Supplemental Document for "A Frequentist Approach to Computer Model Calibration"

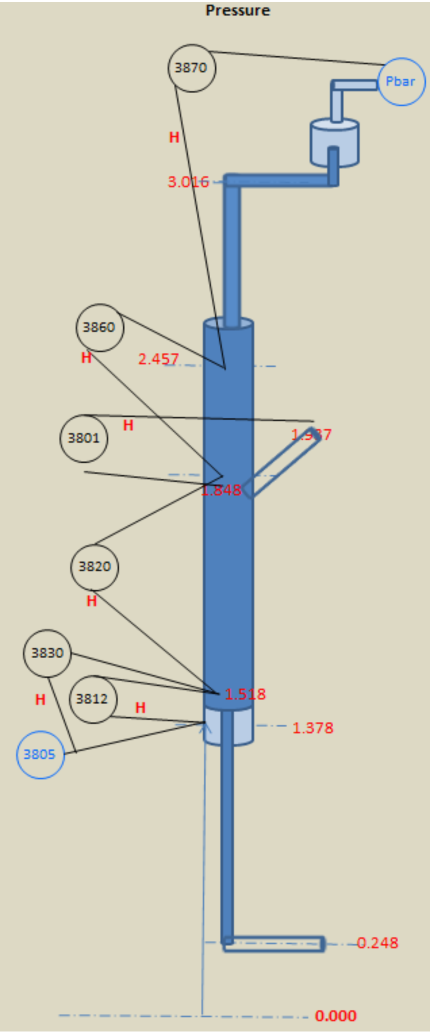
Raymond K. W. Wong[*] Curtis B. Storlie[†]
Thomas C. M. Lee[‡]
[*]*Department of Statistics, Iowa State University*
[†]*Los Alamos National Laboratory*
[‡]*Department of Statistics, University of California, Davis*

March 1, 2016

Abstract

This document provides supplementary material to the article "A Frequentist Approach to Computer Model Calibration" written by the same authors.

S1 Picture of benchscale CO_2 capture unit (C2U)



S2 Simulation Study

A simulation study was conducted to investigate the practical performance of the proposed methodology. The following four simulation configurations were considered:

- Configuration 1:

$$\mathcal{X} = [0,1], \quad \Theta = [0, 0.25] \times [0, 0.5],$$

$$\boldsymbol{\theta}_0 = (0.2, 0.3)^\intercal, \quad \delta_0(x) = \cos(2\pi x - \pi),$$

$$\eta(x, \boldsymbol{\theta}) = 7\{\sin(2\pi\theta_1 - \pi)\}^2 + 2\{(2\pi\theta_2 - \pi)^2 \sin(2\pi x - \pi)\}, \quad \boldsymbol{\theta} = (\theta_1, \theta_2)^\intercal$$

- Configuration 2:

$$\mathcal{X} = [0,1]^2, \quad \Theta = [0,0.25] \times [0,0.5] \times [0,1],$$
$$\boldsymbol{\theta}_0 = (0.2, 0.3, 0.8)^\intercal, \quad \delta_0(\boldsymbol{x}) = \cos(2\pi x_1 - \pi) + 2\left(x_2^2 - x_2 + \frac{1}{6}\right), \quad \boldsymbol{x} = (x_1, x_2)^\intercal,$$
$$\eta(\boldsymbol{x}, \boldsymbol{\theta}) = 7\{\sin(2\pi\theta_1 - \pi)\}^2 + 2\{(2\pi\theta_2 - \pi)^2 \sin(2\pi x_1 - \pi)\} + 6\theta_3(x_2 - 0.5),$$
$$\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^\intercal$$

- Configuration 3:

$$\mathcal{X} = [0,1]^2, \quad \Theta = [0,1]^2,$$
$$\boldsymbol{\theta}_0 = (0.2, 0.4)^\intercal, \quad \delta_0(\boldsymbol{x}) = \exp(-x_1)\left(x_1 - \frac{1}{2}\right)\left(x_2^2 - x_2 + \frac{1}{6}\right), \quad \boldsymbol{x} = (x_1, x_2)^\intercal,$$
$$\eta(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{2}{3}\exp(x_1 + \theta_1) - x_2\sin(\theta_2) + \theta_2, \quad \boldsymbol{\theta} = (\theta_1, \theta_2)^\intercal$$

- Configuration 4:

$$\mathcal{X} = [0,1]^2, \quad \Theta = [0,1]^2,$$
$$\boldsymbol{\theta}_0 = (0.6, 0.2)^\intercal, \quad \delta_0(\boldsymbol{x}) = 0, \quad \boldsymbol{x} = (x_1, x_2)^\intercal,$$
$$\eta(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{2}\theta_1\left[\sqrt{1 + (\theta_2 + x_1^2)\frac{x_2}{\theta_1^2}} - 1\right] + (\theta_1 + 3x_2)\exp\{1 + \sin(x_1)\}, \quad \boldsymbol{\theta} = (\theta_1, \theta_2)^\intercal$$

Note that, for a given pair of ζ and η, we can determine $\boldsymbol{\theta}_0$ (and therefore δ_0) by minimizing (2). Therefore the above configurations essentially correspond to four pairs of ζ and η. Note that Configuration 2 is a modification of Configuration 1. In below, we explain how this modification would be interesting in terms of its effects on parameter estimation and uncertainty quantification. As for Configurations 3 and 4, the physical reality η are test functions used in Park (1991)[1]. Moreover, Configuration 4 is coupled with no discrepancy to mimic the ideal scenario where the physical reality can be completely modeled by the computer model. For all configurations, we used $n = 50$ and $m = 300$. Both ε_i's and τ_j's are independent normal random variables with signal-to-noise ratios (snrs) set to 10 and 55, respectively. These snrs are defined respectively as

$$\int_{\mathcal{X}} \left\{\zeta(\boldsymbol{x}) - \int_{\mathcal{X}} \zeta(\boldsymbol{x}) d\boldsymbol{x}\right\}^2 d\boldsymbol{x} / \mathrm{Var}(\varepsilon_1)$$

and

$$\int_{\mathcal{X}\times\Theta} \left\{\eta(\boldsymbol{x}, \boldsymbol{\theta}) - \int_{\mathcal{X}\times\Theta} \eta(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} d\boldsymbol{\theta}\right\}^2 d\boldsymbol{x} d\boldsymbol{\theta} / \mathrm{Var}(\tau_1).$$

[1] See `http://www.sfu.ca/~ssurjano/calibrat.html` for other references of these test functions.

