Appendix

A Remaining Simulation Results

We include first Table 1, giving a full error comparison of the lasso-random forest baseline, BART, boosting, random forests, and local linear forests, on Friedman's data-generating process: generate $X_1, \dots, X_n$ i.i.d. $U[0,1]^5$ and model Y_i from

$$y = 10\sin(\pi X_{i1} X_{i2}) + 20(X_{i3} - 0.5)^2 + 10X_{i4} + 5X_{i5} + \epsilon,$$

Errors are reported on dimension ranging from 10 to 50, σ from 5 to 20, and $n = 1000$ and 5000, averaged over 50 training runs.

d	n	σ	RF	lasso-RF	LLF	BART	XGBoost
10	1000	5	2.33	2.12	2.03	2.49	**1.98**
10	5000	5	1.90	**1.48**	1.57	1.51	1.52
30	1000	5	2.82	2.41	**2.11**	2.60	**2.11**
30	5000	5	2.08	**1.61**	1.73	2.03	1.64
50	1000	5	3.00	2.48	**2.12**	2.84	2.20
50	5000	5	2.18	1.82	**1.80**	2.11	1.82
10	1000	20	**3.19**	3.41	3.40	6.45	6.73
10	5000	20	2.43	2.35	**2.29**	3.85	4.42
30	1000	20	4.17	3.98	**3.68**	7.60	7.03
30	5000	20	2.97	2.66	**2.40**	4.78	4.85
50	1000	20	4.25	4.45	**3.88**	8.05	7.47
50	5000	20	3.16	2.67	**2.35**	4.95	4.97

Table 1: Root mean square error on Friedman's function, with dimension d from 10 to 50 predictors in increments of 20, and consider error standard deviation σ ranging from 1 to 20, for a variety of signal-to-noise ratios. For this setting, $\text{Var}(\mathbb{E}[Y \mid X]) \approx 23.8$, as approximated over 10,000 Monte Carlo repetitions; so letting $\sigma = 1$ corresponds to a signal-to-noise ratio of about 23.8, while letting $\sigma = 20$ corresponds to a signal-to-noise ratio of about 0.24. We train on $n = 1000$ and $n = 5000$ points, and report test errors from predicting on 1000 test points. All errors reported are averaged over 50 runs and the methods are cross-validated as described in the main document. Minimizing errors are reported in bold.

We include next Table 2, again giving a more complete error comparison of the lasso-random forest baseline, BART, boosting, random forests, and local linear forests, on the data-generating process: simulate $X_1, \dots, X_n$ i.i.d. Uniform $[0,1]^{20}$, with responses

$$y_i = \log\left(1 + \exp(6X_{i1})\right) + \epsilon, \quad \epsilon \sim \mathcal{N}(0,\, 20).$$

Errors are reported on dimension ranging from 5 to 20, σ from 0.1 to 2, and $n = 1000$ and 5000, averaged over 50 training runs.

To close this section, we consider some basic linear and polynomial models in low dimensions, in order to effectively compare local linear forests with local linear regression.

d	n	σ	RF	lasso- RF	LLF	BART	XGBoost
5	1000	0.1	0.10	0.06	**0.02**	0.27	0.07
5	5000	0.1	0.06	**0.02**	**0.02**	0.22	0.06
50	1000	0.1	0.29	0.18	0.11	0.52	**0.07**
50	5000	0.1	0.18	0.10	0.07	0.62	**0.06**
5	1000	1	0.21	0.24	**0.14**	0.47	0.56
5	5000	1	0.15	0.11	**0.09**	0.26	0.52
50	1000	1	0.41	0.39	**0.20**	0.82	0.53
50	5000	1	0.23	0.21	**0.10**	0.57	0.52
5	1000	2	0.31	0.55	**0.26**	0.69	1.21
5	5000	2	0.25	0.28	**0.21**	0.40	1.18
50	1000	2	0.47	0.27	**0.24**	0.89	1.22
50	5000	2	0.33	0.27	**0.15**	0.70	0.96

Table 2: Root mean square error from simulations on random forests, lasso-random forest, local linear forests, BART, and boosting. We vary sample size n, error variance σ, and ambient dimension d, and report test error on 1000 test points. We estimate $\text{Var}[\mathbb{E}[Y \mid X]]$ as 3.52 over 10,000 Monte Carlo repetitions, so that signal-to-noise ratio ranges from 352 at $\sigma = 0.1$ to 0.88 at $\sigma = 2$. All errors are averaged over 50 runs, and minimizing errors are in bold.

