On Semi-parametric Inference with BART (Appendix)

1 Appendix

1.1 Proof of Theorem 4.1theorem.4.1

First, we show that with K large enough $\widehat{\Psi}_K$ will satisfy

$$\sqrt{n}|\Psi_n - \widehat{\Psi}_K| = o_P(1). \tag{1.1}$$

We can write

$$\begin{aligned}\sqrt{n}|\Psi_n - \widehat{\Psi}_K| &= \sqrt{n}|\Psi(f_0) - \Psi(f_0^K) + W_n(a - a^K)| \\ &= \sqrt{n}\left|\frac{1}{n}\sum_{i=1}^n \left[f_0(\boldsymbol{x}_i) - f_0^K(\boldsymbol{x}_i) + \varepsilon_i\right]\left[a(\boldsymbol{x}_i) - a^K(\boldsymbol{x}_i)\right]\right|,\end{aligned}$$

where we used the fact that

$$\sum_{i=1}^n [f_0(\boldsymbol{x}_i) - f_0^K(\boldsymbol{x}_i)]a^K(\boldsymbol{x}_i) = \sum_{k=1}^K \frac{a_k^K}{n}\sum_{\boldsymbol{x}_i\in\Omega_k}\left[f_0(\boldsymbol{x}_i) - \frac{n}{\mu(\Omega_k)}\sum_{\boldsymbol{x}_i\in\Omega_k} f_0(\boldsymbol{x}_i)\right] = 0.$$

Next, we can write

$$\sqrt{n}|\Psi_n - \widehat{\Psi}_K| < \left[\sqrt{n}\|a - a^K\|_L \times \|f_0 - f_0^K\|_L + Z_n^K\right],$$

where

$$Z_n^K = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i[a(\boldsymbol{x}_i) - a^K(\boldsymbol{x}_i)].$$

We assume that the design is regular (according to Definition 3.3 of Rockova and van der Pas (2017) (further referred to as RP17) with $p = q = 1$). Assuming that $a \in \mathcal{H}^\gamma$, it follows from the proof of Lemma 3.2 of Rockova and van der Pas (2017) that

$$\|f_0 - f_0^K\|_L \leq \|f_0\|_{\mathcal{H}^\alpha} C_1/K^\alpha \quad \text{and} \quad \big|\|a\|_L - \|a^K\|_L\big| \leq \|a - a^K\|_L \leq \|a\|_{\mathcal{H}^\gamma} C_2/K^\gamma.$$

We assume that $\|a\|_L^2 \leq \|a\|_\infty^2 < C_a^2$ for some $C_a > 0$ and $\|f_0\|_\infty < C_f$ for some $C_f > 0$. Because

$$\mathsf{Var}\, Z_n^K = \|a - a^K\|_L^2 = \frac{1}{n}\sum_{i=1}^n [a(\boldsymbol{x}_i) - a^K(\boldsymbol{x}_i)]^2 \lesssim 1/K^{2\gamma},$$

we conclude that $Z_n^K = o_P(1)$ when $K \to \infty$ as $n \to \infty$ and thereby

$$\sqrt{n}|\Psi_n - \widehat{\Psi}_K| \lesssim \sqrt{n}K^{-(\alpha+\gamma)} + o_P(1). \tag{1.2}$$

With the choice $K = K_n = \lfloor (n/\log n)^{1/(2\alpha+1)} \rfloor$ for $\alpha > 1/2$ and $\gamma > 1/2$, (1.2) will be satisfied.

