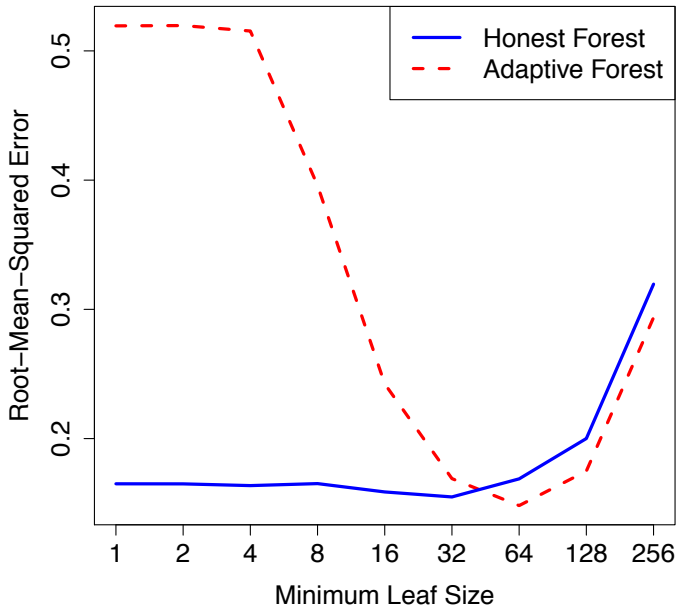


Figure 4: Comparison of the root-mean-squared error of honest and adaptive forests in the setting of Table 2, with $d = 8$. Honest forests use $s = 2500$ (i.e., $|\mathcal{I}| = 1250$) while adaptive forests use $s = 1250$, such that both methods grow trees of the same depth. Both forests have $B = 500$, and results are averaged over 100 simulation replications.

List of Tables

	mean-squared error			coverage		
d	CF	10-NN	100-NN	CF	10-NN	100-NN
2	0.02 (0)	0.21 (0)	0.09 (0)	0.95 (0)	0.93 (0)	0.62 (1)
5	0.02 (0)	0.24 (0)	0.12 (0)	0.94 (1)	0.92 (0)	0.52 (1)
10	0.02 (0)	0.28 (0)	0.12 (0)	0.94 (1)	0.91 (0)	0.51 (1)
15	0.02 (0)	0.31 (0)	0.13 (0)	0.91 (1)	0.90 (0)	0.48 (1)
20	0.02 (0)	0.32 (0)	0.13 (0)	0.88 (1)	0.89 (0)	0.49 (1)
30	0.02 (0)	0.33 (0)	0.13 (0)	0.85 (1)	0.89 (0)	0.48 (1)

Table 1: Comparison of the performance of a causal forests (CF) with that of the k-nearest neighbors (k-NN) estimator with $k = 10, 100$, on the setup (27). The numbers in parentheses indicate the (rounded) standard sampling error for the last printed digit, obtained by aggregating performance over 500 simulation replications.

	mean-squared error			coverage		
d	CF	7-NN	50-NN	CF	7-NN	50-NN
2	0.04 (0)	0.29 (0)	0.04 (0)	0.97 (0)	0.93 (0)	0.94 (0)
3	0.03 (0)	0.29 (0)	0.05 (0)	0.96 (0)	0.93 (0)	0.92 (0)
4	0.03 (0)	0.30 (0)	0.08 (0)	0.94 (0)	0.93 (0)	0.86 (1)
5	0.03 (0)	0.31 (0)	0.11 (0)	0.93 (1)	0.92 (0)	0.77 (1)
6	0.02 (0)	0.34 (0)	0.15 (0)	0.93 (1)	0.91 (0)	0.68 (1)
8	0.03 (0)	0.38 (0)	0.21 (0)	0.90 (1)	0.90 (0)	0.57 (1)

Table 2: Comparison of the performance of a causal forests (CF) with that of the k-nearest neighbors (k-NN) estimator with $k = 7$, 50, on the setup (28). The numbers in parentheses indicate the (rounded) standard sampling error for the last printed digit, obtained by aggregating performance over 25 simulation replications.

	mean-squared error			coverage		
d	CF	10-NN	100-NN	CF	10-NN	100-NN
2	0.02 (0)	0.20 (0)	0.02 (0)	0.94 (0)	0.93 (0)	0.94 (0)
3	0.02 (0)	0.20 (0)	0.03 (0)	0.90 (0)	0.93 (0)	0.90 (0)
4	0.02 (0)	0.21 (0)	0.06 (0)	0.84 (1)	0.93 (0)	0.78 (1)
5	0.02 (0)	0.22 (0)	0.09 (0)	0.81 (1)	0.93 (0)	0.67 (0)
6	0.02 (0)	0.24 (0)	0.15 (0)	0.79 (1)	0.92 (0)	0.58 (0)
8	0.03 (0)	0.29 (0)	0.26 (0)	0.73 (1)	0.90 (0)	0.45 (0)