We simulate $X \sim U[0,1]^3$ and model responses from two models,

$$y_i = 10X_{i1} + 5X_{i12} + X_{i3} + \epsilon \tag{1}$$

$$y_i = 10X_{i1} + 5X_{i2}^2 + X_{i3}^3 + \epsilon, \tag{2}$$

where $\epsilon \sim N(0, \sigma^2)$ and $\sigma \in \{1, 5, 10\}$. Root mean square error on the truth is reported, averaged over 50 runs, for lasso, local linear regression, BART, random forests, adaptive random forests, and local linear forests. In the simple linear case in equation 1, we see that lasso outperforms the other methods, as we would expect; in the polynomial given in equation 2, local linear regression performs the best, followed by BART ($\sigma = 1$ case) and local linear forests ($\sigma = 5, 10$ cases).

B Proof of Theorem 1

Throughout this proof, we use the notation M_λ established in (19), and shorthand $Y_i = \mu(X_i) + \epsilon_i$. Define the diameter (and corresponding radius) of a tree leaf as the length of the longest line segment that can fit completely inside of the leaf. Thanks to our assumed uniform bound on the second derivative of $\mu(\cdot)$, a Taylor expansion of around $\mu(x)$ around

Setup	σ	lasso	LLR	BART	RF	LLF
Equation 1	1	**0.12**	0.15	0.48	0.73	0.22
	5	**0.39**	0.92	1.27	1.25	0.96
	10	**0.70**	1.70	2.37	1.76	1.56
Equation 2	1	1.55	**0.22**	0.50	0.86	0.69
	5	1.55	**0.92**	1.31	1.32	1.28
	10	1.66	**1.44**	1.83	1.70	1.68

Table 3: Root Mean Square Error from simulations on equations 1 and 2 on lasso, local linear regression (LLR), BART, random forests, adaptive random forests, and local linear forests. We vary error variance σ from 1 to 10 and fix $n = 600, d = 3$. All errors are averaged over 50 runs, and minimizing errors are in bold.

x_0 yields the following decomposition starting from (5):

$$\begin{aligned}
\hat{\mu}(x_0) &= e_1^T M_\lambda^{-1} \sum_{i=1}^n \begin{pmatrix} 1 \\ X_i - x_0 \end{pmatrix} \alpha_i(x_0) Y_i = \mu(x_0) + \hat{\gamma}_n(x_0) + Q(x_0) + O\left(\bar{R}^2\right), \\
\hat{\gamma}_n(x_0) &= e_1^T M_\lambda^{-1} \sum_{i=1}^n \begin{pmatrix} 1 \\ X_i - x_0 \end{pmatrix} \alpha_i(x_0) \epsilon_i, \\
Q(x_0) &= e_1^T M_\lambda^{-1} \sum_{i=1}^n \begin{pmatrix} 1 \\ X_i - x_0 \end{pmatrix} \alpha_i(x_0) \left(\nabla \mu(x_0) \cdot (X_i - x_0)\right),
\end{aligned} \tag{3}$$

where $\bar{R}^2$ is the average squared radius of leaves T_b in the forest. In other words, we have decomposed our forest into a variance term $\hat{\gamma}_n(x_0)$, a regularization bias term $Q(x_0)$, and a curvature bias term that's bounded on the order of $\bar{R}^2$. Our main goal is to show that we can approximate $\hat{\gamma}_n(x_0)$ via an (infeasible) regression forest, while the remaining terms are lower order. For simplicity, moving forward we will write $\alpha_i(x_0) = \alpha_i$, dropping the written dependence on x_0.

Curvature bias To control the curvature bias, we need to control the radius R_{T_b} of a typical leaf containing x_0. To do so, we use the following bound. Recall that $X_1, \ldots, X_s \sim U([0,1]^d)$ independently, and that T_b is a regular, random-split tree. By Lemma 2 of Wager and Athey [2018], we then see that for any $0 < \eta < 1$ and for large enough s,

$$\mathbb{P}\left[\operatorname{diam}_j(L(x_0)) \geq \left(\frac{s}{2k-1}\right)^{-\frac{0.99(1-\eta)\log((1-\omega)^{-1})}{\log(\omega^{-1})}\frac{\pi}{d}}\right] \leq \left(\frac{s}{2k-1}\right)^{-\frac{\eta^2}{2}\frac{1}{\log(\omega^{-1})}\frac{\pi}{d}}, \tag{4}$$

where k is (fixed) the tree-depth parameter from Assumption 1. We start by applying (4) with $\eta = 0.49$, and note that $0.99(1 - 49) > 0.5$ and $0.49^2/2/\log(1/0.8) > 0.53$, meaning that for all $\omega \leq 0.2$,

$$\mathbb{P}\left(\operatorname{diam}_j(L(x_0)) \geq r_s\right) \leq r_s^{1.06}, \qquad r_s = s^{-\frac{1}{2}\frac{\log((1-\omega)^{-1})}{\log(\omega^{-1})}\frac{\pi}{d}}. \tag{5}$$