To continue, we introduce the following notation

$$f_t^K = f^K - \frac{t\, a^K}{\sqrt{n}}$$

and write

$$\begin{aligned}
&\ell_n(f_t^K) - \ell_n(f_0^K) - [\ell_n(f^K) - \ell_n(f_0^K)] \\
&\quad = -\frac{n}{2}[\|f_t^K - f_0^K\|_L^2 - \|f^K - f_0^K\|_L^2] + \sqrt{n}\, W_n(f_t^K - f^K) \\
&\quad = -\frac{n}{2}[\|f_t^K - f^K\|_L^2 + 2\langle f_t^K - f^K, f^K - f_0^K\rangle_L] + \sqrt{n}\, W_n(f_t^K - f^K) \\
&\quad = -\frac{t^2}{2}\|a^K\|_L^2 + \sqrt{n}\, t\langle a^K, f^K - f_0^K\rangle_L - t\, W_n(a^K)
\end{aligned}$$

Then we have

$$t\sqrt{n}(\Psi(f^K) - \widehat{\Psi}_K) = t\sqrt{n}(\Psi(f^K) - \Psi(f_0^K)) - t\, W_n(a^K)$$

and thereby, using the fact

$$t\sqrt{n}(\Psi(f^K) - \Psi(f_0^K)) = t\sqrt{n}\langle a^K, f^K - f_0^K\rangle_L,$$

we can write

$$\begin{aligned}
&t\sqrt{n}(\Psi(f^K) - \widehat{\Psi}_K) + \ell_n(f^K) - \ell_n(f_0^K) \\
&\quad = \ell_n(f_t^K) - \ell_n(f_0^K) + \frac{t^2}{2}\|a^K\|_L^2 - \sqrt{n}\, t\langle a^K, f^K - f_0^K\rangle_L + t\sqrt{n}\langle a^K, f^K - f_0^K\rangle_L \\
&\quad = \ell_n(f_t^K) - \ell_n(f_0^K) + \frac{t^2}{2}\|a^K\|_L^2.
\end{aligned}$$

Given sets $A_{n,K} \subset \mathcal{F}[K]$ (to be defined later) such that $\Pi(A_{n,K}^C \,|\, \boldsymbol{Y}^{(n)}) \to 0$ in $\mathbb{P}_0^n$ probability, we define

$$I_{n,K} = \mathbb{E}^\Pi[\mathrm{e}^{t\sqrt{n}(\Psi(f^K) - \widehat{\Psi}_K)} \,|\, \boldsymbol{Y}^{(n)}, A_{n,K}] \tag{1.3}$$

and, using the calculations above, we write

$$I_{n,K} = \mathrm{e}^{\frac{t^2}{2}\|a^K\|_L^2} \times \frac{\int_{A_{n,K}} \mathrm{e}^{\ell_n(f_t^K) - \ell_n(f_0^K)} \mathrm{d}\Pi^K(f^K)}{\int_{A_{n,K}} \mathrm{e}^{\ell_n(f^K) - \ell_n(f_0^K)} \mathrm{d}\Pi^K(f^K)}. \tag{1.4}$$

For $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)' \in \mathbb{R}^K$, define $\beta_k^t = \beta_k - \frac{ta_k^K}{\sqrt{n}}$ and denote $\boldsymbol{\beta}^t = (\beta_1^t, \ldots, \beta_K^t) \in \mathbb{R}^K$ and $\boldsymbol{a}^K = (a_1^K, \ldots, a_K^K)' \in \mathbb{R}^K$. Then we have

$$f^K(\boldsymbol{x}) = \sum_{k=1}^K \mathbb{I}(\boldsymbol{x} \in \Omega_k)\beta_k \quad \text{and} \quad f_t^K(\boldsymbol{x}) = \sum_{k=1}^K \mathbb{I}(\boldsymbol{x} \in \Omega_k)\beta_k^t.$$

Assuming the multivariate Gaussian prior $\pi(\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi|\Sigma|}}\mathrm{e}^{-\frac{1}{2}\boldsymbol{\beta}'\Sigma^{-1}\boldsymbol{\beta}}$ centered at zero with a covariance matrix Σ, we can write

$$\pi(\boldsymbol{\beta}) = \pi(\boldsymbol{\beta}^t)\mathrm{e}^{\frac{t^2}{2n}\boldsymbol{a}^{K\prime}\Sigma^{-1}\boldsymbol{a}^K - \frac{t}{\sqrt{n}}\boldsymbol{a}^{K\prime}\Sigma^{-1}\boldsymbol{\beta}}. \tag{1.5}$$