Their values were chosen to mimic common practical situations that the snrs of the simulator data are significantly higher than that of the experimental data. Both designs in the experimental data and the simulator data are generated by Latin hypercube sampling (McKay et al., 1979). We note that the above configurations do not attempt to cover the huge range of possible practical configurations.

For each of the four configurations, 200 data sets were simulated according to (3), to which the proposed frequentist calibration method was applied to estimate $\boldsymbol{\theta}_0$ and δ_0. Smoothing spline ANOVA (with main effects and two-way interactions) was used as the nonparametric regression model for both η and δ_0, with the corresponding smoothing parameters selected by generalized cross-validation. We used the R package gss (Gu, 2014) for the practical implementation.

Uncertainty measures on $\boldsymbol{\theta}_0$ and δ_0 were constructed using the following two methods:

1. fboot: The proposed frequentist method coupled with the bootstrap procedure of Section 3 *without* re-sampling of the design (i.e., skip Step 1).

2. fboot-rs: The proposed frequentist method coupled with the bootstrap procedure of Section 3 *with* re-sampling of the design (i.e., keep Step 1).

As discussed in Section 2.1, the definition of the ideal parameter value $\boldsymbol{\theta}_*$ is important. While the proposed methods target at $\boldsymbol{\theta}_* = \boldsymbol{\theta}_0$, most Bayesian alternatives target at the physical parameter (whenever possible) and require the the specification of its corresponding prior distribution. Due to the different targets (and the effects of prior distribution), it would be difficult to perform a fair comparison between our method and the Bayesian alternatives. To avoid misleading results, we only demonstrate the practical performances of our methods in Configurations 1, 2 and 3. However, Configuration 4 has no discrepancy function and hence the computer model can be tuned as the physical reality exactly at the parameter value $\boldsymbol{\theta}_0$ which seems to be also a reasonable target of the Bayesian alternatives. Therefore, we apply a state-of-the-art calibration method, Bayesian smoothing spline ANOVA (Storlie et al., 2014) bss-anova, with the suggested generic prior specifications to give a reference comparsion. It is noted that the prior distribution of the discrepancy function is a mean zero Gaussian process which matches with no discrepancy scenario. Since $\boldsymbol{\theta}_0$ is a reasonable target for both methods, it is expected to see that none of the two methods outperform the other in a clear way. The goal of this comparison is to demonstrate the empirical performances of the proposed methods with a Bayesian method as reference. Finally, we note that model parameters, computer model and discrepancy are non-random in each configuration, as similar to Storlie et al. (2014). The Bayesian methods are not intended to have exact nominal coverage under this non-random setting.

The mean squared error (MSE) for each element of $\boldsymbol{\theta}_0 = (\theta_{0,1}, \dots \theta_{0,d})$, where $d = 2$ for Configurations 1, 3 and 4, and $d = 3$ for Configuration 2, are shown in Tables S1 and S2. For bss-anova, the posterior mean of $\boldsymbol{\theta}_0$ is treated as the estimate and the corresponding MSE is computed. In Configuration 4, the MSE of bss-anova is double as that of the proposed method in $\theta_{0,1}$, and vice versa in $\theta_{0,2}$. Therefore, the proposed method and bss-anova do not outperform the other clearly.

For both configurations, $\theta_{0,1}$ and $\theta_{0,2}$ have small MSEs. However, for Configuration 2, $\theta_{0,3}$ has a relatively larger MSE which results from the fact that $\theta_{0,3}$ is fundamentally difficult to estimate. Note that the variability of η, defined as

$$\int_{\mathcal{X}\times\Theta} \left\{ \eta(\boldsymbol{x}, \boldsymbol{\theta}_0) - \int_{\mathcal{X}\times\Theta} \eta(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x} d\boldsymbol{\theta} \right\}^2 d\boldsymbol{x} d\boldsymbol{\theta},$$

of Configurations 1 and 2 are approximately 45.1 and 46.1, meaning the additional signal introduced by $\theta_{0,3}$ is relatively weak. In below, the results of uncertainty quantification show that the proposed method detected this issue and produced wider confidence or credible intervals for $\theta_{0,3}$.

As for the discrepancy function, the averaged MSEs, with standard errors, are also shown in Tables S1 and S2. Each MSE is computed over a grid of $\mathcal{X}$. The estimate of the discrepancy function from the bss-anova procedure was taken to be the posterior mean at each grid point. In Configuration 4, as expected, there is no insignificant difference between the two approaches.

The simulation results pertaining to uncertainty quantification are summarized in Tables S3-S6. In Configuration 2, the increased difficulty in estimating $\theta_{0,3}$ has led to a wider confidence intervals for $\theta_{0,3}$, which matches with general intuition. Overall, our methods provide coverage close to the nominal rate of 95%.

Table S1: The mean squared errors of $\boldsymbol{\theta}_0$ and δ_0 for Configurations 1, 2 and 3 with the corresponding standard errors shown in parentheses.

	Configuration		
	1	2	3
$\theta_{0,1}$	1.61e-04 (0.330e-04)	1.99e-04 (0.350e-04)	1.18e-03 (0.123e-03)
$\theta_{0,2}$	7.07e-05 (0.875e-05)	8.89e-05 (1.17e-05)	6.12e-03 (0.662e-03)
$\theta_{0,3}$	-	1.45e-02 (0.119e-02)	-
δ_0	9.43e-02 (0.525e-02)	3.01e-01 (0.147e-01)	3.60e-03 (0.286e-03)

Table S2: The mean squared errors of $\boldsymbol{\theta}_0$ and δ_0 for Configuration 4 with the corresponding standard errors shown in parentheses.

	proposed method	bss-anova
$\theta_{0,1}$	3.08e-03 (0.409e-03)	6.64e-03 (0.437e-03)
$\theta_{0,2}$	21.2e-02 (1.66e-02)	9.27e-02 (0.199e-02)
δ_0	2.95e-01 (0.213e-01)	2.76e-01 (0.140e-01)

