Supplementary file for “FAST: a Fused and Accurate Shrinkage Tree for Heterogeneous Treatment Effects Estimation”

S1 Additional figures

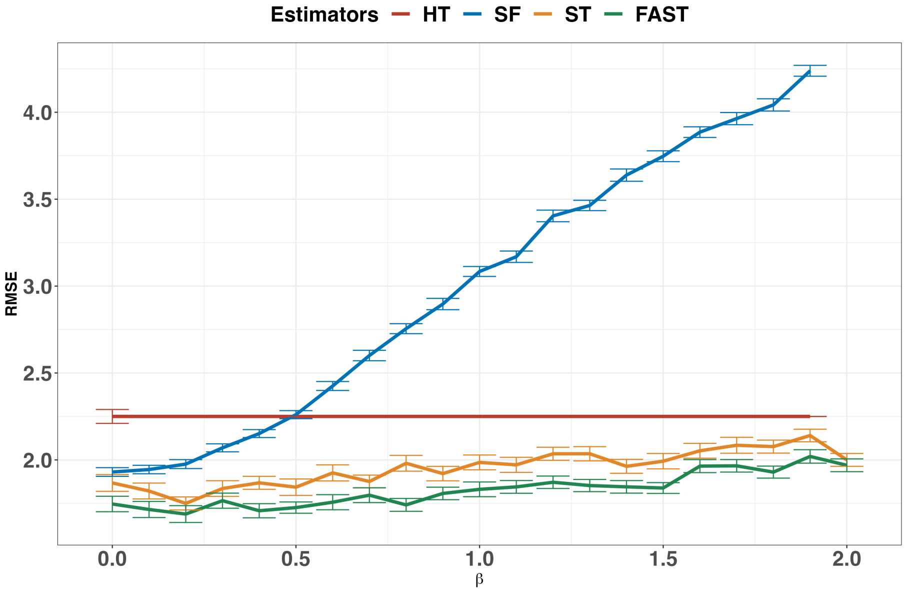

Figure S1: The averaged root mean square error (RMSE) (mean with 2×s.d. error bars) of each algorithm on multiple simulation datasets with different levels of the confounding bias parameter β.

S2 Pre-processing of the real-world data

In the STAR dataset, each of the pre-treatment covariate X_j $(1 \leq j \leq p)$ was standardized to a range of -1 to 1, and the outcome variable Y was standardized to a range of 0 to 100.

S3 Proof of Theorem 1

The proof follows the similar arguments as in Györfi et al. [2002] and Scornet et al. [2015]. It is sufficient to show the result at the root node given the recursive nature of the partitioning. We will use the following notations in the sequel. We denote $\mathbb{E}_T, \mathbb{P}_T$ and $\mathbb{E}_O, \mathbb{P}_O$ as the expectation and probability under trial data and observational data, respectively. We let $Z = (X, \tilde{Y})$. For any $q \in [p]$ and $c \in \mathbb{R}$, let $\theta = (q, c)$ and the corresponding two partitioned notes are denoted as $Q_L(\theta) = \{x | x_q \leq c\}$ and $Q_R(\theta) = \{x | x_q > c\}$. The parameter space of θ is denoted as $\Theta = [p] \times \mathbb{R}$.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Let μ_L and μ_R be the predictions for Y on $Q_L(\theta)$ and $Q_R(\theta)$, respectively and denote $\tau = (\tau_L, \tau_R)$. Let $M^i(\theta)$ and $M^i_{of}(\theta)$ be the MSEs of the conventional HTE estimator and the fused HTE estimator on the child nodes of $Q_i(\theta)$, respectively, for $i \in \{R, L\}$.

(i).