Next, for $\widetilde{\varepsilon}_{n,K} = \sqrt{\frac{K\,\log(n)}{n}}$ and $M > 0$, we define

$$A_{n,K}(M) = \left\{ f^K \in \mathcal{F}[K] : \|f^K - f_0^K\|_L \leq M\,\widetilde{\varepsilon}_{n,K} \right\}. \tag{1.6}$$

We show (Section 1.1.2) that $\Pi(A_{n,K}^C(M) \,|\, \boldsymbol{Y}^{(n)}) \to 0$ in $\mathbb{P}_0^n$-probability as $n \to \infty$ for some suitably large $M > 0$. We note that

$$A_{n,K}(M) \subset \{\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0^K\|_2^2 \leq M^2\, n\, \widetilde{\varepsilon}_{n,K}^2\},$$

where $\boldsymbol{\beta}_0^K \in \mathbb{R}^K$ are coefficients of the projection of f_0 onto $\mathcal{F}[K]$. On the set $A_{n,K}(M)$, we can thus write

$$\begin{aligned}
\frac{t}{\sqrt{n}}\left|\boldsymbol{a}^{K\prime}\Sigma^{-1}\boldsymbol{\beta}\right| &< \frac{t}{\sqrt{n}\,\lambda_{min}}\|a^K\|_2\sqrt{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0^K\|_2^2 + \|\boldsymbol{\beta}_0^K\|_2^2} \\
&< \frac{t}{\sqrt{n}\,\lambda_{min}}\|a^K\|_2\sqrt{K\log n + KC_f^2} \\
&< \frac{t\,K\,C_a}{\sqrt{n}\,\lambda_{min}}\sqrt{\log n + C_f^2},
\end{aligned} \tag{1.7}$$

where λ_{min} is the smallest singular value of Σ and where we used the fact that $\|a\|_\infty < C_a$ and $\|f_0\|_\infty < C_f$. Using (1.7) and (1.5), we have

$$\mathrm{e}^{\frac{t^2}{2}\|a\|_L^2 - \frac{t\,K\,C_a}{\lambda_{min}\sqrt{n}}\sqrt{\log n + C_f^2}} < I_{n,K} < \mathrm{e}^{\frac{t^2}{2}\|a\|_L^2 + \frac{t^2\,K\,C_a^2}{2\,n\,\lambda_{min}} + \frac{t\,K\,C_a}{\lambda_{min}\sqrt{n}}\sqrt{\log n + C_f^2}}.$$

If $\lambda_{min} > c$ for some c and $K\sqrt{(\log n)/n} \to 0$, we have $\lim_{n\to\infty} I_{n,K} = \mathrm{e}^{t^2\|a\|_L^2}$ for each $t \in \mathbb{R}$. This is satisfied if we set, for instance, $K = K_n = \lfloor (n/\log n)^{1/(2\alpha+1)} \rfloor$ with $\alpha > 1/2$. This concludes the proof of the BvM property for a fixed K and a single partition.

1.1.1 The Laplace Prior

For the Laplace prior, we use the reverse triangle inequality $|\beta_k| > |\beta_k^t| - |\frac{ta_k^K}{\sqrt{n}}|$ to find that

$$\pi(\boldsymbol{\beta}^K) = \prod_{k=1}^K \psi(\beta_k;\lambda) < \prod_{k=1}^K \psi(\beta_k^t;\lambda)\mathrm{e}^{\frac{t\lambda|a_k^K|}{\sqrt{n}}} = \pi(\boldsymbol{\beta}_t^K)\mathrm{e}^{\frac{t\lambda}{\sqrt{n}}\|\boldsymbol{a}^K\|_1} < \pi(\boldsymbol{\beta}_t^K)\mathrm{e}^{\frac{t\lambda}{\sqrt{n}}KC_a}.$$