Table 3: Comparison of the performance of a causal forests (CF) with that of the k-nearest neighbors (k-NN) estimator with $k = 10, 100$, on the setup (29). The numbers in parentheses indicate the (rounded) standard sampling error for the last printed digit, obtained by aggregating performance over 40 simulation replications.

n	d	s	MSE	Coverage	Variance	n	d	s	MSE	Coverage	Variance	n	d	s	MSE	Coverage	Variance
1000	2	100	0.16 (1)	0.29 (3)	0.01 (0)	2000	4	200	0.14 (1)	0.25 (3)	0.01 (0)	5000	6	500	0.05 (0)	0.52 (3)	0.01 (0)
1000	2	200	0.09 (1)	0.69 (5)	0.03 (0)	2000	4	400	0.06 (1)	0.70 (4)	0.02 (0)	5000	6	1000	0.03 (0)	0.75 (3)	0.01 (0)
1000	2	250	0.07 (1)	0.79 (4)	0.03 (0)	2000	4	500	0.05 (1)	0.81 (3)	0.02 (0)	5000	6	1250	0.02 (0)	0.79 (3)	0.01 (0)
1000	2	333	0.07 (1)	0.89 (3)	0.04 (0)	2000	4	667	0.04 (0)	0.90 (2)	0.03 (0)	5000	6	1667	0.02 (0)	0.86 (2)	0.02 (0)
1000	2	500	0.07 (1)	0.94 (2)	0.07 (0)	2000	4	1000	0.04 (0)	0.95 (1)	0.04 (0)	5000	6	2500	0.02 (0)	0.92 (1)	0.02 (0)
1000	2	667	0.08 (1)	0.90 (4)	0.08 (1)	2000	4	1333	0.05 (0)	0.96 (1)	0.06 (0)	5000	6	3333	0.03 (0)	0.96 (1)	0.03 (0)
1000	3	100	0.25 (2)	0.16 (2)	0.01 (0)	2000	5	200	0.17 (1)	0.18 (2)	0.01 (0)	5000	8	500	0.06 (1)	0.42 (3)	0.01 (0)
1000	3	200	0.13 (2)	0.53 (5)	0.02 (0)	2000	5	400	0.07 (1)	0.65 (6)	0.02 (0)	5000	8	1000	0.03 (0)	0.69 (2)	0.01 (0)
1000	3	250	0.11 (2)	0.66 (6)	0.03 (0)	2000	5	500	0.05 (1)	0.75 (5)	0.02 (0)	5000	8	1250	0.03 (0)	0.73 (3)	0.01 (0)
1000	3	333	0.09 (1)	0.81 (5)	0.04 (0)	2000	5	667	0.05 (0)	0.84 (3)	0.03 (0)	5000	8	1667	0.03 (0)	0.81 (2)	0.01 (0)
1000	3	500	0.08 (1)	0.90 (3)	0.06 (0)	2000	5	1000	0.04 (0)	0.91 (1)	0.04 (0)	5000	8	2500	0.03 (0)	0.88 (2)	0.02 (0)
1000	3	667	0.08 (1)	0.73 (5)	0.04 (0)	2000	5	1333	0.05 (0)	0.92 (2)	0.05 (0)	5000	8	3333	0.03 (0)	0.92 (1)	0.03 (0)
1000	4	100	0.32 (1)	0.12 (1)	0.01 (0)	2000	6	200	0.22 (1)	0.12 (1)	0.01 (0)	10000	2	1000	0.01 (0)	0.95 (1)	0.01 (0)
1000	4	200	0.16 (1)	0.43 (3)	0.03 (0)	2000	6	400	0.09 (1)	0.52 (5)	0.02 (0)	10000	2	2000	0.01 (0)	0.96 (1)	0.02 (0)
1000	4	250	0.13 (1)	0.58 (4)	0.03 (0)	2000	6	500	0.07 (1)	0.68 (4)	0.02 (0)	10000	2	2500	0.02 (0)	0.96 (0)	0.02 (0)
1000	4	333	0.09 (1)	0.76 (3)	0.04 (0)	2000	6	667	0.05 (0)	0.82 (3)	0.03 (0)	10000	2	3333	0.02 (0)	0.96 (0)	0.02 (0)
1000	4	500	0.07 (1)	0.91 (2)	0.06 (0)	2000	6	1000	0.04 (0)	0.89 (2)	0.03 (0)	10000	2	5000	0.03 (0)	0.97 (0)	0.04 (0)
1000	4	667	0.07 (1)	0.79 (3)	0.04 (1)	2000	6	1333	0.05 (0)	0.94 (1)	0.05 (0)	10000	2	6667	0.04 (0)	0.98 (0)	0.06 (0)
1000	5	100	0.34 (2)	0.10 (1)	0.01 (0)	2000	8	200	0.24 (1)	0.12 (1)	0.01 (0)	10000	3	1000	0.01 (0)	0.84 (1)	0.01 (0)
1000	5	200	0.16 (2)	0.41 (6)	0.02 (0)	2000	8	400	0.08 (0)	0.61 (4)	0.02 (0)	10000	3	2000	0.01 (0)	0.91 (1)	0.01 (0)
1000	5	250	0.12 (2)	0.59 (5)	0.03 (0)	2000	8	500	0.06 (0)	0.78 (2)	0.02 (0)	10000	3	2500	0.01 (0)	0.92 (1)	0.01 (0)
1000	5	333	0.09 (1)	0.80 (4)	0.04 (0)	2000	8	667	0.05 (0)	0.85 (1)	0.02 (0)	10000	3	3333	0.02 (0)	0.94 (1)	0.02 (0)