Table S3: Simulation results of 95% confidence (credible) intervals of the elements of $\boldsymbol{\theta}_0$ for Configuration 1: Average coverages and lengths of 95% confidence (credible) intervals. The standard errors are shown in parentheses.

		fboot	fboot-rs
coverage	$\theta_{0,1}$	97.5% (1.11%)	98.5% (0.862%)
	$\theta_{0,2}$	94.5% (1.62%)	96.0% (1.39%)
length	$\theta_{0,1}$	4.07e-02 (0.105e-02)	4.10e-02 (0.102e-02)
	$\theta_{0,2}$	3.17e-02 (0.0345e-02)	3.14e-02 (0.0339e-02)

Table S4: Simulation results of 95% confidence (credible) intervals of the elements of $\boldsymbol{\theta}_0$ for Configuration 2: Same format as that in Table S3.

		fboot	fboot-rs
coverage	$\theta_{0,1}$	98.5% (0.862%)	96.5% (1.30%)
	$\theta_{0,2}$	98.0% (0.992%)	99.0% (0.705%)
	$\theta_{0,3}$	96.5% (1.30%)	98.5% (0.862%)
length	$\theta_{0,1}$	4.70e-02 (0.103e-02)	4.65e-02 (0.0977e-02)
	$\theta_{0,2}$	3.56e-02 (0.0462e-02)	3.64e-02 (0.0489e-02)
	$\theta_{0,3}$	4.25e-01 (0.0687e-01)	4.39e-01 (0.0595e-01)

Table S5: Simulation results of 95% confidence (credible) intervals of the elements of $\boldsymbol{\theta}_0$ for Configuration 3: Same format as that in Table S3.

		fboot	fboot-rs
coverage	$\theta_{0,1}$	96.5% (1.30%)	95.0% (1.54%)
	$\theta_{0,2}$	96.0% (1.39%)	95.5% (1.47%)
length	$\theta_{0,1}$	1.39e-01 (0.0222e-01)	1.41e-01 (0.0225e-01)
	$\theta_{0,2}$	3.32e-01 (0.0536e-01)	3.38e-01 (0.0546e-01)

Table S6: Simulation results of 95% confidence (credible) intervals of the elements of $\boldsymbol{\theta}_0$ for Configuration 4: Same format as that in Table S3.

		fboot	fboot-rs
coverage	$\theta_{0,1}$	92.5% (1.87%)	92.5% (1.87%)
	$\theta_{0,2}$	99.0% (0.705%)	98.5% (0.862%)
length	$\theta_{0,1}$	2.05e-01 (0.0218e-01)	2.08e-01 (0.0231e-01)
	$\theta_{0,2}$	9.86e-01 (0.0319e-01)	9.87e-01 (0.0370e-01)

S3 Technical details

Lemma 1 (Consistency of $\hat{\boldsymbol{\theta}}_n$)**.** *Suppose that Assumptions 1, 2, 3(a), 4(a), 5, and 6(a) hold. Then $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$. That means, $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E \xrightarrow{P} 0$ as $n \to \infty$.*

Proof of Lemma 1. Recall that $g_{\boldsymbol{\theta}}(\boldsymbol{x}) = \eta(\boldsymbol{x}, \boldsymbol{\theta})$ and $\zeta(\boldsymbol{x}) = g(\boldsymbol{x}, \boldsymbol{\theta}_0) + \delta_0(\boldsymbol{x})$. We have

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 + \|\zeta - g_{\boldsymbol{\theta}}\|_n^2 + 2\langle \varepsilon, \zeta - g_{\boldsymbol{\theta}} \rangle_n. \tag{S1}$$

Define

$$M_{0,n}(\boldsymbol{\theta}) = \|\zeta - g_{\boldsymbol{\theta}}\|_n^2 + \sigma^2.$$

To show the uniform convergence of $M_n(\boldsymbol{\theta}) - M_{0,n}(\boldsymbol{\theta})$ over $\boldsymbol{\theta} \in \Theta$, we first obtain the uniform convergence of $\langle \varepsilon, \zeta - g_{\boldsymbol{\theta}} \rangle_n$. Recall that $\mathcal{G} = \{g_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ and $\mathcal{G} - \zeta = \{g_{\boldsymbol{\theta}} - \zeta : \boldsymbol{\theta} \in \Theta\}$. For any $\boldsymbol{\theta} \in \Theta$, $\|g_{\boldsymbol{\theta}} - \zeta\|_n \leq \|g_{\boldsymbol{\theta}}\|_n + \|g_{\boldsymbol{\theta}_0}\|_n + \|\delta_0\|_n < \infty$ due to Assumptions 3(a) and 6(a). Using Lemma 2.5 of van de Geer (2000) with Assumption 2 and 3(a),

$$H(u, \mathcal{G} - \zeta, F_n) \leq d \log\left(\frac{4R_0 c_0 + u}{u}\right),$$

where $H(u, \mathcal{G} - \zeta, F_n)$ is the u-entropy of $\mathcal{G} - \zeta$ for the $L_2(F_n)$-metric (Definition 2 of van de Geer, 2000). Thus the entropy integral converges:

$$\int_0^1 H^{1/2}(u, \mathcal{G} - \zeta, F_n) du < \infty.$$

Hence, using Corollary 8.3 of van de Geer (2000) with Assumptions 1, we have

$$\sup_{\boldsymbol{\theta} \in \Theta} |\langle \varepsilon, \zeta - g_{\boldsymbol{\theta}} \rangle_n| = o_p(1).$$

By Bernstein's inequality, we have $(1/n)\sum_{i=1}^{n} \varepsilon_i^2 \xrightarrow{P} \sigma^2$ and thus (S1) implies $\sup_{\boldsymbol{\theta} \in \Theta} |M_n(\boldsymbol{\theta}) - M_{0,n}(\boldsymbol{\theta})| = o_p(1)$. Consider

$$\sup_{\boldsymbol{\theta} \in \Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| \leq \sup_{\boldsymbol{\theta} \in \Theta} |M_n(\boldsymbol{\theta}) - M_{0,n}(\boldsymbol{\theta})| + \sup_{\boldsymbol{\theta} \in \Theta} |M_{0,n}(\boldsymbol{\theta}) - M(\boldsymbol{\theta})|,$$

where $M(\boldsymbol{\theta}) = \|\zeta - g_{\boldsymbol{\theta}}\|^2 + \sigma^2$. By Assumption 4(a), $\sup_{\boldsymbol{\theta} \in \Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| = o_p(1)$. By Theorem 5.7 of van der Vaart (2000) and Assumption 5, $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E = o_p(1)$. □

Lemma 2. *Suppose Assumptions 1, 2 and 3(a) hold. There exist constants $\tilde{C}$ and c (that only depend on the constants appearing in Assumptions 1, 2 and 3(a)) such that for $T > 0$ and $\xi \geq \xi_n$,*

$$\Pr\left(\sup_{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_E > \xi} \frac{|\langle \varepsilon, g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0} \rangle_n|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_E^2} \geq T\right) \leq c \exp\left(-\frac{n\xi^2 T^2}{c^2}\right),$$

where $\sqrt{n}\xi_n \geq \tilde{C}/T$.