Then we find that for $K = K_n$

$$\mathrm{e}^{\frac{t^2}{2}(\|a\|_L^2+o(1))} \times \mathrm{e}^{-\frac{t\lambda}{\sqrt{n}} K_n(\alpha) C_a} < I_{n,K} < \mathrm{e}^{\frac{t^2}{2}(\|a\|_L^2+o(1))} \times \mathrm{e}^{\frac{t\lambda}{\sqrt{n}} K_n(\alpha) C_a}.$$

Since $K_n(\alpha) = \lfloor (n/\log n)^{1/(2\alpha+1)} \rfloor$, we have for $\alpha > 1/2$ and $t \in \mathbb{R}$.

$$\lim_{n\to\infty} \mathbb{E}^{\Pi}[\mathrm{e}^{t\sqrt{n}(\Psi(f^K)-\Psi_n)} \mid \boldsymbol{Y}^{(n)}] = \mathrm{e}^{\frac{t^2}{2}\|a\|_L^2}.$$

It is interesting to note that with the Laplace prior, one can obviate the proof of posterior concentration around the projections, which was needed for the Gaussian case.

1.1.2 The Set $A_{n,K}$

We want to show that the posterior distribution concentrates around f_0^K, the projection of f_0 onto $\mathcal{F}[K]$, at the following contraction rate

$$\widetilde{\varepsilon}_{n,K} = \sqrt{\frac{K\log(n)}{n}}.$$

For $K \leq n/\log(n)$ and $A_{n,K}(M)$ defined in (1.6), we show that

$$\lim_{n\to\infty} \Pi(A_{n,K}^C(M) \mid \boldsymbol{Y}^{(n)}) = 0 \quad \text{in } \mathbb{P}_0^n\text{-probability} \tag{1.8}$$

for some sufficiently constant $M > 0$. We show this statement by verifying conditions (2.4) and (2.5) of Theorem 2.1 of Kelijn and van der Vaart (2006). We start with the entropy condition (2.5). In our model, the covering number for testing under misspecification can be bounded by the classical local entropy (according to Lemma 2.1 by Kelijn and van der Vaart (2006). It follows from Section 8.1 of Rockova and van der Pas (2017) that the local entropy satisfies

$$N\Big(\tfrac{\varepsilon}{36}, \Big\{f^K \in \mathcal{F}[K] : \|f^K - f_0^K\|_L < \varepsilon\Big\}, \|.\|_L\Big) \leq \left(\frac{108}{\bar{C}}\sqrt{n}\right)^K,$$

where $\bar{C}$ is such that $\mu(\Omega_k) > \bar{C}/n$. The entropy condition (2.5) will be met since

$$K\log n \lesssim n\,\widetilde{\varepsilon}_{n,K}^2.$$

Regarding the prior concentration condition (2.4), we note (similarly as in Section 8.2 of Rockova and van der Pas (2017)) that

$$\big\{f^K \in \mathcal{F}[K] : \|f^K - f_0^K\|_L \leq M\,\widetilde{\varepsilon}_{n,K}\big\} \supset \{\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0^K\|_2 \leq M\,\widetilde{\varepsilon}_{n,K}\}$$