For any $\theta \in \Theta$, according to Equation (2) in the main paper we have

$$M^i_{of}(\theta) = (1 - w_i(\theta))M^i(\theta),$$

for $i \in \{R, L\}$, where the weight $w_i(\theta)$ satisfies

$$w_i(\theta) \asymp \frac{\sigma^2_u(Q_i)/n}{\sigma^2_u(Q_i)/n + b^2(\theta)},$$

by Equation (5) in the main paper, which is lower bounded by $\frac{\sigma^2_{\min}}{\sigma^2_{\min}+nb^2_{\max}}$, where $\sigma^2_{\min} < \mathrm{Var}(\tilde{Y}|\boldsymbol{X} = \boldsymbol{x}, S = 0)$ and $b_{\max} = \sup_{\boldsymbol{x}\in Q_j} |\{\mathbb{E}(\tilde{Y}|\boldsymbol{X} = \boldsymbol{x}, S = 0) - \mathbb{E}(\tilde{Y}|\boldsymbol{X} = \boldsymbol{x}, S = 1)\}|$ Therefore, we conclude that

$$\frac{M_{of}(\theta)}{M(\theta)} - 1 \leq -\frac{\sigma^2_{\min}}{\sigma^2_{\min} + nb^2_{\max}},$$

which reveals the MSE reduction effect of the proposed split criterion.

(ii). The proof includes two parts. In Part 1, we will derive the bounds for the discrepancies between the MSEs under the empirically estimated split and the oracle split under the conventional criterion, and in Part 2 the similar results under the proposed split criterion.

Part 1. We define the following criterion function:

$$\begin{aligned}\ell_n(\theta, \tau, \mathcal{R}^t_n) &= \frac{1}{n}\sum_{i=1}^{n}\Big\{(\tilde{Y}_{0,i} - \tau_L)^2 I\{X_{0,i} \in Q_L(\theta)\} + (\tilde{Y}_{0,i} - \tau_R)^2 I\{X_{0,i} \in Q_R(\theta)\}\Big\} \\ &=: \ell^L_n(\theta, \tau_L, \mathcal{R}^t_n) + \ell^R_n(\theta, \tau_R, \mathcal{R}^t_n).\end{aligned}$$

For $i \in \{L, R\}$, let

$$\mathcal{L}^i(\theta, \tau_i) = \mathbb{E}_T\left\{\ell^i_n(\theta, \tau_i, \mathcal{R}^t_n)\right\} \text{ and } \mathcal{L}(\theta, \tau) = \mathcal{L}^L_n(\theta, \tau_L) + \mathcal{L}^R_n(\theta, \tau_R) \tag{S1}$$

Then $\mathcal{L}^i(\theta, \tau_i)$ represents the MSE of τ_i on the region $Q_i(\theta)$. For a given split $\theta = (q, c)$, it is straightforward to see that the optimal $\tau(\theta) = (\tau_L(\theta), \tau_R(\theta))$ is given by

$$\tau_i(\theta) = \underset{\tau_i\in\mathbb{R}}{\arg\min}\, \ell^i_n(\theta, \tau_i, \mathcal{R}^t_n) = \mathbb{E}_n\left\{\tilde{Y}_0|X_0 \in Q_i(\theta)\right\}$$

for $i \in \{L, R\}$, which is the sample mean of Y on the region $Q_i(\theta)$. Therefore, by the definition of $M^i(\theta)$, it holds that $\mathcal{L}^i(\theta, \tau_i(\theta)) = M^i(\theta)$ for $i \in \{L, R\}$. The optimal split $\theta_0 = (q_0, c_0)$ on the population level is defined via minimizing the profiled criterion function:

$$(q_0, c_0) = \underset{q\in[p],c\in\mathbb{R}}{\arg\min}\left\{M^L(\theta) + M^R(\theta)\right\} = \underset{q\in[p],c\in\mathbb{R}}{\arg\min}\, M(\theta).$$

Define $M^i_n(\theta) = \ell^i_n(\theta, \tau_i(\theta), \mathcal{R}^t_n)$ for $i \in \{L, R\}$ and the empirical optimal split $\widehat{\theta} = (\widehat{q}, \widehat{c})$ is defined via minimizing the sample criterion function:

$$(\widehat{q}, \widehat{c}) = \underset{q\in[p],c\in\mathbb{R}}{\arg\min}\left\{M^L_n(\theta) + M^R_n(\theta)\right\} =: \underset{q\in[p],c\in\mathbb{R}}{\arg\min}\, M_n(\theta).$$

Step 1 (Main error decomposition).