1000	5	500	0.07 (1)	0.89 (3)	0.06 (0)	2000	8	1000	0.04 (0)	0.91 (1)	0.04 (0)	10000	3	5000	0.02 (0)	0.95 (0)	0.03 (0)
1000	5	667	0.07 (1)	0.77 (4)	0.04 (0)	2000	8	1333	0.04 (0)	0.89 (3)	0.04 (0)	10000	3	6667	0.03 (0)	0.97 (0)	0.04 (0)
1000	6	100	0.41 (3)	0.07 (1)	0.01 (0)	5000	2	500	0.02 (0)	0.86 (2)	0.01 (0)	10000	4	1000	0.02 (0)	0.73 (3)	0.01 (0)
1000	6	200	0.22 (3)	0.31 (4)	0.02 (0)	5000	2	1000	0.02 (0)	0.92 (1)	0.02 (0)	10000	4	2000	0.02 (0)	0.85 (2)	0.01 (0)
1000	6	250	0.17 (3)	0.48 (6)	0.03 (0)	5000	2	1250	0.02 (0)	0.94 (1)	0.02 (0)	10000	4	2500	0.02 (0)	0.87 (1)	0.01 (0)
1000	6	333	0.12 (2)	0.68 (7)	0.04 (0)	5000	2	1667	0.03 (0)	0.95 (1)	0.03 (0)	10000	4	3333	0.02 (0)	0.90 (1)	0.01 (0)
1000	6	500	0.08 (2)	0.89 (3)	0.05 (0)	5000	2	2500	0.04 (0)	0.96 (0)	0.05 (0)	10000	4	5000	0.02 (0)	0.93 (1)	0.02 (0)
1000	6	667	0.07 (1)	0.75 (6)	0.04 (0)	5000	2	3333	0.05 (0)	0.97 (0)	0.06 (0)	10000	4	6667	0.02 (0)	0.95 (0)	0.03 (0)
1000	8	100	0.51 (2)	0.06 (0)	0.01 (0)	5000	3	500	0.02 (0)	0.75 (3)	0.01 (0)	10000	5	1000	0.02 (0)	0.65 (3)	0.01 (0)
1000	8	200	0.29 (1)	0.20 (2)	0.02 (0)	5000	3	1000	0.02 (0)	0.89 (2)	0.01 (0)	10000	5	2000	0.02 (0)	0.79 (2)	0.01 (0)
1000	8	250	0.23 (1)	0.31 (2)	0.03 (0)	5000	3	1250	0.02 (0)	0.91 (1)	0.02 (0)	10000	5	2500	0.02 (0)	0.83 (2)	0.01 (0)
1000	8	333	0.16 (1)	0.53 (4)	0.04 (0)	5000	3	1667	0.02 (0)	0.94 (1)	0.02 (0)	10000	5	3333	0.02 (0)	0.87 (1)	0.01 (0)
1000	8	500	0.10 (1)	0.86 (2)	0.06 (0)	5000	3	2500	0.03 (0)	0.96 (1)	0.03 (0)	10000	5	5000	0.02 (0)	0.92 (1)	0.02 (0)
1000	8	667	0.08 (1)	0.70 (5)	0.04 (0)	5000	3	3333	0.03 (0)	0.97 (0)	0.05 (0)	10000	5	6667	0.02 (0)	0.94 (0)	0.02 (0)
2000	2	200	0.05 (0)	0.64 (4)	0.01 (0)	5000	4	500	0.03 (0)	0.61 (3)	0.01 (0)	10000	6	1000	0.02 (0)	0.62 (3)	0.00 (0)
2000	2	400	0.03 (0)	0.88 (1)	0.02 (0)	5000	4	1000	0.02 (0)	0.84 (2)	0.01 (0)	10000	6	2000	0.02 (0)	0.75 (2)	0.01 (0)
2000	2	500	0.03 (0)	0.92 (1)	0.03 (0)	5000	4	1250	0.02 (0)	0.88 (2)	0.01 (0)	10000	6	2500	0.02 (0)	0.79 (2)	0.01 (0)
2000	2	667	0.04 (0)	0.95 (1)	0.04 (0)	5000	4	1667	0.02 (0)	0.91 (1)	0.02 (0)	10000	6	3333	0.02 (0)	0.84 (1)	0.01 (0)
2000	2	1000	0.05 (0)	0.97 (1)	0.06 (0)	5000	4	2500	0.03 (0)	0.95 (1)	0.03 (0)	10000	6	5000	0.02 (0)	0.90 (1)	0.01 (0)
2000	2	1333	0.06 (0)	0.98 (0)	0.08 (0)	5000	4	3333	0.03 (0)	0.96 (1)	0.04 (0)	10000	6	6667	0.02 (0)	0.93 (1)	0.02 (0)
2000	3	200	0.09 (1)	0.40 (5)	0.01 (0)	5000	5	500	0.04 (0)	0.56 (4)	0.01 (0)	10000	8	1000	0.03 (0)	0.56 (2)	0.00 (0)
2000	3	400	0.04 (1)	0.78 (4)	0.02 (0)	5000	5	1000	0.02 (0)	0.81 (2)	0.01 (0)	10000	8	2000	0.02 (0)	0.70 (2)	0.01 (0)
2000	3	500	0.04 (1)	0.85 (3)	0.02 (0)	5000	5	1250	0.02 (0)	0.86 (3)	0.01 (0)	10000	8	2500	0.02 (0)	0.74 (2)	0.01 (0)
2000	3	667	0.04 (1)	0.90 (2)	0.03 (0)	5000	5	1667	0.02 (0)	0.90 (2)	0.02 (0)	10000	8	3333	0.02 (0)	0.80 (2)	0.01 (0)
2000	3	1000	0.04 (0)	0.94 (1)	0.05 (0)	5000	5	2500	0.02 (0)	0.94 (1)	0.02 (0)	10000	8	5000	0.02 (0)	0.87 (1)	0.01 (0)
2000	3	1333	0.05 (0)	0.96 (1)	0.06 (0)	5000	5	3333	0.03 (0)	0.96 (1)	0.03 (0)	10000	8	6667	0.02 (0)	0.91 (1)	0.02 (0)