With the Gaussian prior $\boldsymbol{\beta} \mid K \sim \mathcal{N}_K(0, \Sigma)$, we have

$$\Pi\left(\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0^K\|_2 \leq M\,\widetilde{\varepsilon}_{n,K}\right) \geq \Pi\left(\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^K : \|\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_0^K\|_2 \leq M\,\frac{\widetilde{\varepsilon}_{n,K}}{\sqrt{\lambda_{max}}}\right),$$

where λ_{max} is the maximal eigenvalue of Σ and where $\widetilde{\boldsymbol{\beta}}_0^K = \Sigma^{-1/2}\boldsymbol{\beta}_0^K$ and $\widetilde{\boldsymbol{\beta}} = \Sigma^{-1/2}\boldsymbol{\beta} \sim \mathcal{N}_K(0, \mathrm{I}_K)$. The right-hand side can be further lower-bounded with

$$\frac{2^{-K}\mathrm{e}^{-\|\widetilde{\boldsymbol{\beta}}_0^K\|_2^2 - M^2\widetilde{\varepsilon}_{n,K}^2/(4\lambda_{max})}}{\Gamma\left(\frac{K}{2}\right)\left(\frac{K}{2}\right)}\left(\frac{M\widetilde{\varepsilon}_{n,K}}{\sqrt{\lambda_{max}}}\right)^K.$$

With λ_{min} denoting the minimal eigenvalue of Σ, we obtain the following lower bound for the above:

$$\mathrm{e}^{-C_1\,K\log\left(C_2\,K\,\lambda_{max}/\widetilde{\varepsilon}_{n,K}\right) - K\|f_0\|_\infty^2/\lambda_{min} - C_3\widetilde{\varepsilon}_{n,K}^2/\lambda_{max}}$$

With $\lambda_{min} > c$ for some $c > 0$, $\lambda_{max} \lesssim n$ and $\|f_0\|_\infty \leq \log^{1/2}(n)$, the above is bounded from below by $\mathrm{e}^{-D\,K\log(n)}$ for some suitable $D > 0$. It then follows from Theorem 2.1 of Kleijn and van der Vaart (2006) that (1.8) holds.

1.2 Posterior Concentration for the Laplace Prior

Ročková and van der Pas (2017) and Ročková and Saha (2019) show posterior concentration for BART under (a) the conditionally uniform prior (8equation.3.8) and (9equation.3.9) and (b) the Galton Watson Process prior (10equation.3.10). Both of these results apply for Gaussian step heights. Here, we formally show that the Laplace prior gives rise to optimal posterior concentration as well.

Theorem 1.1. *Assume $f_0 \in \mathcal{H}_p^\alpha$ with $0 < \alpha \leq 1$ where $p \lesssim \log^{1/2} n$ and $\|f_0\|_\infty \lesssim \log^{1/2} n$. Moreover, we assume that $\mathcal{X} \equiv \{\boldsymbol{x}_i\}_{i=1}^n$ is regular. We assume priors* (8equation.3.8) *and* (9equation.3.9) *or the Galton Watson process* (10equation.3.10) *and Laplace step heights* (12equation.3.12) *where $1/\lambda \lesssim \sqrt{K}$. Then with $\varepsilon_n = n^{-\alpha/(2\alpha+p)}\log^{1/2} n$ we have*

$$\Pi\left(f_{\mathcal{T},\boldsymbol{\beta}} : \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_n > M_n\,\varepsilon_n \mid \boldsymbol{Y}^{(n)}\right) \to 0,$$

for any $M_n \to \infty$ in $\mathbb{P}_0^n$-probability, as $n, p \to \infty$.

Proof. It suffices to show the prior concentration condition (2.2) in Rockova and van der Pas (2017), i.e.

$$\Pi(f_{\mathcal{T},\boldsymbol{\beta}} : \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0\|_n \leq \varepsilon_n) \geq \mathrm{e}^{-d\,n\,\varepsilon_n^2}$$

for some $d > 2$. Using similar considerations as in Section 8.2. of RP17, this boils down to showing that

$$\Pi(\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_2 \leq \varepsilon_n/2) \geq \Pi(\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_1 \leq \varepsilon_n/2),$$

where $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^K$ are the step heights of the projection of f_0 onto a partition supported by the k-d tree with K steps, where $K \lesssim n\varepsilon_n^2/\log n$. The right hand side equals