Now we will bound $M(\widehat{\theta}) - M(\theta_0)$, which represents the discrepancy of the MSEs of the oracle and empirical split. To apply empirical process theories for stochastic error analysis, we will use a truncation argument. We let $M^i_{n,\beta_n}(\theta, \pi_i, \mathcal{R}^t_n) = \mathbb{E}_n(T_{\beta_n}\tilde{Y} - T_{\beta_n}\pi_i(\theta))^2 I(X \in Q_i(\theta))$

and $M^i_{\beta_n}(\theta) = \mathbb{E}_T\left\{M^i_{n,\beta_n}(\theta, \pi_i(\theta), \mathcal{R}^t_n)\right\}$, where $T_{\beta_n}x =: (|x| \wedge \beta_n)\text{sign}(x)$ for any $\beta_n > 0$. Correspondingly, let $M_{\beta_n}(\theta) = M^L_{\beta_n}(\theta) + M^R_{\beta_n}(\theta)$ and $M_{n,\beta_n}(\theta) = M^L_{n,\beta_n}(\theta) + M^R_{n,\beta_n}(\theta)$. Then we have the following error decomposition:

$$
\begin{aligned}
&0 < M(\widehat{\theta}) - M(\theta_0) \\
=&M(\widehat{\theta}) - M_{\beta_n}(\widehat{\theta}) - M(\theta_0) + M_{\beta_n}(\theta_0) \\
&+ M_{\beta_n}(\widehat{\theta}) - M_{\beta_n}(\theta_0) - 2M_{n,\beta_n}(\widehat{\theta}) + 2M_{n,\beta_n}(\theta_0) \\
&+ 2M_{n,\beta_n}(\widehat{\theta}) - 2M_n(\widehat{\theta}) - 2M_{n,\beta_n}(\theta_0) + 2M_n(\theta_0) \\
&+ 2M_n(\widehat{\theta}) - 2M_n(\theta_0) \\
=:&S_{1,n} + S_{2,n} + S_{3,n} + S_{4,n}.
\end{aligned}
$$

By the definition of $\widehat{\theta}$, we have $S_{4,n} \leq 0$. In following steps, we will bound $S_{1,n}, S_{2,n}$ and $S_{3,n}$, respectively. The truncation level β_n is chosen as $\beta_n = \beta_0 \log(n)$ for $\beta_0 \geq 2\sigma_Y$.

Step 2 (Bounding $S_{1,n}$).

For any θ, it holds that

$$
\begin{aligned}
M^i(\theta) - M^i_{\beta_n}(\theta) =&\mathbb{E}_T\left\{(\tilde{Y} - \widehat{\tau}_i(\theta))^2 - (T_{\beta_n}\tilde{Y} - T_{\beta_n}\widehat{\tau}_i(\theta))^2 I\{X \in Q_i(\theta)\}\right\} \\
=&\mathbb{E}_T\left\{(\tilde{Y} - T_{\beta_n}\tilde{Y})(\tilde{Y} + T_{\beta_n}\tilde{Y} - 2\widehat{\pi}_i(\theta))I\{X \in Q_i(\theta)\}\right\} \\
&+ \mathbb{E}_T\left\{(T_{\beta_n}\widehat{\pi}_i(\theta) - \widehat{\pi}_i(\theta))(T_{\beta_n}\tilde{Y} + T_{\beta_n}\widehat{\pi}_i(\theta) - 2\widehat{\pi}_i(\theta))I\{X \in Q_i(\theta)\}\right\} \\
=:&S_{5,n} + S_{6,n}.
\end{aligned}
$$