Table 4: Simulations for the generative model described in (28), while varying s and n. The results presented in Table 2 are a subset of these results. The numbers in parentheses indicate the (rounded) estimates for standard sampling error for the last printed digit, obtained by aggregating performance over 10 simulation replications. Mean-squared error (MSE) and coverage denote performance for estimating τ on a random test set; the variance column denotes the mean variance estimate obtained by the infinitesimal jackknife on the test set. Target coverage is 0.95. We always grew $B = n$ trees.

		mean-squared error			coverage		
q	d	CF	10-NN	100-NN	CF	10-NN	100-NN
2	6	0.04 (0)	0.24 (0)	0.13 (0)	0.92 (1)	0.92 (0)	0.59 (1)
4	6	0.06 (0)	0.22 (0)	0.07 (0)	0.87 (1)	0.93 (0)	0.72 (1)
6	6	0.08 (0)	0.22 (0)	0.05 (0)	0.75 (1)	0.93 (0)	0.78 (1)
2	12	0.04 (0)	0.38 (0)	0.34 (0)	0.86 (1)	0.88 (0)	0.45 (0)
4	12	0.08 (0)	0.30 (0)	0.18 (0)	0.76 (1)	0.90 (0)	0.51 (1)
6	12	0.12 (0)	0.26 (0)	0.12 (0)	0.59 (1)	0.91 (0)	0.59 (1)

Table 5: Results for a data-generating design where we vary both the number of signal features q and the number of ambient features d. All simulations have $n = 5,000$, $B = 2,000$ and a minimum leaf size of 1, and are aggregated over 20 simulation replicates.