Submitted to *INFORMS Journal on Computing*
manuscript (Please, provide the manuscript number!)

Online Supplement to "Convergence Analysis of Stochastic Kriging-Assisted Simulation with Random Covariates"

Cheng Li
Department of Statistics and Data Science, National University of Singapore, Singapore 117546, stalic@nus.edu.sg

Siyang Gao
Department of Advanced Design and Systems Engineering and School of Data Science, City University of Hong Kong, Hong Kong, siyangao@cityu.edu.hk

Jianzhong Du
School of Management, Fudan University, China, jianzhodu2-c@my.cityu.edu.hk

This document provides supplementary materials for the paper "Convergence Analysis of Stochastic Kriging-Assisted Simulation with Random Covariates".

1. Notation

We summarize the key notation used in this paper in the following table.

Table 1: Table of notation.

Symbol	i folded[1]	Meaning
$\|\cdot\|$		Euclidean norm of a vector
$\|\|\|\cdot\|\|\|$		operator norm of a matrix, defined as $\sup_{\|\mathbf{v}\|=1} \|\cdot \mathbf{v}\|$
$\|\cdot\|_2^2$		L_2 norm of a function

[1]For simplicity of notation, in this research, we have folded the design index i in circumstances with no ambiguity. This column shows the symbol if its subscript i has been folded in the paper.

$a_l \lesssim b_l$		mean that $\limsup_{l\to\infty} a_l/b_l < \infty$
$a_l \asymp b_l$		mean that $a_l \lesssim b_l$ and $b_l \lesssim a_l$
k		number of system designs
d		dimension of the covariate space
m		number of covariate points
n		number of replications for each pair of covariate point and design
q		dimension of the regressors $\mathbf{f}$ and the regression coefficient $\boldsymbol{\beta}$
$\mathcal{X}$		support of covariate points
$y_i(\cdot)$	$y(\cdot)$	mean of design i
$Y_{il}(\cdot)$		the l-th simulation sample from design i
$\epsilon_{il}(\cdot)$		simulation noise of the l-th sample of design i
$\overline{\epsilon}_i(\cdot)$		averaged simulation errors, defined as $n^{-1}\sum_{l=1}^{n}\epsilon_{il}(\cdot)$
$\sigma_i^2(\cdot)$		variance of $\epsilon_{il}(\cdot)$
$\overline{Y}_i(\cdot)$	$\overline{Y}(\cdot)$	sample mean of design i
$\overline{\mathbf{Y}}_i$	$\overline{\mathbf{Y}}$	vector of samples means at m covariate points
$\mathbf{x}, \mathbf{X}$		vector of covariates with support $\mathcal{X} \subseteq \mathbb{R}^d$
$\mathbf{x}_0, \mathbf{X}_0$		test covariate point for the SK model
$\mathbf{x}_j, \mathbf{X}_j$		the j-th covariate point
$\mathbf{x}^m, \mathbf{X}^m$		vector of m covariate points
$\mathbf{f}_i(\cdot)$	$\mathbf{f}(\cdot)$	vector of known basis functions
$\boldsymbol{\beta}_i$	$\boldsymbol{\beta}$	vector of unknown parameters for $\mathbf{f}_i(\cdot)$
$M_i(\cdot)$	$M(\cdot)$	realization of a mean zero stationary Gaussian process for design i
$\boldsymbol{\Sigma}_{M,i}(\mathbf{x},\mathbf{x}')$	$\boldsymbol{\Sigma}_M(\mathbf{x},\mathbf{x}')$	covariance function, defined as $\mathrm{Cov}\left[M_i(\mathbf{x}), M_i(\mathbf{x}')\right]$
$\boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^m)$	$\boldsymbol{\Sigma}_{\epsilon}(\mathbf{x}^m)$	covariance matrix of the averaged simulation errors in design i
$\widehat{y}_i(\cdot)$	$\widehat{y}(\cdot)$	MSE-optimal linear predictor of the i-th SK model
$\mathrm{MSE}_{i,\mathrm{opt}}(\cdot)$	$\mathrm{MSE}_{\mathrm{opt}}(\cdot)$	the MSE of predictor $\widehat{y}_i(\cdot)$
$\lambda_{\max}(\cdot), \lambda_{\min}(\cdot)$		the largest and smallest eigenvalues of a matrix