Proof of Lemma 2. Due to Assumption 2, $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_E \le 2R_0$ for any $\boldsymbol{\theta} \in \Theta$. Let $S = \min\{s \in \{0, 1, \dots\} : 2^{s+1}\xi \ge 2R_0\}$. Using the peeling device (see, e.g., Section 5.3 of van de Geer (2000)),

$$\begin{aligned}
\Pr\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_E>\xi} \frac{|\langle \varepsilon, g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0}\rangle_n|}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_E^2} \ge T \right) &\le \sum_{s=0}^{S} \Pr\left(\sup_{\boldsymbol{\theta}:2^s\xi<\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_E\le 2^{s+1}\xi} \frac{|\langle \varepsilon, g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0}\rangle_n|}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_E^2} \ge T \right) \\
&\le \sum_{s=0}^{S} \Pr\left(\sup_{\boldsymbol{\theta}:2^s\xi<\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_E\le 2^{s+1}\xi} |\langle \varepsilon, g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0}\rangle_n| \ge 2^{2s}\xi^2 T \right) \\
&\le \sum_{s=0}^{S} \Pr\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_E\le 2^{s+1}\xi} |\langle \varepsilon, g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0}\rangle_n| \ge 2^{2s}\xi^2 T \right)
\end{aligned}$$

To deal with each probability term in the summation, we apply Corollary 8.3 of van de Geer (2000). Let $\mathcal{G}(z) = \{g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0} : \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_E \le z, \boldsymbol{\theta} \in \Theta\}$. Due to Assumptions 2 and 3(a), $\sup_{g\in\mathcal{G}(z)} \|g\|_n \le c_0 z$. By Assumptions 2 and 3(a), and Lemma 2.5 of van de Geer (2000),

$$\begin{aligned}
\int_0^{c_0 z} H^{1/2}(u, \mathcal{G}_n(z), F_n) du &\le \int_0^{c_0 z} d^{1/2} \left\{ \log\left(\frac{4c_0 z + u}{u} \right) \right\}^{1/2} du \\
&= 4c_0 d^{1/2} z \int_0^{1/4} \left\{ \log\left(\frac{1}{\tilde{u}} + 1 \right) \right\}^{1/2} d\tilde{u} \\
&\le \tilde{K} z,
\end{aligned}$$

for a constant $\tilde{K}$ that also satisfies $\tilde{K} \ge c_0$. Take $\sqrt{n}\xi_n \ge 4C\tilde{K}/T$ where C is a constant (only depends on K and σ_0 in Assumption 1) specified in Corollary 8.3 of van de Geer (2000). For $\xi \ge \xi_n$, Corollary 8.3 of van de Geer (2000) gives

$$\Pr\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_E\le 2^{s+1}\xi} |\langle \varepsilon, g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0}\rangle_n| \ge 2^{2s}\xi^2 T \right) \le C \exp\left(-\frac{n\xi^2 T^2 2^{2s}}{16 C^2 c_0^2} \right)$$

for $s = 0, \dots, S$. Thus, for $\xi \ge \xi_n$,

$$\Pr\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_E>\xi} \frac{|\langle \varepsilon, g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0}\rangle_n|}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_E^2} \ge T \right) \le c \exp\left(-\frac{n\xi^2 T^2}{c^2} \right),$$

for some constant c. □

Proof of Theorem 1. First, we derive a basic inequality. As $\hat{\boldsymbol{\theta}}_n$ minimizes M_n, we have $M_n(\hat{\boldsymbol{\theta}}_n) \le M_n(\boldsymbol{\theta}_0)$ which leads to

$$\|g_{\hat{\boldsymbol{\theta}}_n} - g_{\boldsymbol{\theta}_0}\|_n^2 \le 2\langle \delta_0, g_{\hat{\boldsymbol{\theta}}_n} - g_{\boldsymbol{\theta}_0}\rangle_n + 2\langle \varepsilon, g_{\hat{\boldsymbol{\theta}}_n} - g_{\boldsymbol{\theta}_0}\rangle_n. \tag{S2}$$

Next, we handle the first term in the right hand side of (S2). This term arises from the misspecification of the regression function, which results in non-mean-zero "errors" $(\delta_0(\boldsymbol{x}_i) + \varepsilon_i)$, when compared to typical least square estimation. Write the second derivative of $M(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_0$ as $A = A_1 - A_2$ where

$$A_1 = \int_{\mathcal{X}} g^{(1)}_{\boldsymbol{\theta}_0}(\boldsymbol{x}) g^{(1)}_{\boldsymbol{\theta}_0}(\boldsymbol{x})^{\intercal} dF(\boldsymbol{x}) \qquad \text{and} \qquad A_2 = \int_{\mathcal{X}} \delta_0(\boldsymbol{x}) g^{(2)}_{\boldsymbol{\theta}_0}(\boldsymbol{x}) dF(\boldsymbol{x}).$$

From the identification assumption (Assumption 5), A is strictly positive definite. By Taylor expansion, we also have, for $\boldsymbol{\theta} \in \Theta$ close to $\boldsymbol{\theta}_0$ and $\boldsymbol{x} \in \mathcal{X}$,

$$g_{\boldsymbol{\theta}}(\boldsymbol{x}) = g_{\boldsymbol{\theta}_0}(\boldsymbol{x}) + g^{(1)}_{\boldsymbol{\theta}_0}(\boldsymbol{x})^{\intercal}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\intercal} g^{(2)}_{\boldsymbol{\theta}_0}(\boldsymbol{x})(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \gamma_{\boldsymbol{\theta}}(\boldsymbol{x}), \tag{S3}$$

where

$$\gamma_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\intercal}\{g^{(2)}_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{x}) - g^{(2)}_{\boldsymbol{\theta}_0}(\boldsymbol{x})\}(\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