For $T_{1,n}$, by Cauchy-Schwarz inequality we have

$$
|S_{5,n}| \leq \sqrt{\mathbb{E}_T(\tilde{Y} - T_{\beta_n}\tilde{Y})^2}\sqrt{\mathbb{E}_T(\tilde{Y} + T_{\beta_n}\tilde{Y} - 2\widehat{\pi}_i(\theta))^2} \lesssim \sqrt{\mathbb{E}_T(\tilde{Y} - T_{\beta_n}\tilde{Y})^2},
$$

where the second inequality is because $\mathbb{E}_T(\tilde{Y}^2) \leq \infty$ and $\mathbb{E}_T\left\{\widehat{\pi}^2_i(\theta)\right\} \leq \mathbb{E}_T(\tilde{Y}^2)/|Q_i(\theta)|$. Since

$$
I(|\tilde{Y}| > \beta_n) \leq \frac{\exp(\sigma_Y|Y|^2/2)}{\sigma_Y\beta_n^2/2},
$$

therefore,

$$
|T_{1,n}| \lesssim \sqrt{\mathbb{E}_T(\tilde{Y} - T_{\beta_n}\tilde{Y})^2} \leq \sqrt{\mathbb{E}_T\left\{|Y|^2\frac{\exp(\sigma_Y|Y|^2/2)}{\sigma_Y\beta_n^2/2}\right\}} \leq \sqrt{\frac{2}{\sigma_Y}\mathbb{E}_T\exp(\sigma_Y|Y|^2)}\exp(-\frac{\sigma_Y\beta_n^2}{4}).
$$

Since $\mathbb{E}_T\exp(\sigma_Y|Y|^2) < \infty$ and $\beta_n = \beta_0\log(n)$, we conclude that $|S_{5,n}| \lesssim \frac{1}{n}$. With the same argument, we have $S_{6,n} \lesssim \frac{1}{n}$, implying that

$$
M(\theta) - M_{\beta_n}(\theta) \lesssim \frac{1}{n} \tag{S2}
$$

for any $\theta \in \Theta$. Therefore, the truncation error $S_{1,n} \lesssim \frac{1}{n}$.

Step 3 (Bounding $S_{2,n}$).

Let $M_{N,of} = \left\{f = (T_{\beta_n}\tilde{Y} - T_{\beta_n}\pi)I(X \in Q_i(\theta)) : \theta = (q,c) \in [p] \times \mathbb{R}\right\}$. By applying Lemma 2 we obtain

$$
\mathcal{N}_1(\delta, M_{N,of}, z_1^n) \leq (pn)^2\left(\frac{c\beta_n}{\delta}\right)^4,
$$

where z_1^n is any set $\{z_1, \cdots, z_n\} \in [0,1]^p \times \mathcal{Y}$ and $c > 0$ is a universal constant. It follows from Lemma 1 that

$$\begin{aligned}
&\mathbb{P}_T \left\{ \exists \theta \in \Theta : |M_{\beta_n}(\theta) - M_{n,\beta_n}(\theta)| \geq \frac{1}{2}(\alpha + \gamma + M_{\beta_n}(\theta)) \right\} \\
\leq & 28(pn)^2 \left(\frac{80c\beta_n^2}{\gamma} \right)^4 \exp\left(-\frac{\alpha n}{1284\beta_n^4} \right) \\
\lesssim & \exp\left(-\frac{\alpha n}{\beta_n^4} + \log(pn) - \log(\gamma) \right).
\end{aligned}$$

Taking $\gamma = 1/n$ and $\alpha = (t + \log(pn))\beta_n^4/n$ implies that with probability at least $1 - C_1 e^{-t}$ for some universal constant $C_1 > 0$,

$$\forall \theta \in \Theta, |M_{\beta_n}(\theta) - 2M_{n,\beta_n}(\theta)| \lesssim \frac{t + \log(pn)\log^4(n)}{n}. \tag{S3}$$

Therefore, we conclude that with probability at least $1 - C_1 e^{-t}$, the stochastic error $S_{2,n} \lesssim \left\{ t + \log(pn)\log^4(n) \right\}/n$.