$A_1 \prec A_2, A_2 \succ A_1$		mean that $A_2 - A_1$ is positive definite
$A_1 \preceq A_2, A_2 \succeq A_1$		mean that $A_2 - A_1$ is positive semi-definite
$\mathbb{1}(\cdot)$		indicator function
$\mathbb{P}_{\mathbf{X}}, \mathrm{E}_{\mathbf{X}}$		a probability distribution/expectation over $\mathcal{X}$
$L_2(\mathbb{P}_{\mathbf{X}})$		L_2 space under $\mathbb{P}_{\mathbf{X}}$
$\langle \cdot, \cdot \rangle_{L_2(\mathbb{P}_{\mathbf{X}})}$		inner product in $L_2(\mathbb{P}_{\mathbf{X}})$, defined as $\mathrm{E}_{\mathbf{X}}(\cdot)$
$[T_{\mathbf{\Sigma}_M} f](\mathbf{x})$		a linear operator of $\mathbf{x} \in \mathcal{X}$, defined as $\int_{\mathcal{X}} \mathbf{\Sigma}_M(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}')$
$\{\phi_{i,l}(\mathbf{x}) : l = 1, \dots\}$	$\phi_l(\mathbf{x})$	the orthonormal basis for $\mathbf{\Sigma}_{M,i}$ (from Mercer's theorem)
$\mathrm{tr}(\cdot)$		trace of a kernel (matrix)
$\{\mu_{i,l} : l = 1, \dots\}$	μ_l	eigenvalues of $\mathbf{\Sigma}_{M,i}$
$\mathbb{H}_i$	$\mathbb{H}$	reproducing kernel Hilbert space attached to $\mathbf{\Sigma}_{M,i}$
$\langle \cdot, \cdot \rangle_{\mathbb{H}}$		$\mathbb{H}$-inner product
ρ_*, r_*		parameters made for $\phi_{i,l}(\mathbf{x})$ in Assumption A.3
$\kappa_i, \nu_i, \tau_i, \varphi_i$	$\kappa, \nu, \tau, \varphi$	kernel parameters of the i-th SK model
κ_*, ν_*		parameters in rate functions, defined as $\kappa_* = \min_{i \in \{1,2,\dots,k\}} \kappa_i$ and $\nu_* = \min_{i \in \{1,2,\dots,k\} \nu_i}$
$\lesssim_{\mathbb{P}_{\mathbf{X}^m}}$		mean bounding in $\mathbb{P}_{\mathbf{X}^m}-$ probability
δ_0		indifference-zone parameter
$\underline{\sigma}_0^2, \overline{\sigma}_0^2$		lower and upper bounds for $\sigma_i^2(\mathbf{x})$ for all i and all $\mathbf{x} \in \mathcal{X}$
$i^\circ(\cdot), \widehat{i^\circ}(\cdot)$		the real and estimated optimal designs
$R^F(m,n)$		rate function of the maximal IMSE for finite-rank kernels
$R_i^E(m,n)$	$R^E(m,n)$	rate function of the maximal IMSE for exponentially decaying kernels and design i
$R_i^P(m,n)$	$R^P(m,n)$	rate function of the maximal IMSE for polynomially decaying kernels and design i
$R(m,n)$		rate function of IMSE
$\mu_i^{\mathcal{X}}$		expected mean performance of design i over the covariate space
$q_{i,\alpha}^{\mathcal{X}}$		α-quantile of the performance of design i over the covariate space

$w_i^{\mathcal{X}}$	proportion of design i being the best over the covariate space

2. Technical Proofs and Additional Theoretical Results

In this section, we first prove Theorems 1 to 5 in the main text. Next, we present a new theorem (Theorem 6) about the restrictiveness of Assumptions A.6 and A.7.

We reinstate some useful notation and relations. For any finite dimensional vector $\mathbf{v}$, we let $\|\mathbf{v}\|$ be its Euclidean norm. For any generic matrix A, we use A_{ab} to denote its (a,b)-entry, cA to denote the matrix whose (a,b)-entry is cA_{ab} for any constant $c \in \mathbb{R}$, and $|||A||| = \sup_{\|\mathbf{v}\|=1} \|A\mathbf{v}\|$ to denote its matrix operator norm. For any positive definite matrix A, let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ be its largest and smallest eigenvalues. For two positive definite matrices A_1, A_2, $A_1 \prec A_2$ and $A_2 \succ A_1$ mean that $A_2 - A_1$ is positive definite; $A_1 \preceq A_2$ and $A_2 \succeq A_1$ mean that $A_2 - A_1$ is positive semi-definite. For two sequences of positive numbers $\{a_l\}_{l\geq 1}$ and $\{b_l\}_{l\geq 1}$, $a_l \lesssim b_l$ means that $\limsup_{l\to\infty} a_l/b_l < \infty$, and $a_l \asymp b_l$ means that both $a_l \lesssim b_l$ and $b_l \lesssim a_l$ hold true. Let $\mathbb{1}(\cdot)$ be the indicator function and $\mathbf{I}_k$ be the $k \times k$ identity matrix.

Any function $f \in L_2(\mathbb{P}_{\mathbf{X}})$ has the series expansion $f(\mathbf{x}) = \sum_{l=1}^{\infty} \theta_l \phi_l(\mathbf{x})$, where $\theta_l = \langle f, \phi_l \rangle_{L_2(\mathbb{P}_{\mathbf{X}})}$. The L_2 norm of f is given by $\|f\|_2^2 = \sum_{l=1}^{\infty} \theta_l^2$. The reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ attached to $\mathbf{\Sigma}_M$ is the space of all functions $f \in L_2(\mathbb{P}_{\mathbf{X}})$ such that its $\mathbb{H}$-norm $\|f\|_{\mathbb{H}}^2 = \sum_{l=1}^{\infty} \theta_l^2/\mu_l < \infty$. For any two generic functions $\mathrm{h}_1, \mathrm{h}_2 \in \mathbb{H}$, let their $L_2(\mathbb{P}_{\mathbf{X}})$ expansions be $\mathrm{h}_s(\mathbf{x}) = \sum_{l=1}^{\infty} h_{sl}\phi_l(\mathbf{x})$ for $s = 1, 2$. Their $\mathbb{H}$-inner product is given by $\langle \mathrm{h}_1, \mathrm{h}_2 \rangle_{\mathbb{H}} = \sum_{l=1}^{\infty} h_{1l}h_{2l}/\mu_l$. For any $\mathrm{h} \in \mathbb{H}$, the reproducing property of $\mathbb{H}$ says that for any $\mathbf{x} \in \mathcal{X}$, $\langle \mathbf{\Sigma}_M(\mathbf{x}, \cdot), \mathrm{h}(\cdot) \rangle_{\mathbb{H}} = \mathrm{h}(\mathbf{x})$.