$$\int_{|\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}|_1 \leq \varepsilon_n/2}\left(\frac{\lambda}{2}\right)^K \mathrm{e}^{-\lambda|\boldsymbol{\beta}|_1}\mathrm{d}\boldsymbol{\beta} > \mathrm{e}^{-\lambda|\widehat{\boldsymbol{\beta}}|_1}\int_{|\boldsymbol{\beta}|_1 \leq \varepsilon_n/2}\left(\frac{\lambda}{2}\right)^K \mathrm{e}^{-\lambda|\boldsymbol{\beta}|_1}\mathrm{d}\boldsymbol{\beta}$$

$$> \mathrm{e}^{-\lambda|\widehat{\boldsymbol{\beta}}|_1}\left(\frac{\varepsilon_n\lambda}{2K}\right)^K \mathrm{e}^{-\lambda\varepsilon_n/2}.$$

Assuming that $\|f_0\|_\infty \lesssim \log n$, we have $|\widehat{\boldsymbol{\beta}}|_1 \leq K \log n \lesssim n\varepsilon_n^2$. Next, for $1/\lambda \lesssim \sqrt{K}$ we have $K \log[4K/(\lambda\varepsilon_n^2)] \lesssim K \log n \lesssim n\varepsilon_n^2$. □

1.3 Proof of Theorem 5.1theorem.5.1

According to Lemma 5.1lemma.5.1, the posterior concentrates on the set $\mathcal{R}(\mathcal{K}_n)$. All the following arguments will be thus conditional on $\mathcal{R}(\mathcal{K}_n)$. The conditional posterior decomposes into a mixture of laws with weights $\pi(\mathcal{T} \mid \boldsymbol{Y}^{(n)}, \mathcal{R}(\mathcal{K}_n))$ in the sense that

$$\begin{aligned}
&\mathbb{E}^\Pi[\mathrm{e}^{t\sqrt{n}(\Psi(f_{\mathcal{T},\boldsymbol{\beta}})-\Psi_n)} \mid \boldsymbol{Y}^{(n)}, \mathcal{R}(\mathcal{K}_n)] \\
&\quad= \sum_{\mathcal{T}\in\mathcal{R}(\mathcal{K}_n)} \pi[\mathcal{T} \mid \boldsymbol{Y}^{(n)}, \mathcal{R}(\mathcal{K}_n)] \mathbb{E}^\Pi[\mathrm{e}^{t\sqrt{n}(\Psi(f_{\mathcal{T},\boldsymbol{\beta}})-\Psi_n)} \mid \boldsymbol{Y}^{(n)}, \mathcal{T}] \\
&\quad= \mathrm{e}^{t\times o_P(1)} \sum_{\mathcal{T}\in\mathcal{R}(\mathcal{K}_n)} \pi[\mathcal{T} \mid \boldsymbol{Y}^{(n)}, \mathcal{R}(\mathcal{K}_n)] I_{n,\mathcal{T}}
\end{aligned}$$

where

$$I_{n,\mathcal{T}} = \mathbb{E}^\Pi\left[\mathrm{e}^{t\sqrt{n}(\Psi(f_{\mathcal{T},\boldsymbol{\beta}})-\widehat{\Psi}_{\mathcal{T}})} \middle| \boldsymbol{Y}^{(n)}, \mathcal{T}\right]$$

and where we used the fact that $\sqrt{n}|\widehat{\Psi}_{\mathcal{T}} - \Psi_n| = o_P(1)$ under the assumption of self-similarity (as we showed earlier in the Section 5.2subsection.5.2). Under the Laplace prior (12equation.3.12), we can write for $\mathcal{T} \in \mathcal{R}(\mathcal{K}_n)$

$$\mathrm{e}^{\frac{t^2}{2}(\|a\|_L^2+o(1))} \times \mathrm{e}^{-\frac{t\lambda}{\sqrt{n}}K_n C_a} < I_{n,\mathcal{T}} < \mathrm{e}^{\frac{t^2}{2}(\|a\|_L^2+o(1))} \times \mathrm{e}^{\frac{t\lambda}{\sqrt{n}}K_n C_a}.$$