Step 4 (Bounding $S_{3,n}$). According to (S2), we have

$$\forall \theta \in \Theta : \mathbb{E}_T \left\{ M_{n,\beta_n}(\theta) - M_n(\theta) \right\} \lesssim \frac{1}{n}$$

Since $\tilde{Y}$ is sub-Gaussian by assumption, it is straightforward to see that $(T_{\beta_n}\tilde{Y} - T_{\beta_n}\pi_i(\theta))^2 I(X \in Q_i(\theta))$ and $(\tilde{Y} - \pi_i(\theta))^2 I(X \in Q_i(\theta))$ are sub-exponential for $i \in \{L, R\}$. Suppose $\left\| (T_{\beta_n}\tilde{Y} - T_{\beta_n}\pi_i(\theta))^2 I(X \in Q_i(\theta)) \right\|_{\psi_1} \leq \sigma_0$ and $\left\| (\tilde{Y} - \pi_i(\theta))^2 I(X \in Q_i(\theta)) \right\|_{\psi_1} \leq \sigma_0$ for all $\theta \in \Theta$, where $\|\cdot\|_{\psi_1}$ is the sub-exponential norm operator. By applying Bernstein's inequality, for any $s > 0$, we have

$$\begin{aligned}
&\mathbb{P}_T \left\{ \left| M_{n,\beta_n}^i(\theta) - M_n^i(\theta) - \mathbb{E}_T \left\{ M_{n,\beta_n}^i(\theta) - M_n^i(\theta) \right\} \geq s \right| \right\} \\
\leq & 2\exp\left(-c \min\left(\frac{ns^2}{\sigma_0^2}, \frac{ns}{\sigma_0} \right) \right),
\end{aligned}$$

for $i \in \{R, L\}$, where $c > 0$ is a universal constant. Taking $s = \frac{\sigma_0 t}{cn} = C_2 t$, for any $t \geq 0$ we obtain

$$\mathbb{P}_T \left\{ \left| M_{n,\beta_n}^i(\theta) - M_n^i(\theta) - \mathbb{E}_T \left\{ M_{n,\beta_n}^i(\theta) - M_n^i(\theta) \right\} \geq C_2 t \right| \right\} \leq 2\exp(-t) \tag{S4}$$

for any $n > t/c$. Since the above result holds for any $\theta \in \Theta$, we conclude that for any $t > 0$, with probability at least $1 - 4e^{-t}$, we have $S_{3,n} \lesssim (t+1)/n$.

Combining the results on $S_{1,n}, S_{2,n}$ and $S_{3,n}$, we conclude that for any $t > 0$, with probability at least $1 - C_3 e^{-t}$, it holds that

$$\mathcal{L}_i(\widehat{\theta}, \widehat{\pi}_i(\widehat{\theta})) - \mathcal{L}_i(\theta_0, \pi(\theta_0)) \lesssim \frac{t + \log(pn)\log^4(n)}{n}, \tag{S5}$$

for some universal constants $C_3, C_4 > 0$.