Here $\tilde{\boldsymbol{\theta}}$ lies between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. By Assumptions 3(c) and 6(a), and Lemma 1, we have

$$\langle \delta_0, g_{\hat{\boldsymbol{\theta}}_n} - g_{\boldsymbol{\theta}_0} \rangle_n = \boldsymbol{c}_n^{\intercal}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\intercal} A_{2,n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E^2),$$

where $\boldsymbol{c}_n = (1/n)\sum_{i=1}^n \delta_0(\boldsymbol{x}_i) g^{(1)}_{\boldsymbol{\theta}_0}(\boldsymbol{x}_i)$ and $A_{2,n} = (1/n)\sum_{i=1}^n \delta_0(\boldsymbol{x}_i) g^{(2)}_{\boldsymbol{\theta}_0}(\boldsymbol{x}_i)$. Moreover, by Assumption 3(c) and Lemma 1,

$$\|g_{\hat{\boldsymbol{\theta}}_n} - g_{\boldsymbol{\theta}_0}\|_n^2 = (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\intercal} A_{1,n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E^2),$$

where $A_{1,n} = (1/n)\sum_{i=1}^n g^{(1)}_{\boldsymbol{\theta}_0}(\boldsymbol{x}_i) g^{(1)}_{\boldsymbol{\theta}_0}(\boldsymbol{x}_i)^{\intercal}$. Elements of $A_{1,n} - A_{2,n}$ converge in probability to those of $A_1 - A_2$ due to Assumption 4(b). Hence (S2) implies

$$a\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E^2 \le (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\intercal} A(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \le o_p(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E^2) + 2\boldsymbol{c}_n^{\intercal}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + 2\langle \varepsilon, g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0} \rangle_n,$$

where a is the smallest eigenvalue of A. Since A is strictly positive definite, $a > 0$. By Cauchy-Schwarz inequality,

$$\frac{a}{2} + o_p(1) \le \frac{\boldsymbol{c}_n^{\intercal}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \langle \varepsilon, g_{\hat{\boldsymbol{\theta}}_n} - g_{\boldsymbol{\theta}_0} \rangle_n}{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E^2} \le \frac{\|\boldsymbol{c}_n\|_E}{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E} + \frac{|\langle \varepsilon, g_{\hat{\boldsymbol{\theta}}_n} - g_{\boldsymbol{\theta}_0} \rangle_n|}{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E^2},$$

Therefore, either

$$\frac{a}{4} + o_p(1) \le \frac{|\langle \varepsilon, g_{\hat{\boldsymbol{\theta}}_n} - g_{\boldsymbol{\theta}_0} \rangle_n|}{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E^2}, \tag{S4}$$

or

$$\frac{a}{4} + o_p(1) \le \frac{\|\boldsymbol{c}_n\|_E}{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E}. \tag{S5}$$

Let $\mathcal{E}_{1,n}$ be the event that (S4) occurs. On $\mathcal{E}_{1,n}$, $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E > \xi$ implies

$$\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_E > \xi} \frac{|\langle \varepsilon, g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0} \rangle_n|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_E^2} \ge \frac{a}{4} + o_p(1).$$

By Lemma 2, for sufficiently large n and $\xi \geq \xi_n$,

$$\Pr\left(\{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E > \xi\} \cap \mathcal{E}_{1,n}\right) \leq \Pr\left(\sup_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_E > \xi} \frac{|\langle \varepsilon, g_{\boldsymbol{\theta}} - g_{\boldsymbol{\theta}_0}\rangle_n|}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_E^2} \geq \frac{a}{8}\right) \leq c\exp\left(-\frac{na^2\xi^2}{64c^2}\right), \quad \text{(S6)}$$

where $\sqrt{n}\xi_n \geq 64\tilde{C}/a$.

Let $\mathcal{E}_{2,n}$ be the event that (S5) occurs. For sufficiently large n,

$$\Pr\left(\{\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E > \xi\} \cap \mathcal{E}_{2,n}\right) \leq \Pr\left(\|\boldsymbol{c}_n\|_E \geq \frac{a\xi}{8}\right). \quad \text{(S7)}$$

Combining (S6), (S7) and Assumption 4(c), we have $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_E = \mathcal{O}_p(n^{-1/2})$. By Assumption 3(a), $\|g_{\hat{\boldsymbol{\theta}}_n} - g_{\boldsymbol{\theta}_0}\|_n = \mathcal{O}_p(n^{-1/2})$. □

Lemma 3. *Assume that ε_i's are uniformly sub-Gaussian random variables and z_n is a function of $\boldsymbol{x} \in \mathcal{X}$ such that $\|z_n\|_n = \mathcal{O}_p(n^{-1/2})$. Moreover, assume that Assumption 6(b-d) holds. Let*

$$\tilde{\delta}_n = \arg\min_{\delta\in\mathcal{H}}\left[\frac{1}{n}\sum_{i=1}^{n}\{\tilde{y}_i - \delta(\boldsymbol{x}_i)\}^2 + \lambda_n^2 J^v(\delta)\right], \quad \text{(S8)}$$

$\tilde{y}_i = \delta_0(\boldsymbol{x}_i) + z_n(\boldsymbol{x}_i) + \varepsilon_i$ for $i = 1,\ldots,n$, with $v > (2\alpha)/(2+\alpha)$.