Proof of Theorem 1:

According to Mercer's theorem (e.g. Theorem 4.2 of Rasmussen and Williams 2006), the series expansion of the kernel function $\mathbf{\Sigma}_M(\mathbf{x}, \mathbf{x}') = \sum_{l=1}^{\infty} \mu_l \phi_l(\mathbf{x})\phi_l(\mathbf{x}')$ holds almost surely for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and hence

$$\mathbf{\Sigma}_M(\mathbf{x}_0, \mathbf{x}_0) = \sum_{a=1}^{\infty} \mu_a \phi_a^2(\mathbf{x}_0), \quad \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}_0) = \sum_{a=1}^{\infty} \mu_a \phi_a(\mathbf{x}_j)\phi_a(\mathbf{x}_0), \text{ for } j = 1, \dots, m,$$
$$\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}_0) = \left[\mathbf{\Sigma}_M(\mathbf{x}_1, \mathbf{x}_0), \dots, \mathbf{\Sigma}_M(\mathbf{x}_m, \mathbf{x}_0)\right]^\top. \tag{1}$$

Under the orthonormal property, if $\mathbf{X} \sim \mathbb{P}_{\mathbf{X}}$, then $\mathrm{E}_{\mathbf{X}}[\phi_a^2(\mathbf{X})] = 1$ and $\mathrm{E}_{\mathbf{X}}[\phi_a(\mathbf{X})\phi_b(\mathbf{X})] = 0$ for $a \neq b$. Therefore,

$$
\begin{aligned}
&\mathrm{E}_{\mathbf{X}^m}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] \\
&= \mathrm{E}_{\mathbf{X}_0}\left[\mathbf{\Sigma}_M(\mathbf{X}_0,\mathbf{X}_0)\right] - \mathrm{E}_{\mathbf{X}^m}\mathrm{E}_{\mathbf{X}_0}\left\{\mathbf{\Sigma}_M^\top(\mathbf{X}^m,\mathbf{X}_0)\left[\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}_0)\right\} \\
&\overset{(i)}{=} \sum_{a=1}^{\infty}\mu_a \mathrm{E}_{\mathbf{X}_0}\left[\phi_a^2(\mathbf{X}_0)\right] - \\
&\quad \mathrm{E}_{\mathbf{X}^m}\mathrm{E}_{\mathbf{X}_0}\sum_{j=1}^{m}\sum_{j'=1}^{m}\sum_{a=1}^{\infty}\sum_{b=1}^{\infty}\mu_a\mu_b\left\{\left[\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{X}_j)\phi_a(\mathbf{X}_0)\phi_b(\mathbf{X}_j')\phi_b(\mathbf{X}_0) \\
&\overset{(ii)}{=} \sum_{a=1}^{\infty}\mu_a \mathrm{E}_{\mathbf{X}_0}\left[\phi_a^2(\mathbf{X}_0)\right] - \mathrm{E}_{\mathbf{X}^m}\sum_{j=1}^{m}\sum_{j'=1}^{m}\sum_{a=1}^{\infty}\sum_{b=1}^{\infty}\mu_a\mu_b\left\{\left[\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'} \\
&\quad \cdot \phi_a(\mathbf{X}_j)\phi_b(\mathbf{X}_{j'})\mathrm{E}_{\mathbf{X}_0}\left[\phi_a(\mathbf{X}_0)\phi_b(\mathbf{X}_0)\right] \\
&= \sum_{a=1}^{\infty}\mu_a - \mathrm{E}_{\mathbf{X}^m}\sum_{j=1}^{m}\sum_{j'=1}^{m}\sum_{a=1}^{\infty}\mu_a^2\left\{\left[\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{X}_j)\phi_b(\mathbf{X}_{j'}) \\
&= \sum_{a=1}^{\zeta}\mu_a - \mathrm{E}_{\mathbf{X}^m}\sum_{a=1}^{\zeta}\sum_{j=1}^{m}\sum_{j'=1}^{m}\mu_a^2\left\{\left[\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{X}_j)\phi_a(\mathbf{X}_{j'}) \\
&\quad + \mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right) - \mathrm{E}_{\mathbf{X}^m}\sum_{a=\zeta+1}^{\infty}\sum_{j=1}^{m}\sum_{j'=1}^{m}\mu_a^2\left\{\left[\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{X}_j)\phi_a(\mathbf{X}_{j'}) \\
&\overset{(iii)}{\leq} \sum_{a=1}^{\zeta}\left\{\mu_a - \mathrm{E}_{\mathbf{X}^m}\sum_{j=1}^{m}\sum_{j'=1}^{m}\mu_a^2\left\{\left[\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{X}_j)\phi_a(\mathbf{X}_{j'})\right\} + \mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right).
\end{aligned}
\tag{2}
$$

In the derivation above, we exchange the expectation and the summation in several steps.