Part 2. The proposed scale criterion can reformulated as follows. For $i \in \{L, R\}$, let

$$\begin{aligned}
F_{0,i}(\theta) =& \{1 - w_i(\theta)\} (\tilde{Y}_0 - \tau_{0,i}(\theta))^2 I(X_0 \in Q_i(\theta)) \\
F_{1,i}(\theta) =& w_i(\theta)(\tilde{Y}_1 - \tau_{1,i}(\theta))^2 I(X_1 \in Q_i(\theta)) \text{ and}
\end{aligned}$$

where $\tau_{0,i}(\theta) = \mathbb{E}_n(\tilde{Y}_0 | X_0 \in Q_i(\theta))$ and $\tau_{1,i}(\theta) = \mathbb{E}_m(\tilde{Y}_1 | X_1 \in Q_i(\theta))$, and

$$w_i(\theta) = \sigma_u^2(Q_i(\theta)) / \left\{ \sigma_u^2(Q_i(\theta)) + \sigma_b^2(Q_i(\theta)) + b^2(Q_i(\theta)) \right\},$$

where $\sigma_u^2(Q_i(\theta)) = \text{Var}_n(\tau_{0,i}(\theta)), \sigma_b^2(Q_i(\theta)) = \text{Var}_m(\tau_{1,i}(\theta))$ and $b(Q_i(\theta)) = \tau_{1,i}(\theta) - \tau_{0,1}(\theta)$. Let $\mathcal{F}_{s,i}(\theta) = \mathbb{E}_s(F_{s,i}(\theta))$ for $s \in \{0, 1\}$ and $\mathcal{F}_s(\theta) = \mathcal{F}_{s,L}(\theta) + \mathcal{F}_{s,R}(\theta)$, the population criterion is defined as $M_{of}(\theta) = \mathcal{F}_0(\theta) + \mathcal{F}_1(\theta)$. For the empirical criterion, we first define $\mathcal{F}_{n,i}(\theta) =$

$\mathbb{E}_n(F_i^R(\theta))$ and $\mathcal{F}_{m,i}(\theta) = \mathbb{E}_m(F_i^R(\theta))$. Let $M_{N,of}(\theta) = \mathcal{F}_{n,L}(\theta) + \mathcal{F}_{n,R}(\theta)$ and $\mathcal{F}_m(\theta) = \mathcal{F}_{m,L}(\theta) + \mathcal{F}_{m,R}(\theta)$, the empirical criterion is the denoted as $M_{N,of}(\theta) = M_{N,of}(\theta) + \mathcal{F}_m(\theta)$. The population and empirical optimal splits are defined by

$$\theta_{of} = \underset{\theta \in \Theta}{\arg\min}\, M_{of}(\theta) \text{ and } \widehat{\theta}_f = \underset{\theta \in \Theta}{\arg\min}\, M_{N,of}(\theta).$$

We first have the following error decomposition:

$$\begin{aligned} M_{of}(\widehat{\theta}) - M_{of}(\theta_0) =& M_{of}(\widehat{\theta}) - M_{of,\beta_n}(\widehat{\theta}_{of}) + M_{of}(\theta_0) - M_{of,\beta_n}(\theta_{of}) \\ &+ M_{of,\beta_n}(\widehat{\theta}_{of}) + M_{of,\beta_n}(\theta_{of}) - 2M_{N,of,\beta_n}(\widehat{\theta}_{of}) + 2M_{N,of,\beta_n}(\theta_{of}) \\ &+ 2M_{N,of,\beta_n}(\widehat{\theta}_{of}) - 2M_{N,of}(\widehat{\theta}_{of}) - 2M_{N,of,\beta n}(\theta_{of}) + 2M_{N,of}(\theta_{of}) \\ &+ 2M_{N,of}(\widehat{\theta}_{of}) - 2M_{N,of}(\theta_{of}) \\ =:& T_{1,n} + T_{2,n} + T_{3,n} + T_{4,n}. \end{aligned}$$