(i) If $J(\delta_0) > 0$ and $\lambda_n \asymp n^{-1/(2+\alpha)}$, we have

$$\|\tilde{\delta}_n - \delta_0\|_n = \mathcal{O}_p\left(n^{-1/(2+\alpha)}\right).$$

(ii) If $J(\delta_0) = 0$ and $J(\delta) > 0$ for all $\delta \in \mathcal{H}\setminus\{\delta_0\}$, we have

$$\|\tilde{\delta}_n - \delta_0\|_n = \mathcal{O}_p\left(\max\left\{n^{-1/2}, \lambda_n^{-2\alpha/(2v-2\alpha+v\alpha)}n^{-v/(2v-2\alpha+v\alpha)}\right\}\right).$$

Proof of Lemma 3. The proof is similar to that of Theorem 10.2 in van de Geer (2000), with modification due to the contamination z_n. Since $\tilde{\delta}_n$ minimizes (S8),

$$\begin{aligned}\|\tilde{\delta}_n - \delta_0\|_n^2 + \lambda_n^2 J^v(\tilde{\delta}_n) &\leq \langle \varepsilon, \tilde{\delta}_n - \delta_0\rangle_n + \langle z_n, \tilde{\delta}_n - \delta_0\rangle_n + \lambda_n^2 J^v(\delta_0)\\ &\leq \langle \varepsilon, \tilde{\delta}_n - \delta_0\rangle_n + \mathcal{O}_p(n^{-1/2})\|\tilde{\delta}_n - \delta_0\|_n + \lambda_n^2 J^v(\delta_0), \quad \text{(S9)}\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz inequality and $\|z_n\|_n = \mathcal{O}_p(n^{-1/2})$. By Lemma 8.4 of van de Geer (2000), under Assumption 6(b-d), we have

$$\sup_{\delta\in\bar{\mathcal{H}}} \frac{|\langle \varepsilon, \delta - \delta_0\rangle_n|}{\|\delta - \delta_0\|_n^{1-\alpha/2}\{J(\delta) + J(\delta_0)\}^{\alpha/2}} = \mathcal{O}_p(n^{-1/2}), \quad \text{(S10)}$$

where $\bar{\mathcal{H}} = \{\delta \in \mathcal{H} : J(\delta) + J(\delta_0) > 0\}$. Next, we analyze (S9) under various cases. To improve readibility, we underline the corresponding rates of convergences derived from each (sub)case.

Case (i): Suppose $J(\tilde{\delta}_n) > J(\delta_0)$. Thus $J(\tilde{\delta}_n) > 0$ and $\tilde{\delta}_n \in \bar{\mathcal{H}}$. We study the following two subcases, (a) $J(\delta_0) = 0$ and (b) $J(\delta_0) > 0$, seperately.

Case (i)(a): For $J(\delta_0) = 0$, using (S10), (S9) becomes

$$\|\tilde{\delta}_n - \delta_0\|_n^2 + \lambda_n^2 J^v(\tilde{\delta}_n) \leq \mathcal{O}_p(n^{-1/2})\|\tilde{\delta}_n - \delta_0\|_n^{1-\alpha/2} J^{\alpha/2}(\tilde{\delta}_n) + \mathcal{O}_p(n^{-1/2})\|\tilde{\delta}_n - \delta_0\|_n.$$

Either

$$\|\tilde{\delta}_n - \delta_0\|_n^2 + \lambda_n^2 J^v(\tilde{\delta}_n) \leq \mathcal{O}_p(n^{-1/2})\|\tilde{\delta}_n - \delta_0\|_n, \tag{S11}$$

or

$$\|\tilde{\delta}_n - \delta_0\|_n^2 + \lambda_n^2 J^v(\tilde{\delta}_n) \leq \mathcal{O}_p(n^{-1/2})\|\tilde{\delta}_n - \delta_0\|_n^{1-\alpha/2} J^{\alpha/2}(\tilde{\delta}_n). \tag{S12}$$

Solving (S11) and (S12) yield

$$\|\tilde{\delta}_n - \delta_0\|_n = \mathcal{O}_p(n^{-1/2}) \quad \text{and} \quad \|\tilde{\delta}_n - \delta_0\|_n = \mathcal{O}_p(\lambda_n^{-2\alpha/(2v-2\alpha+v\alpha)} n^{-v/(2v-2\alpha+v\alpha)})$$

respectively. So

$$\underline{\|\tilde{\delta}_n - \delta_0\|_n = \mathcal{O}_p\Big(\max\Big\{n^{-1/2}, \lambda_n^{-2\alpha/(2v-2\alpha+v\alpha)} n^{-v/(2v-2\alpha+v\alpha)}\Big\}\Big)}.$$

Case (i)(b): Suppose $J(\delta_0) > 0$. Using (S10), (S9) becomes

$$\begin{aligned}\|\tilde{\delta}_n - \delta_0\|_n^2 + \lambda_n^2 J^v(\tilde{\delta}_n) \leq \mathcal{O}_p(n^{-1/2})\|\tilde{\delta}_n - \delta_0\|_n^{1-\alpha/2} J^{\alpha/2}(\tilde{\delta}_n) \\ + \lambda_n^2 J^v(\delta_0) + \mathcal{O}_p(n^{-1/2})\|\tilde{\delta}_n - \delta_0\|_n.\end{aligned} \tag{S13}$$

Let $\mathcal{A}_n$ be the event that the last term of (S13) is the largest term of the right hand side of (S13). On $\mathcal{A}_n$, we have $\lambda_n^2 \leq \mathcal{O}_p(n^{-1/2})\|\tilde{\delta}_n - \delta_0\|_n$ and

$$\|\tilde{\delta}_n - \delta_0\|_n^2 + \lambda_n^2 J^v(\tilde{\delta}_n) \leq \mathcal{O}_p(n^{-1/2})\|\tilde{\delta}_n - \delta_0\|_n,$$

which leads to $\|\tilde{\delta}_n - \delta_0\|_n \leq \mathcal{O}_p(n^{-1/2})$. Together, we have $\lambda_n \leq \mathcal{O}_p(n^{-1/2})$. However, we assume $\lambda_n \asymp n^{-1/(2+\alpha)}$ for the case of $J(\delta_0) > 0$. Thus, $\Pr(\mathcal{A}_n) \leq \Pr(\lambda_n \leq \mathcal{O}_p(n^{-1/2})) \to 0$ as $n \to \infty$. By focusing on $\mathcal{A}_n^c$,

$$\underline{\|\tilde{\delta}_n - \delta_0\|_n = \mathcal{O}_p(n^{-1/(2+\alpha)})}$$

follows from the same arguments of Theorem 10.2 of van de Geer (2000).

Case (ii): Suppose $J(\tilde{\delta}_n) \leq J(\delta_0)$. Again, we study the two subcases, (a) $J(\delta_0) = 0$ and (b) $J(\delta_0) > 0$, separately.