- For Step (i), because $\left\{\sum_{a=1}^{N}\mu_a\phi_a^2(\mathbf{X}_0),\quad N=1,2,\ldots\right\}$ is a non-decreasing sequence of functions, by the monotone convergence theorem, we have $\mathrm{E}_{\mathbf{X}_0}\left[\mathbf{\Sigma}_M(\mathbf{X}_0,\mathbf{X}_0)\right] = \mathrm{E}_{\mathbf{X}_0}\left[\sum_{a=1}^{\infty}\mu_a\phi_a^2(\mathbf{X}_0)\right] = \sum_{a=1}^{\infty}\mu_a\mathrm{E}_{\mathbf{X}_0}\left[\phi_a^2(\mathbf{X}_0)\right]$.
- For Step (ii), for any $\mathbf{x}^m$, every $j, j' = 1,\ldots,m$, and $N_1, N_2 = 1,2,\ldots$,

$$
\begin{aligned}
&\left|\sum_{a=1}^{N_1}\sum_{b=1}^{N_2}\mu_a\mu_b\left\{\left[\mathbf{\Sigma}_M(\mathbf{x}^m,\mathbf{x}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0)\right| \\
&\leq \sum_{a=1}^{N_1}\sum_{b=1}^{N_2}\mu_a\mu_b\left\{\left[\mathbf{\Sigma}_M(\mathbf{x}^m,\mathbf{x}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0) \\
&\quad \cdot \mathrm{sgn}\left(\left\{\left[\mathbf{\Sigma}_M(\mathbf{x}^m,\mathbf{x}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\right\}_{jj'}\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0)\right),
\end{aligned}
\tag{3}
$$

where $\mathrm{sgn}(x)=1$ for $x>0$, $\mathrm{sgn}(x)=-1$ for $x<0$, and $\mathrm{sgn}(x)=0$ if $x=0$. By Assumption A.3 and Hölder's inequality,

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}_0}\left\{\phi_a(\mathbf{X}_0)\phi_b(\mathbf{X}_0)\mathrm{sgn}\left(\left\{[\boldsymbol{\Sigma}_M(\mathbf{x}^m,\mathbf{x}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)]^{-1}\right\}_{jj'}\phi_a(\mathbf{x}_j)\phi_a(\mathbf{X}_0)\phi_b(\mathbf{x}_{j'})\phi_b(\mathbf{X}_0)\right)\right\}\\
&\le \mathrm{E}_{\mathbf{X}_0}\{|\phi_a(\mathbf{X}_0)\phi_b(\mathbf{X}_0)|\} \le \left(\mathrm{E}_{\mathbf{X}_0}\left\{\phi_a^2(\mathbf{X}_0)\right\}\right)^{1/2}\left(\mathrm{E}_{\mathbf{X}_0}\left\{\phi_b^2(\mathbf{X}_0)\right\}\right)^{1/2}\\
&\le \left(\mathrm{E}_{\mathbf{X}_0}\left\{\phi_a^{2r_*}(\mathbf{X}_0)\right\}\right)^{1/(2r_*)}\left(\mathrm{E}_{\mathbf{X}_0}\left\{\phi_b^{2r_*}(\mathbf{X}_0)\right\}\right)^{1/(2r_*)} \le \rho_*^2. \qquad (4)
\end{aligned}$$

We apply the dominated convergence theorem using (3) and (4) to obtain that

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}_0}\left\{\sum_{j=1}^m\sum_{j'=1}^m\sum_{a=1}^\infty\sum_{b=1}^\infty \mu_a\mu_b\left\{[\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)]^{-1}\right\}_{jj'}\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0)\right\}\\
&=\sum_{j=1}^m\sum_{j'=1}^m \mathrm{E}_{\mathbf{X}_0}\Bigg\{\lim_{N_1,N_2\to\infty}\sum_{a=1}^{N_1}\sum_{b=1}^{N_2}\mu_a\mu_b\left\{[\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)]^{-1}\right\}_{jj'}\\
&\qquad\cdot\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0)\Bigg\}\\
&=\sum_{j=1}^m\sum_{j'=1}^m \lim_{N_1,N_2\to\infty}\sum_{a=1}^{N_1}\sum_{b=1}^{N_2}\mu_a\mu_b\left\{[\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)]^{-1}\right\}_{jj'}\\
&\qquad\cdot\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\mathrm{E}_{\mathbf{X}_0}\left[\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0)\right]\\
&=\sum_{j=1}^m\sum_{j'=1}^m\sum_{a=1}^\infty\sum_{b=1}^\infty\mu_a\mu_b\left\{[\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)]^{-1}\right\}_{jj'}\phi_a(\mathbf{x}_j)\phi_b(\mathbf{x}_{j'})\mathrm{E}_{\mathbf{X}_0}\left[\phi_a(\mathbf{x}_0)\phi_b(\mathbf{x}_0)\right],
\end{aligned}$$

which gives the right-hand side of Step (ii).

- For Step (iii), we make the left-hand side larger by dropping the negative quadratic term in the summation $\sum_{a=\zeta+1}^\infty\sum_{j=1}^\infty\sum_{j'=1}^\infty$.

To proceed from (2), we define some useful quantities:

$$\begin{aligned}
&\mathbf{M}=\mathrm{diag}\left(\mu_1,\ldots,\mu_\zeta\right),\quad \mathbf{M}^{\mathrm{rem}}=\mathrm{diag}\left(\mu_{\zeta+1},\mu_{\zeta+2},\ldots\right),\\
&\boldsymbol{\phi}_a=\left[\phi_a(\mathbf{X}_1),\ldots,\phi_a(\mathbf{X}_m)\right]^\top,\ \text{for } a=1,2,\ldots,\\
&\boldsymbol{\Phi}=\left[\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_\zeta\right],\quad \boldsymbol{\Phi}^{\mathrm{rem}}=\left[\boldsymbol{\phi}_{\zeta+1},\boldsymbol{\phi}_{\zeta+2},\ldots\right],\\
&\mathbf{B}=\mathbf{M}-\mathbf{M}\boldsymbol{\Phi}^\top\left[\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\boldsymbol{\Phi}\mathbf{M},
\end{aligned}$$

such that $\boldsymbol{\Phi}$ is a $m\times\zeta$ matrix, and $\mathbf{B}$ is a $\zeta\times\zeta$ positive definite matrix. From this definition and (2), we have

$$\mathrm{tr}(\mathbf{B})=\sum_{a=1}^\zeta\mu_a-\sum_{a=1}^\zeta\sum_{j=1}^m\sum_{j'=1}^m\mu_a^2\left\{[\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)]^{-1}\right\}_{jj'}\phi_a(\mathbf{X}_j)\phi_b(\mathbf{X}_{j'}),$$

$$\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_{\mathbf{X}^m}\left[\mathrm{MSE}_{\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] \leq \mathrm{E}_{\mathbf{X}^m}\operatorname{tr}(\mathbf{B}) + \operatorname{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right). \tag{5}$$