By the definition of $\widehat{\theta}_{of}$, we have $T_{4,n} \leq 0$. In the following steps, we will bound $T_{1,n}$, $T_{2,n}$ and $T_{3,n}$, respectively. Following the same argument as for $S_{1,n}$, it can be obtained that $T_{1,n} \lesssim \frac{1}{n}$. We now bound $T_{2,n}$ Let $\mathcal{G}_n = \{g : g(y,x) = \sqrt{1-w(\theta)}\tilde{y} - \tau)I(x \in \mathcal{Q}(\theta), \theta \in \Theta_n)\}$ and $\mathcal{H}_n = \{h : h(y,x) = \sqrt{w(\theta)}(\tilde{y} - \tau)I(x \in \mathcal{Q}(\theta), \theta \in \Theta_n)\}$, then via Lemma 2 we have

$$\mathcal{N}_1(\delta, \mathcal{G}_n, z_1^n) \leq (pn)^2 \left(\frac{c\beta_n}{\delta}\right)^4 \text{ and } \mathcal{N}_1(\delta, \mathcal{H}_n, z_1^n) \leq (pn)^2 \left(\frac{c\beta_n}{\delta}\right)^4,$$

for any $\delta > 0$, where z_1^n is any set $\{z_1, \cdots, z_n\} \in [0,1]^p \times \mathcal{Y}$ and $c > 0$ is a universal constant. It follows from Lemma 1 that for any $\alpha_1, \gamma_1 > 0$

$$\begin{aligned} &\mathbb{P}_T \left\{ \exists \theta \in \Theta_n : |\mathcal{F}_{0,\beta_n}(\theta) - \mathcal{F}_{n,\beta_n}(\theta)| \geq \frac{1}{2}(\alpha_1 + \gamma_1 + M_{of,\beta_n}(\theta)) \right\} \\ \leq& 28(pn)^2 \left(\frac{80c\beta_n^2}{\gamma_1}\right)^4 \exp\left(-\frac{\alpha_1 n}{1284\beta_n^4}\right) \\ \lesssim& \exp\left(-\frac{\alpha_1 n}{\beta_n^4} + \log(pn) - \log(\gamma_1)\right). \end{aligned}$$

Taking $\gamma_1 = 1/n$ and $\alpha_1 = (t + \log(pn))\beta_n^4/n$ implies that with probability at least $1 - C_4 e^{-t}$ for some universal constant $C_4 > 0$,

$$\forall \theta \in \Theta_n, |M_{of,\beta_n}(\theta) - 2\mathcal{F}_{n,\beta_n}(\theta)| \lesssim \frac{t + \log(pn)\log^4(n)}{n}. \tag{S6}$$

Similary, for any $\alpha_2, \gamma_2 > 0$,

$$\begin{aligned} &\mathbb{P}_O \left\{ \exists \theta \in \Theta_n : |\mathcal{F}_{1,\beta_n}(\theta) - \mathcal{F}_{m,\beta_n}(\theta)| \geq \frac{1}{2}(\alpha_1 + \gamma_1 + \mathcal{F}_{1,\beta_n}(\theta)) \right\} \\ \lesssim& \exp\left(-\frac{\alpha_1 m}{\beta_n^4} + \log(pn) - \log(\gamma_1)\right). \end{aligned}$$

Taking $\gamma_2 = 1/n$ and $\alpha_2 = (t + \log(pn))\beta_n^4/m$ implies that with probability at least $1 - C_5 e^{-t}$ for some universal constant $C_5 > 0$,

$$\forall \theta \in \Theta_n, |\mathcal{F}_{1,\beta_n}(\theta) - 2\mathcal{F}_{m,\beta_n}(\theta)| \lesssim \frac{t + \log(pn)\log^4(n)}{m}. \tag{S7}$$

Combining (S6) and (S7) delivers that with probability at least $1 - 2C_1 e^{-t}$,

$$T_{2,n} \lesssim \frac{\log(pn)\log^4(n)}{m} + \frac{t + \log(pn)\log^4(n)}{n}, \tag{S8}$$

for ant $t > 0$, since $\widehat{\theta}_{of}, \theta_f \in \Theta_n$ and $M_{of,\beta_n}(\theta) = \mathcal{F}_{0,\beta_n}(\theta) + \mathcal{F}_{1,\beta_n}(\theta)$ and $M_{N,of,\beta_n}(\theta) = \mathcal{F}_{n,\beta_n}(\theta) + \mathcal{F}_{m,\beta_n}(\theta)$ for any $\theta \in \Theta_n$.