Case (ii)(a): If $J(\delta_0) = 0$, we assume that $J(\delta) > 0$ for all $\delta \in \mathcal{H} \setminus \{\delta_0\}$. Thus $\underline{\tilde{\delta}_n = \delta_0}$.

Case (ii)(b): If $J(\delta_0) > 0$, we utilize (S9) and (S10) to obtain

$$\|\tilde{\delta}_n - \delta_0\|_n^2 \leq \mathcal{O}_p(n^{-1/2})\|\tilde{\delta}_n - \delta_0\|_n^{1-\alpha/2} J^{\alpha/2}(\delta_0) + \lambda_n^2 J^v(\delta_0) + \mathcal{O}_p(n^{-1/2})\|\tilde{\delta}_n - \delta_0\|_n. \tag{S14}$$

Let $\mathcal{B}_n$ be the event that the last term of (S14) is the largest term of the right hand side of (S14). Using similar argument for $\mathcal{A}_n$, we can show that $\Pr(\mathcal{B}_n) \to 0$ as $n \to \infty$. By focusing on $\mathcal{B}_n^c$,

$$\underline{\|\tilde{\delta}_n - \delta_0\|_n = \mathcal{O}_p(n^{-1/(2+\alpha)})}$$

follows from the arguments of Theorem 10.2 of van de Geer (2000).

The proof is completed by collecting the dominating terms for the two different scenarios, $J(\delta_0) > 0$ (cases (i)(b) and (ii)(b)) and $J(\delta_0) = 0$ (cases (i)(a) and (ii)(a)), seperately. □

Proof of Theorem 2. This follows from Theorem 1 and Lemma 3. □

Proof of Corollary 1. The key idea is the same as Section 10.1.1 of van de Geer (2000). Similar to Example 9.3.2 of van de Geer (2000), define $\psi_k(x) = x^{k-1}$ for $0 \leq x \leq 1$, $k = 1, \ldots, m$; $\beta_u = \delta^{(m)}(u)$ and $\tilde{\phi}_u(x) = \phi_u(x) - \bar{\phi}_u(x)$ for $0 < u \leq 1$, where

$$\phi_u(x) = \frac{(x-u)^{m-1}}{(m-1)!} 1\{u \leq x\} \quad \text{and} \quad \bar{\phi}_u(x) = \sum_{k=1}^{m} \gamma_{k,u}\psi_k(x)$$

such that $\langle \tilde{\phi}_u, \psi_k \rangle_n = 0$ for $k = 1, \ldots, m$. That means

$$\begin{pmatrix} \langle \phi_u, \psi_1 \rangle_n \\ \langle \phi_u, \psi_2 \rangle_n \\ \vdots \\ \langle \phi_u, \psi_m \rangle_n \end{pmatrix} = \Sigma_n \begin{pmatrix} \gamma_{1,u} \\ \gamma_{2,u} \\ \vdots \\ \gamma_{m,u} \end{pmatrix}, \tag{S15}$$

where $\Sigma_n = \int \boldsymbol{\psi}\boldsymbol{\psi}^\intercal dF_n$. For any $\delta \in \mathcal{H}$, we can decompose $\delta = \delta_1 + \delta_2$ (orthogonally via inner product $\langle \cdot, \cdot \rangle_n$) where

$$\delta_1 = \sum_{k=1}^{m} \alpha_k \psi_k(x) \in \mathcal{H}_1 \quad \text{and} \quad \delta_2 = \int_0^1 \beta_u \tilde{\phi}_u du \in \mathcal{H}_2.$$

Here, $\mathcal{H}_1 = \{\sum_{k=1}^m \alpha_k \psi_k : \alpha_k \in \mathbb{R}\}$ and $\mathcal{H}_2 = \{\gamma \in \mathcal{H} : \langle \gamma, \psi_k \rangle_n = 0, k = 1, \ldots, m\}$. Note that J does not depend on δ_1. That means $J(\delta) = J(\delta_2)$.

We can then show that $\hat{\delta}_n$ can be estimated via two separate estimations. Write the least squares criterion in (5) as

$$\|y - g_{\hat{\boldsymbol{\theta}}_n} - \delta\|_n^2 = \|y - g_{\hat{\boldsymbol{\theta}}_n} - \delta_0\|_n^2 + \|\delta_0 - \delta\|_n^2 + 2\langle y - g_{\hat{\boldsymbol{\theta}}_n} - \delta_0, \delta_0 - \delta \rangle_n.$$

Here, the first term is a constant with respect to δ. The second term can be written as

$$\|\delta_0 - \delta\|_n^2 = \|\delta_{0,1} - \delta_1\|_n^2 + \|\delta_{0,2} - \delta_2\|_n^2$$

where $\delta_0 = \delta_{0,1} + \delta_{0,2}$ and $\delta = \delta_1 + \delta_2$ with $\delta_{0,1}, \delta_1 \in \mathcal{H}_1$ and $\delta_{0,2}, \delta_2 \in \mathcal{H}_2$. In addition,

$$\langle y - g_{\hat{\boldsymbol{\theta}}_n} - \delta_0, \delta_0 - \delta \rangle_n = \langle \varepsilon + g_{\boldsymbol{\theta}_0} - g_{\hat{\boldsymbol{\theta}}_n}, \delta_{0,1} - \delta_1 \rangle_n + \langle \varepsilon + g_{\boldsymbol{\theta}_0} - g_{\hat{\boldsymbol{\theta}}_n}, \delta_{0,2} - \delta_2 \rangle_n.$$

The estimator can be written as $\hat{\delta}_n = \hat{\delta}_{1,n} + \hat{\delta}_{2,n}$, where

$$\begin{aligned}
\hat{\delta}_{1,n} &= \arg\min_{\delta_1 \in \mathcal{H}_1} \left\{ \|\delta_1 - \delta_{0,1}\|_n^2 - 2\langle \varepsilon + g_{\boldsymbol{\theta}_0} - g_{\hat{\boldsymbol{\theta}}_n}, \delta_1 - \delta_{0,1} \rangle_n \right\}, \\
\hat{\delta}_{2,n} &= \arg\min_{\delta_2 \in \mathcal{H}_2} \left\{ \|\delta_2 - \delta_{0,2}\|_n^2 - 2\langle \varepsilon + g_{\boldsymbol{\theta}_0} - g_{\hat{\boldsymbol{\theta}}_n}, \delta_2 - \delta_{0,2} \rangle_n + \lambda_n^2 J^2(\delta_2) \right\}.
\end{aligned}$$