Let $\boldsymbol{\Sigma}_M^{\mathrm{rem}} = \boldsymbol{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) - \boldsymbol{\Phi}\mathbf{M}\boldsymbol{\Phi}^\top = \boldsymbol{\Phi}^{\mathrm{rem}}\mathbf{M}^{\mathrm{rem}}\boldsymbol{\Phi}^{\mathrm{rem}\top}$, which is a $m \times m$ positive semi-definite matrix. Then by the Woodbury formula (Rasmussen and Williams 2006, Appendix A.3), the matrix $\mathbf{B}$ can be written as

$$\begin{aligned}\mathbf{B} &= \mathbf{M} - \mathbf{M}\boldsymbol{\Phi}^\top\left[\boldsymbol{\Sigma}_M(\mathbf{X}^m, \mathbf{X}^m) + \boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right]^{-1}\boldsymbol{\Phi}\mathbf{M} \\ &= \left[\mathbf{M}^{-1} + \boldsymbol{\Phi}^\top\left\{\boldsymbol{\Sigma}_M^{\mathrm{rem}} + \boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right\}^{-1}\boldsymbol{\Phi}\right]^{-1}.\end{aligned} \tag{6}$$

By Assumption A.1 and the definition of n, we have that $\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m)$ is diagonal and $\boldsymbol{\Sigma}_\epsilon(\mathbf{x}^m) \preceq \frac{\overline{\sigma}_0^2}{n}\mathbf{I}_m$ for any value of $\mathbf{x}^m$, where $\mathbf{I}_m$ is the $m \times m$ identity matrix. Therefore, from (6), we can apply the Woodbury formula again to obtain that

$$\begin{aligned}\mathbf{B} &\preceq \left[\mathbf{M}^{-1} + \boldsymbol{\Phi}^\top\left\{\boldsymbol{\Sigma}_M^{\mathrm{rem}} + \frac{\overline{\sigma}_0^2}{n}\mathbf{I}_m\right\}^{-1}\boldsymbol{\Phi}\right]^{-1} \\ &= \frac{\overline{\sigma}_0^2}{mn}\left[\mathbf{I}_\zeta + \frac{\overline{\sigma}_0^2}{mn}\mathbf{M}^{-1} + \frac{1}{m}\boldsymbol{\Phi}^\top\left(\frac{n}{\overline{\sigma}_0^2}\boldsymbol{\Sigma}_M^{\mathrm{rem}} + \mathbf{I}_m\right)^{-1}\boldsymbol{\Phi} - \mathbf{I}_\zeta\right]^{-1} \\ &= \frac{\overline{\sigma}_0^2}{mn}\mathbf{Q}^{-2}\left\{\mathbf{I}_\zeta + \mathbf{Q}^{-1}\left[\frac{1}{m}\boldsymbol{\Phi}^\top\left(\frac{n}{\overline{\sigma}_0^2}\boldsymbol{\Sigma}_M^{\mathrm{rem}} + \mathbf{I}_m\right)^{-1}\boldsymbol{\Phi} - \mathbf{I}_\zeta\right]\mathbf{Q}^{-1}\right\}^{-1},\end{aligned} \tag{7}$$

where $\mathbf{Q} = \left(\mathbf{I}_\zeta + \frac{\overline{\sigma}_0^2}{mn}\mathbf{M}^{-1}\right)^{1/2}$.

Define the event $\mathcal{E}_2 = \left\{\frac{n}{\overline{\sigma}_0^2}\boldsymbol{\Sigma}_M^{\mathrm{rem}} \preceq \delta_2\mathbf{I}_m\right\}$. Then since $\boldsymbol{\Sigma}_M^{\mathrm{rem}}$ is positive semi-definite, we have the relation that

$$\left\{\operatorname{tr}\left(\tfrac{n}{\overline{\sigma}_0^2}\boldsymbol{\Sigma}_M^{\mathrm{rem}}\right) \leq \delta_2\right\} \subseteq \left\{\lambda_{\max}\left(\tfrac{n}{\overline{\sigma}_0^2}\boldsymbol{\Sigma}_M^{\mathrm{rem}}\right) \leq \delta_2\right\} \subseteq \mathcal{E}_2.$$

Therefore, by Markov's inequality and the monotone convergence theorem, we have that

$$\begin{aligned}\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_2^c) &\leq \mathbb{P}_{\mathbf{X}^m}\left\{\operatorname{tr}\left(\tfrac{n}{\overline{\sigma}_0^2}\boldsymbol{\Sigma}_M^{\mathrm{rem}}\right) > \delta_2\right\} \leq \frac{1}{\delta_2}\mathrm{E}_{\mathbf{X}^m}\operatorname{tr}\left(\tfrac{n}{\overline{\sigma}_0^2}\boldsymbol{\Sigma}_M^{\mathrm{rem}}\right) \\ &= \frac{n}{\overline{\sigma}_0^2\delta_2}\sum_{i=1}^m\sum_{a=\zeta+1}^\infty \mu_a\mathrm{E}_{\mathbf{X}^m}\phi_a^2(\mathbf{X}_i) = \frac{mn}{\overline{\sigma}_0^2\delta_2}\operatorname{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right).\end{aligned} \tag{8}$$