Since $K_n = \lfloor M_2 n^{1/(2\alpha+1)} \rfloor$, we have for $1/2 < \alpha \leq 1$ and for all $t \in \mathbb{R}$

$$\lim_{n\to\infty} \mathbb{E}^\Pi[\mathrm{e}^{t\sqrt{n}(\Psi(f_{\mathcal{T},\boldsymbol{\beta}})-\Psi_n)} \mid \boldsymbol{Y}^{(n)}, \mathcal{R}(\mathcal{K}_n)] = \mathrm{e}^{\frac{t^2}{2}\|a\|_L^2}.$$

For the Gaussian prior, one can proceed analogously as before. For each $\mathcal{T} \in \mathcal{R}(\mathcal{K}_n) \cap \mathcal{V}^K$, we denote with $A_{n,\mathcal{T}}(M) = \{f_{\mathcal{T},\boldsymbol{\beta}} \in \mathcal{F}[\mathcal{T}] : \|f_{\mathcal{T},\boldsymbol{\beta}} - f_0^{\mathcal{T}}\|_L \leq M\sqrt{K \log n/n}\}$. Using the same arguments as in Section 1.1.2, one can show that, given $\mathcal{T} \in \mathcal{R}(\mathcal{K}_n)$, the posterior concentrates on $A_{n,\mathcal{T}}(M)$. We then define $I_{n,\mathcal{T}} = \mathbb{E}^\Pi\left[\mathrm{e}^{t\sqrt{n}(\Psi(f_{\mathcal{T},\boldsymbol{\beta}})-\widehat{\Psi}_{\mathcal{T}})} \middle| \boldsymbol{Y}^{(n)}, \mathcal{T}, A_{n,\mathcal{T}}(M)\right]$ and using the same arguments show that $\lim_{n\to\infty} I_{n,\mathcal{T}} = \mathrm{e}^{\frac{t^2}{2}\|a\|_L^2}$ uniformly for all $\mathcal{T} \in \mathcal{R}(K_n)$.

1.4 Proof of Theorem 5.2theorem.5.2

Let $a_{\mathcal{E}}$ denote the projection of a onto $\mathcal{F}[\mathcal{E}]$ (the set of all forest mappings (7equation.3.7) supported on a given ensemble $\mathcal{E}$). The no-bias condition (16equation.4.16) is satisfied automatically since $a_{\mathcal{E}} = a$.

Similarly as in Rockova and van der Pas (2017) (Corollary 5.1), one can show that the posterior concentrates on $\mathcal{E}$, whose trees are not too big (i.e. $\sum_t K^t \lesssim n\varepsilon_n^2/\log n$ for

$\varepsilon_n = n^{-\alpha/(2\alpha+p)} \log n$). The next step is to show that the prior is sufficiently diffuse in the sense that it does not change much under a small perturbation. To this end, we introduce a local shift, for some $s > 0$,

$$f_{\mathcal{E},\mathcal{B}_s} = \sum_{t=1}^{T} \left(f_{\mathcal{T}^t,\boldsymbol{\beta}^t} - \frac{s}{T\sqrt{n}} \right) = f_{\mathcal{E},\mathcal{B}} - \frac{s}{\sqrt{n}},$$

where $\mathcal{B} = [\boldsymbol{\beta}_s^1, \dots, \boldsymbol{\beta}_s^T]$ and where $\boldsymbol{\beta}_s^t = \boldsymbol{\beta}^t - s/(T\sqrt{n})$. Now we have to perform the change of measures from $f_{\mathcal{E},\mathcal{B}}$ to $f_{\mathcal{E},\mathcal{B}_s}$. We start with the independent Laplace prior (12equation.3.12) for each $\boldsymbol{\beta}^t$ with a penalty $\lambda_t \asymp 1/\sqrt{K^t}$. Similarly as in Section 1.1.1, we have