As for $\hat{\delta}_{1,n}$, by Theorem 1, we have

$$\begin{aligned}
\|\hat{\delta}_{1,n} - \delta_{0,1}\|_n^2 &\leq 2\langle \varepsilon, \hat{\delta}_{1,n} - \delta_{0,1} \rangle_n + 2\langle g_{\boldsymbol{\theta}_0} - g_{\hat{\boldsymbol{\theta}}_n}, \hat{\delta}_{1,n} - \delta_{0,1} \rangle_n \\
&\leq 2\langle \varepsilon, \hat{\delta}_{1,n} - \delta_{0,1} \rangle_n + \mathcal{O}_p(n^{-1/2}) \|\hat{\delta}_{1,n} - \delta_{0,1}\|_n
\end{aligned}$$

Thus, either

$$\|\hat{\delta}_{1,n} - \delta_{0,1}\|_n^2 \leq \mathcal{O}_p(n^{-1/2}) \|\hat{\delta}_{1,n} - \delta_{0,1}\|_n, \tag{S16}$$

or

$$\|\hat{\delta}_{1,n} - \delta_{0,1}\|_n^2 \leq 4\langle \varepsilon, \hat{\delta}_{1,n} - \delta_{0,1} \rangle_n. \tag{S17}$$

The first inequality (S16) results in $\|\hat{\delta}_{1,n} - \delta_{0,1}\|_n = \mathcal{O}_p(n^{-1/2})$. Note that $\mathcal{H}_1$ is just the function space for a linear regression, we can obtain its entropy result from Example 9.3.1 of van de Geer (2000). Then the second inequalty (S17) also leads to $\|\hat{\delta}_{1,n} - \delta_{0,1}\|_n = \mathcal{O}_p(n^{-1/2})$ by the peeling device, similarly as in Lemma 2, and the arguments in Theorem 1 (see (S6)). Since the argument is typical (see, e.g., the proofs of Lemma 3.4 and Theorem 4 of van de Geer (1990), and Theorem 9.1 of van de Geer (2000)), we skip the details.

As for $\hat{\delta}_{2,n}$, we apply Lemma 3. Note that with smallest eigenvalue of $\int \boldsymbol{\psi}\boldsymbol{\psi}^\intercal dF_n$ bounded away from zero, Assumption 6(c) is fulfilled for $\mathcal{H}_2$ (Mammen, 1991; van de Geer, 2000, pp. 171) with $\alpha = 1/m$. We now focus on Assumption 6(d). For any $\delta_2 \in \mathcal{H}_2$,

$$\begin{aligned}
|\delta_2(x)| = \left|\int_0^1 \beta_u \tilde{\phi}_u(x)du\right| &= \left|\int_0^1 \beta_u \phi_u(x)du - \int_0^1 \beta_u \bar{\phi}_u(x)du\right| \\
&\leq \left|\int_0^1 \beta_u \phi_u(x)du\right| + \left|\int_0^1 \beta_u \bar{\phi}_u(x)du\right|.
\end{aligned}$$

First, there exists $K_2 < \infty$ such that for all $0 \leq x \leq 1$, $|\int_0^1 \beta_u \phi_u(x)du| \leq \sqrt{\int_0^1 \beta_u^2 du \int_0^1 \phi_u^2(x)du} \leq K_2 J(\delta_2)$. As for the second term, for $0 \leq x \leq 1$,

$$\begin{aligned}
\left|\int_0^1 \beta_u \bar{\phi}_u(x)du\right| &= \left|\int_0^1 \beta_u \sum_{k=1}^m \gamma_{k,u}\psi_k(x)du\right| \\
&\leq \sum_{k=1}^m |\psi_k(x)| \left|\int_0^1 \beta_u \gamma_{k,u} du\right| \\
&\leq \sum_{k=1}^m \left|\int_0^1 \beta_u \gamma_{k,u} du\right|,
\end{aligned}$$

where the last inequality follows from $|\psi_k(x)| = |x^{k-1}| \leq 1$. By (S15),

$$\left|\int_0^1 \beta_u \gamma_{k,u} du\right| \leq \sum_{l=1}^m B \left|\left\langle \int_0^1 \beta_u \phi_u, \psi_l \right\rangle_n\right| \leq \sum_{k=1}^m B K_2 J(\delta_2) \|\psi_l\|_n \leq m B K_2 J(\delta_2),$$

where $B < \infty$ is the max norm of Σ_n^{-1}. Note that the existence of Σ_n^{-1} is garanteed by that the smallest eigenvalue of Σ_n is bounded away from zero. Thus $\sup_x |\delta_2(x)| \leq (m^2 B + 1) K_2 J(\delta_2)$ which implies Assumption 6(d). Moreover, this result also implies that $J(\delta_2) > 0$ for all $\delta_2 \in \mathcal{H}_2 \setminus \{0\}$ and $\delta_{0,2} \equiv 0$ if $J(\delta_0) = J(\delta_{0,2}) = 0$. Then the corollary follows. □

References

Gu, C. (2014). Smoothing spline anova models: R package gss. *Journal of Statistical Software 58*, 1–25.

Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *The Annals of Statistics 19*, 741–759.

McKay, M., R. Beckman, and W. Conover (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics 21*, 239–245.

Park, J. S. (1991). *Turning complex computer codes to data and optimal designs.* Ph. D. thesis, University of Illlinois, Champaign-Urbana.

Storlie, C. B., W. A. Lane, E. M. Ryan, J. R. Gattiker, and D. M. Higdon (2014). Calibration of computational models with categorical parameters and correlated outputs via Bayesian smoothing spline ANOVA. *Journal of the American Statistical Association*. To appear.

van de Geer, S. (1990). Estimating a regression function. *The Annals of Statistics 18*(2), 907–924.

van de Geer, S. (2000). *Empirical Processes in M-estimation.* New York: Cambridge University Press.

van der Vaart, A. W. (2000). *Asymptotic Statistics.* New York: Cambridge University Press.