On the other hand, we consider the event defined in Lemma 3 with $\delta = \delta_1$, i.e.

$$\mathcal{E}_1 = \left\{\left|\!\left|\!\left|\mathbf{Q}^{-1}\left(\frac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \mathbf{I}_\zeta\right)\mathbf{Q}^{-1}\right|\!\right|\!\right| \leq \delta_1\right\}.$$

On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have that

$$\begin{aligned}
&\mathbf{I}_\zeta + \mathbf{Q}^{-1}\left\{\frac{1}{m}\boldsymbol{\Phi}^\top\left(\tfrac{n}{\overline{\sigma}_0^2}\boldsymbol{\Sigma}_M^{\mathrm{rem}} + \mathbf{I}_m\right)^{-1}\boldsymbol{\Phi} - \mathbf{I}_\zeta\right\}\mathbf{Q}^{-1} \\
&\overset{(i)}{\succeq} \mathbf{I}_\zeta + \mathbf{Q}^{-1}\left\{\frac{1}{m}\boldsymbol{\Phi}^\top\left(\delta_2\mathbf{I}_m + \mathbf{I}_m\right)^{-1}\boldsymbol{\Phi}^\top - \mathbf{I}_\zeta\right\}\mathbf{Q}^{-1} \\
&= \mathbf{I}_\zeta - \left(1 - \frac{1}{1+\delta_2}\right)\mathbf{Q}^{-2} + \frac{1}{1+\delta_2}\mathbf{Q}^{-1}\left\{\frac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \mathbf{I}_\zeta\right\}\mathbf{Q}^{-1} \\
&\overset{(ii)}{\succeq} \mathbf{I}_\zeta - \left(1 - \frac{1}{1+\delta_2}\right)\mathbf{I}_\zeta - \frac{1}{1+\delta_2}\cdot\delta_1\mathbf{I}_\zeta = \frac{1-\delta_1}{1+\delta_2}\mathbf{I}_\zeta,
\end{aligned} \tag{9}$$

where (i) follows on the event $\mathcal{E}_2$, and (ii) holds on the event $\mathcal{E}_1$ and from the fact $\mathbf{Q}^{-2} \preceq \mathbf{I}_\zeta$.

Therefore, by combining (8), (9), and the upper bound for $\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_1^c)$ given in Lemma 3 under our assumptions A.1-A.3, we obtain that

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}^m}\operatorname{tr}(\mathbf{B}) \le \mathrm{E}_{\mathbf{X}^m}\left\{\operatorname{tr}(\mathbf{B})\mathbb{1}(\mathcal{E}_1\cap\mathcal{E}_2)\right\} + \mathrm{E}_{\mathbf{X}^m}\left[\operatorname{tr}(\mathbf{B})\left\{\mathbb{1}(\mathcal{E}_1^c) + \mathbb{1}(\mathcal{E}_2^c)\right\}\right] \\
&\overset{(i)}{\le} \frac{1+\delta_2}{1-\delta_1}\frac{\overline{\sigma}_0^2}{mn}\operatorname{tr}\left(\mathbf{Q}^{-2}\right) + \operatorname{tr}(\boldsymbol{\Sigma}_M)\left\{\mathbb{P}(\mathcal{E}_1^c) + \mathbb{P}(\mathcal{E}_2^c)\right\} \\
&\overset{(ii)}{\le} \frac{1+\delta_2}{1-\delta_1}\frac{\overline{\sigma}_0^2}{mn}\gamma\left(\frac{\overline{\sigma}_0^2}{mn}\right) + \frac{mn}{\overline{\sigma}_0^2\delta_2}\operatorname{tr}(\boldsymbol{\Sigma}_M)\operatorname{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right) + \operatorname{tr}(\boldsymbol{\Sigma}_M)\left\{100\rho_*^2\frac{b(m,\zeta,r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\delta_1\sqrt{m}}\right\}^{r_*},
\end{aligned} \tag{10}$$

where (i) follows from (9), and (ii) follows from (8), Lemma 3, and the fact that

$$\operatorname{tr}\left(\mathbf{Q}^{-2}\right) = \operatorname{tr}\left\{\left(\mathbf{I}_\zeta + \frac{\overline{\sigma}_0^2}{mn}\mathbf{M}^{-1}\right)^{-1}\right\} = \sum_{a=1}^{\zeta}\left(1 + \frac{\overline{\sigma}_0^2}{mn\mu_a}\right)^{-1} = \sum_{a=1}^{\zeta}\frac{\mu_a}{\mu_a + \frac{\overline{\sigma}_0^2}{mn}} \le \gamma\left(\frac{\overline{\sigma}_0^2}{mn}\right).$$

Finally, we combine (5) and (10) to obtain that

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}^m}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] \le \mathrm{E}_{\mathbf{X}^m}\operatorname{tr}(\mathbf{B}) + \operatorname{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right) \\
&\le \frac{1+\delta_2}{1-\delta_1}\frac{\overline{\sigma}_0^2}{mn}\gamma\left(\frac{\overline{\sigma}_0^2}{mn}\right) + \left\{\frac{mn}{\overline{\sigma}_0^2\delta_2}\operatorname{tr}(\boldsymbol{\Sigma}_M) + 1\right\}\operatorname{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right) + \operatorname{tr}(\boldsymbol{\Sigma}_M)\left\{100\rho_*^2\frac{b(m,\zeta,r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\delta_1\sqrt{m}}\right\}^{r_*}.
\end{aligned}$$