Now we turn to $T_{3,n}$, the truncation error for the empirical loss. With the similar argument as in (S3), we have with probability at least $1-4e^{-t}$, it holds that $T_{3,n} \lesssim (t+1)/n + (t+1)/m$ for any $t>0$.

Combining the results for $T_{1,n}, T_{2,n}$ and $T_{3,n}$, we conclude that for any $t>0$, with probability at least $1-C_6e^{-t}$,

$$M_{of}(\widehat{\theta}_{of}) - M_{of}(\theta_{of}) \lesssim \frac{t+\log(pn)\log^4(n)}{m} + \frac{t+\log(pn)\log^4(n)}{n}, \tag{S9}$$

which completes our proof.

S4 Supporting lemmas

The following to lemmas are from Section 11.3 and Section 13.1 of Györfi et al. [2002], which are useful for our proofs.

Lemma S1. *(Deviation inequality of quadratic process). Suppose that $\mathcal{G}$ is a class of uniformly bounded functions $\mathcal{G} = \left\{g : \mathbb{R}^d \to \mathbb{R} \, \|g\|_\infty \le M\right\}$. Let $\mathcal{F} = \left\{g^2 : g \in \mathcal{G}\right\}$. Then for any $n \ge 1$, it holds that*

$$\begin{aligned}&\mathbb{P}\left\{\exists f \in \mathcal{F} : \left|\mathbb{E}\left\{f(z)\right\} - \mathbb{E}_n\left\{f(z)\right\}\right| \ge \varepsilon(\alpha+\gamma) + \mathbb{E}\left\{f(z)\right\}\right\} \\ \le& 28 \sup_{z_1^n} \mathcal{N}_1(\frac{\gamma\varepsilon}{20M}, \mathcal{G}, x_1^n) \exp\left(-\frac{\varepsilon^2(1-\varepsilon)\alpha n}{214(1+\varepsilon)M^4}\right),\end{aligned}$$

where $z_1^n = (z_1, \cdots, z_n) \in \mathbb{R}^d$, $\alpha, \gamma > 0$ and $0 < \varepsilon \le 1/2$.

Lemma S2. *(Covering number of piece-wise constant functions). Let Π be the family of partitions of $[0,1]^p$. For any set $x_1^n = \{x_1, \cdots, x_n\} \subset [0,1]^p$, let $\Delta(\Pi)$ be the maximal number of partitions of x_1^n induced by elements of Π. Let $M(\Pi)$ be the maximal number of sets contained in a partition $\mathcal{P} \in \Pi$. Denote the piece-wise constant functions on $[0,1]^p$ be $\mathcal{F}(\Pi)$ with $\|f\|_\infty \le \beta_n$ for any $f \in \mathcal{F}(\Pi)$. Then using Lemma 13.31 and Theorem 9.4 of Györfi et al. [2002] we have*

$$\mathcal{N}_1(\delta, \mathcal{F}(\Pi), x_1^n) \le \Delta_n(\Pi) \left(\frac{c_1\beta_n}{\delta}\right)^{2M(\Pi)},$$

for any $\delta > 0$, where $c_1 > 0$ is some univiersal constant. Specifically, in each partition for a node $\mathcal{C}_k$ of a tree, we have $M(\Pi) = 2$ and $\Delta_n(\Pi) \le (pa_n)^2$, where a_n is the sample size of $\mathcal{C}_k$.

References

L. Györfi, M. Köhler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716 – 1741, 2015. doi: 10.1214/15-AOS1321. URL `https://doi.org/10.1214/15-AOS1321`.