This suggests that most leaves should have radius bounded on the order of r_s. To get a useful bound on the second moment of leaf radii via $\bar{R}^2$, though, we need to use chaining: Setting $\eta = 0.71$, we find that

$$\mathbb{P}\left(\operatorname{diam}_j(L(x_0)) \geq r_s^{0.57}\right) \leq r_s^{2.2}.$$

Then, applying Markov's inequality twice, we see that $\bar{R}^2 = O_p(r_s^2)$.

Regularization bias The term $Q(x_0)$ in (3) has more intricate behavior. We note that, if we had no regularization at all, then the local linear correction would perfectly adjust for the slope of $\mu(\cdot)$ and x_0, and so we would have $Q(x_0) = 0$; unfortunately, however, we need positive regularization in other parts of the proof so we cannot directly use this fact. Conversely, as $\lambda \to \infty$, the local linear forest becomes a regression forest, and $Q(x_0)$ becomes a bias term on the order of $\bar{R}$; and this was the dominant bias term in the analysis of Wager and Athey [2018].

The derivation shows that, given a reasonable amount of regularization $0 < \lambda < \infty$, the term $Q(x_0)$ is non-zero but still much smaller than $\bar{R}$. Recall our notation Δ_i denoting a $p+1$-dimensional vector consisting of a 1 stacked with $X_i - x_0$, and let $v = (0, \nabla\mu(x_0))$. Then, writing Δ for the matrix with rows Δ_i and plugging in the expression 19 for M_λ, we see that

$$\begin{aligned} Q(x_0) &= e_1' \left(\Delta' A \Delta + \lambda J\right)^{-1} \Delta' A \Delta v \\ &= -e_1' \left(\Delta' A \Delta + \lambda J\right)^{-1} \lambda J v \\ &= -\lambda e_1' \left(\Delta' A \Delta + \lambda J\right)^{-1} v \\ &= \lambda \left(1 - d_\alpha' \left(S_\alpha + \lambda I\right)^{-1} d_\alpha\right)^{-1} d_\alpha' \left(S_\alpha + \lambda I\right)^{-1} \nabla\mu(x_0), \end{aligned}$$

where the last line followed from the Schur formula, with notation $d_\alpha = \sum_{i=1}^n \alpha_i (X_i - x_0)$ and $S_\alpha = \sum_{i=1}^n \alpha_i (X_i - x_0)^{\otimes 2}$ as used in Assumption 3. We now make some observations. First, by Assumption 3

$$\left(1 - d_\alpha' \left(S_\alpha + \lambda I\right)^{-1} d_\alpha\right)^{-1} = O_p(1)$$

is of constant order in probability. Second, by Cauchy-Schwarz,

$$\begin{aligned} d_\alpha' \left(S_\alpha + \lambda I\right)^{-1} \nabla\mu(x_0) &\leq \sqrt{d_\alpha \left(S_\alpha + \lambda I\right)^{-1} d_\alpha} \sqrt{\nabla\mu(x_0)' \left(S_\alpha + \lambda I\right)^{-1} \nabla\mu(x_0)} \\ &\leq \lambda^{-1/2} \left\lVert \nabla\mu(x_0) \right\rVert_2, \end{aligned}$$

noting that $d_\alpha' S_\alpha^{-1} d_\alpha \leq 1$ by Jensen's inequality. Combining all these facts together, we find that $Q(x_0) = O_p(\sqrt{\lambda})$.

The variance term Finally, we turn to the variance term $\hat{\gamma}_n(x_0)$. To do so, our main task is to couple $\hat{\gamma}_n$ with an approximation $\tilde{\gamma}_n$, defined as

$$\tilde{\gamma}_n(x_0) = \sum_{i=1}^{n} \alpha_i \tilde{Y}_i, \quad \text{where} \;\; \tilde{Y}_i = e_1^T \mathbb{E}[M_\lambda]^{-1} \begin{pmatrix} 1 \\ X_i - x_0 \end{pmatrix} \epsilon_i. \tag{6}$$

Now, we note that $\tilde{Y}_i$ is independent of α_i conditionally on X_i (because the problematic associations discussed at the beginning of Section 4 were mediated by M_λ), and so $\tilde{\gamma}_n(x_0)$ is just the prediction made by a "regression forest" with outcome $\tilde{Y}_i$. Consequently $\tilde{\gamma}_n$ can be characterized via standard tools used to study random forests.