$$\prod_{t=1}^{T} \pi(\boldsymbol{\beta}^t) < \prod_{t=1}^{T} \pi(\boldsymbol{\beta}_s^t) \times \exp\left(\frac{s}{T\sqrt{n}} \sum_{t=1}^{T} \lambda_t K^t \right) < \prod_{t=1}^{T} \pi(\boldsymbol{\beta}_s^t) \times \exp\left(\frac{s\, c_1}{\sqrt{n}} \max_t \sqrt{K^t} \right)$$

for some $c_1 > 0$. Since $\max_t K^t \lesssim n\varepsilon_n^2$, the exponential term converges to one. A similar argument holds also for the lower bound. For the Gaussian prior $\boldsymbol{\beta}^t \,|\, K^t \sim \mathcal{N}_{K^t}(0, K^t \times \mathrm{I}_{K^t})$, we have

$$\prod_{t=1}^{T} \pi(\boldsymbol{\beta}^t) = \prod_{t=1}^{T} \pi(\boldsymbol{\beta}_s^t) \times \exp\left(\frac{s^2}{2nT^2} + \frac{s\|\mathcal{B}\|_2}{\sqrt{n}T} \right),$$

where $\mathcal{B} = (\boldsymbol{\beta}^{1\prime}, \dots, \boldsymbol{\beta}^{T\prime}) \in \mathbb{R}^{\sum_t K^t}$ is the vector of all step heights in the ensemble. Similarly as before, one can show that the conditional posterior distribution of $\mathcal{B}$, given $\mathcal{E}$, concentrates around the projection of f_0 onto $\mathcal{F}[\mathcal{E}]$ at the rate $\sum_t K^t \log n$. Since $\sqrt{\sum_t K^t \log n/n} \lesssim \sqrt{\varepsilon_n^2 \log n} \to 0$, we can use similar arguments as in Section **??** to conclude the BvM property.

1.5 Proof of Lemma 1

Proof. With $\varepsilon_n = n^{-\alpha/(2\alpha+p)}\sqrt{\log n}$, the assumption (18equation.5.18) implies $\|f_0^{\mathcal{T}} - f_0\|_L > M_n/M_3\, \varepsilon_n$ when $\mathrm{diam}(\mathcal{T}) > d_n(\alpha)$, where $f_0^{\mathcal{T}}$ is the $\|\cdot\|_L$ projection of f_0 onto $\mathcal{F}[\mathcal{T}]$. The posterior distribution under the Bayesian CART prior concentrates at the rate ε_n in the $\|\cdot\|_L$ sense, i.e. $\Pi(\|f - f_0\|_L > M_n\varepsilon_n \,|\, \boldsymbol{Y}^{(n)}) \to 0$ in $\mathbb{P}_0^n$-probability for any arbitrarily slowly increasing sequence M_n. This follows from Ročková and van der Pas (2017) for the conditionally uniform tree partition prior and from Ročková and Saha (2019) for the tree-branching process prior. Both of these papers study Gaussian step heights. In the Appendix (Section 1.2), we extend these results to Laplace step heights. From the near-minimaxity of the posterior, it then follows that partitions that are *not* regular are *not* supported by the posterior. The dimensionality part regarding K follows from Ročková and van der Pas (2017) and Ročková and Saha (2019). □

References

Kleijn, B. and A. van der Vaart (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics 34*, 837–877.

Ročková, V. and E. Saha (2019). On theory for BART. 22^{nd} *International Conference on Artificial Intelligence and Statistics*.

Ročková, V. and S. van der Pas (2017+). Posterior concentration for Bayesian regression trees and forests. *Annals of Statistics (In Revision)*, 1–40.