Taking the infimum with respect to ζ and setting $\delta_1 = \delta_2 = 1/3$ leads to the conclusion. □

Proof of Theorem 2:

We define some additional notation. For abbreviation, we write $\sigma_j^2 = \sigma^2(\mathbf{x}_j)$, $j = 1,\dots,m$. Let $\mathbf{F} = (\mathbf{f}(\mathbf{X}_1),\dots,\mathbf{f}(\mathbf{X}_m))^\top = (\mathrm{f}_1(\mathbf{X}^m),\dots,\mathrm{f}_q(\mathbf{X}^m))$ be the partition of $\mathbf{F}$ according to rows and columns, respectively. For the "bias" defined in (6) of the manuscript, let $\eta(\mathbf{x}) = (\eta_1(\mathbf{x}),\ldots,\eta_q(\mathbf{x}))^\top$ for any $\mathbf{x}\in\mathcal{X}$, where $\eta_s(\mathbf{x}) = \mathrm{f}_s(\mathbf{x}) - \mathrm{f}_s(\mathbf{x}^m)^\top(\mathbf{\Sigma}_M(\mathbf{x}^m,\mathbf{x}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{x}^m))^{-1}\mathbf{\Sigma}_M(\mathbf{x}^m,\mathbf{x})$. Since by Assumption A.4, $\mathrm{f}_s(\cdot)\in\mathbb{H}$ for each $s=1,\ldots,q$ and $\mathbf{\Sigma}(\mathbf{x}_j,\cdot)\in\mathbb{H}$ for each $j=1,\ldots,m$, we have that the function $\eta_s(\cdot)$ also lies in $\mathbb{H}$. In the following, we investigate and provide upper bound for $\|\eta_s\|_2$, $s=1,\ldots,q$. We first expand the function $\mathrm{f}_s(\mathbf{x})$ and $\eta_s(\mathbf{x})$ in terms of the orthonormal basis $\{\phi_l(\mathbf{x}): l=1,2,\ldots\}$:

$$\mathrm{f}_s(\mathbf{x}) = \sum_{l=1}^{\infty}\theta_{sl}\phi_l(\mathbf{x}),\qquad \eta_s(\mathbf{x}) = \sum_{l=1}^{\infty}\delta_{sl}\phi_l(\mathbf{x}), \tag{11}$$

for any $\mathbf{x}\in\mathcal{X}$ and $s=1,\ldots,q$. For a fixed $\zeta\in\mathbb{N}$, define $\theta_s^{\downarrow} = (\theta_{s1},\ldots,\theta_{s\zeta})^\top$, $\theta_s^{\uparrow} = (\theta_{s,\zeta+1},\theta_{s,\zeta+2},\ldots)^\top$, $\delta_s^{\downarrow} = (\delta_{s1},\ldots,\delta_{s\zeta})^\top$, $\delta_s^{\uparrow} = (\delta_{s,\zeta+1},\delta_{s,\zeta+2},\ldots)^\top$. We also define the following quantities:

$$\begin{aligned}
&\mathbf{M} = \mathrm{diag}\left(\mu_1,\ldots,\mu_\zeta\right),\\
&\boldsymbol{\phi}_l = \left[\phi_l(\mathbf{X}_1),\ldots,\phi_l(\mathbf{X}_m)\right]^\top,\ \text{for } l=1,2,\ldots,\\
&\mathbf{\Phi} = \left[\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_\zeta\right],\\
&\mathbf{v}_s = (v_{s1},\ldots,v_{sm})^\top,\quad v_{sj} = \sum_{l=\zeta+1}^{\infty}\delta_{sl}\phi_l(\mathbf{X}_j),\ \text{for } j=1,\ldots,m.
\end{aligned}$$

Then based on Assumptions A.1-A.4, we can prove Lemma 1 and Lemma 2. On the other hand, from the definition of $\mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{x}_0)$ in (6) of the manuscript, we have that

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] &= \mathrm{E}_{\mathbf{X}_0}\left[\eta(\mathbf{X}_0)^\top\left[\mathbf{F}^\top(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m))^{-1}\mathbf{F}\right]^{-1}\eta(\mathbf{X}_0)\right]\\
&\le \lambda_{\max}\left(\left[\mathbf{F}^\top(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m))^{-1}\mathbf{F}\right]^{-1}\right)\cdot\mathrm{E}_{\mathbf{X}_0}\left[\eta(\mathbf{X}_0)^\top\eta(\mathbf{X}_0)\right]\\
&= \left[\lambda_{\min}\left(\mathbf{F}^\top(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m))^{-1}\mathbf{F}\right)\right]^{-1}\cdot\mathrm{E}_{\mathbf{X}_0}\left[\sum_{s=1}^{q}\eta_s(\mathbf{X}_0)^\top\eta_s(\mathbf{X}_0)\right]\\
&= \left[\lambda_{\min}\left(\mathbf{F}^\top(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m))^{-1}\mathbf{F}\right)\right]^{-1}\cdot\sum_{s=1}^{q}\|\eta_s\|_2^2.
\end{aligned} \tag{12}$$