We sketch out an argument below, based on the fact that M_λ concentrates around its expectation. Following the line of argumentation in Wager and Athey [2018], we see that M_λ is a U-statistic with kernel size s. Moreover, by (5), we see that the stochastic fluctuations of the terms forming M_λ are of order r_s^2. Thus, we can use concentration inequalities for U-statistics following Hoeffding [1963] to verify that (to use this concentration inequality, we need to perform several steps of chaining following (5), going up to $\eta = 0.98$)

$$\left\| M_\lambda - \mathbb{E}[M_\lambda] \right\|_\infty = O_p\left(r_s^2 \sqrt{s/n} \right). \tag{7}$$

Next, note that

$$\hat{\gamma}_n(x_0) - \tilde{\gamma}_n(x_0) = e_1 \left(M_\lambda^{-1} - \mathbb{E}[M_\lambda]^{-1} \right) \Delta' A \epsilon. \tag{8}$$

Thus, because ϵ is independent of all other terms in (8), we see that the discrepancy between $\hat{\gamma}_n(x_0)$ and $\tilde{\gamma}_n(x_0)$ is bounded on the order of $\left\| e_1 \left(M_\lambda^{-1} - \mathbb{E}[M_\lambda]^{-1} \right) \Delta' A \right\|_2$; an application of the Schur formula together with (7) then implies that

$$\hat{\gamma}_n(x_0) - \tilde{\gamma}_n(x_0) = O_p\left(\lambda^{-2} r_s^4 \, s/n \right) \tag{9}$$

for all $\lambda \gg r_s^2 \sqrt{s/n}$.

Wrapping up We are now ready to put everything together. Given everything we've seen so far, we've established that

$$\hat{\mu}(x_0) - \mu(x_0) = \tilde{\gamma}(x_0) + O_p\left(r_s^2 + \sqrt{\lambda} + \lambda^{-2} r_s^4 \frac{s}{n} \right)$$

for all $\lambda \gg r_s^2 \sqrt{s/n}$. Thus, setting $\lambda = \Theta(r_s^{1.98} \sqrt[4]{s/n})$ as in (17), we get

$$\hat{\mu}(x_0) - \mu(x_0) = \tilde{\gamma}(x_0) + O_p\left(r_s^2 + r_s^{0.99} \sqrt[8]{s/n} + r_s^{0.04} \sqrt{s/n} \right).$$

Now, recall that we have chose $s = n^\beta$ for some $\beta \geq \beta_{\min}$, meaning that

$$\sqrt[3/8]{s/n} = s^{\frac{3(1-\beta^{-1})}{8}} \geq s^{-\frac{3 \times 1.3}{8} \frac{\log((1-\omega)^{-1})}{\log(\omega^{-1})} \frac{\pi}{d}} \gg r_s^{0.99},$$

and so the above expression simplifies to

$$\hat{\mu}(x_0) - \mu(x_0) = \tilde{\gamma}(x_0) + o_p\left(\sqrt{s/n}\right). \tag{10}$$

It remains to show that $\tilde{\gamma}(x_0)$ is asymptotically centered and Gaussian with errors on the scale of $\sqrt{s/n}$, meaning that $\tilde{\gamma}(x_0)$ is in fact the dominant error term in $\hat{\mu}(x_0)$.

But now, recall that $\tilde{\gamma}(x_0)$ is simply a regression forest with outcome $\tilde{Y}_i$. Thus, Theorem 8 of Wager and Athey [2018] directly implies that there is sequence $\sigma_n(x_0) \to 0$ such that

$$\frac{\tilde{\gamma}_n(x_0)}{\sigma_n(x_0)} \Rightarrow \mathcal{N}(0, 1); \tag{11}$$

here, we used the fact that the ϵ_i are all mean-zero conditionally on the tree construction, and so $\mathbb{E}[\tilde{\gamma}(x_0)] = 0$. Finally, from Theorem 5 of Wager and Athey [2018], we see that $\sigma_n(x_0) = \sqrt{s/n}\,\text{polylog}(s)$, and we note that our above argument in fact established a polynomial gap between the error term in (10) and $\sqrt{s/n}$. Thus (11) in fact captures the dominant error term of our estimator.

References

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, 58(301):13–30, 1963.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018.