For simplicity, we define Γ_m to be the quantity inside the bracelets in Theorem 2:

$$\begin{aligned}
\Gamma_m = 8C_{\mathrm{f}}^2\frac{\overline{\sigma}_0^2}{mn} + \inf_{\zeta\in\mathbb{N}}\Bigg[&8C_{\mathrm{f}}^2\frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right) + C_{\mathrm{f}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\\
&+ C_{\mathrm{f}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\left\{200\rho_*^2\frac{b(m,\zeta,r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}\Bigg].
\end{aligned}$$

From the upper bound of $\mathrm{E}_{\mathbf{X}^m}\|\eta_s\|_2^2$ in Lemma 1, it is clear that $\mathrm{E}_{\mathbf{X}^m}\|\eta_s\|_2^2 \le \Gamma_m$ for all $s=1,\dots,q$ since we can make the upper bound in Lemma 1 larger by replacing each $\|\mathrm{f}_s\|_{\mathbb{H}}$ with C_{f}. From the Markov's inequality, for any $\xi \in (0,1/4)$,

$$\mathbb{P}_{\mathbf{X}^m}\left(\sum_{s=1}^{q}\|\eta_s\|_2^2 \ge q\Gamma_m/\xi\right) \le \frac{\sum_{s=1}^{q}\mathrm{E}_{\mathbf{X}^m}\|\eta_s\|_2^2}{q\Gamma_m/\xi} \le \frac{q\Gamma_m}{q\Gamma_m/\xi} = \xi. \tag{13}$$

Then from Lemma 2, we have that for any $\xi \in (0,1/4)$, for all $m > m_0$ (with m_0 dependent on $\xi, \mathbf{\Sigma}_M, \mathbf{f}, n, \overline{\sigma}_0^2, \rho_*$),

$$\mathbb{P}_{\mathbf{X}^m}\left(\left[\lambda_{\min}\left\{\mathbf{F}^\top\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathbf{F}\right\}\right]^{-1} > \frac{8\,\mathrm{tr}(\mathbf{\Sigma}_M)}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\right) < \xi. \tag{14}$$

We combine (12), (13) and (14) together to conclude that for any $\xi \in (0,1/4)$, for all $m > m_0$, there exists a constant $c_\xi = 1/\xi$, such that

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X}^m}\left(\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] > c_\xi \cdot \frac{8q\,\mathrm{tr}(\mathbf{\Sigma}_M)}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\Gamma_m\right)\\
&\le \mathbb{P}_{\mathbf{X}^m}\left(\left[\lambda_{\min}\left(\mathbf{F}^\top\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathbf{F}\right)\right]^{-1}\cdot\sum_{s=1}^{q}\|\eta_s\|_2^2 > \frac{8\,\mathrm{tr}(\mathbf{\Sigma}_M)}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\cdot\frac{q\Gamma_m}{\xi}\right)\\
&\le \mathbb{P}_{\mathbf{X}^m}\left(\sum_{s=1}^{q}\|\eta_s\|_2^2 \ge q\Gamma_m/\xi\right)\\
&\qquad + \mathbb{P}_{\mathbf{X}^m}\left(\left[\lambda_{\min}\left\{\mathbf{F}^\top\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\mathbf{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathbf{F}\right\}\right]^{-1} > \frac{8\,\mathrm{tr}(\mathbf{\Sigma}_M)}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\right)\\
&< \xi + \xi = 2\xi.
\end{aligned} \tag{15}$$

This has proved that $\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{8q\,\mathrm{tr}(\mathbf{\Sigma}_M)}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\Gamma_m$, which is the conclusion of Theorem 2. □

LEMMA 1. *Under Assumptions A.1-A.4, we have that for each $s=1,\dots,q$,*

$$\mathrm{E}_{\mathbf{X}^m}\|\eta_s\|_2^2 \le \frac{8\|\mathrm{f}_s\|_{\mathbb{H}}^2\overline{\sigma}_0^2}{mn} + \inf_{\zeta\in\mathbb{N}}\left[\frac{8\|\mathrm{f}_s\|_{\mathbb{H}}^2 mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right) + \|\mathrm{f}_s\|_{\mathbb{H}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\right.$$
$$\left. + \|\mathrm{f}_s\|_{\mathbb{H}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\left\{200\rho_*^2\frac{b(m,\zeta,r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}\right].$$

Proof of Lemma 1:

By Assumption A.1, we have that $\mathbf{\Sigma}_\epsilon(\mathbf{x}^m) = \operatorname{diag}(\sigma_1^2/n, \dots, \sigma_m^2/n)$, where we let $\sigma_j^2 = \sigma^2(\mathbf{x}_j)$ for $j = 1, \dots, m$. For any $\mathbf{x} \in \mathcal{X}$ and any $s \in \{1, \dots, q\}$, we have the following relation:

$$
\begin{aligned}
&\sum_{j=1}^m \frac{n}{\sigma_j^2} \eta_s(\mathbf{x}_j) \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}) \\
&= \sum_{j=1}^m \frac{n}{\sigma_j^2} \left\{ \mathrm{f}_s(\mathbf{x}_j) - \mathrm{f}_s(\mathbf{x}^m)^\top \left(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m) \right)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}_j) \right\} \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}) \\
&= \sum_{j=1}^m \frac{n}{\sigma_j^2} \mathrm{f}_s(\mathbf{x}_j) \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}) - \sum_{j=1}^m \frac{n}{\sigma_j^2} \mathrm{f}_s(\mathbf{x}^m)^\top \left(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m) \right)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}_j) \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}) \\
&= \mathrm{f}_s(\mathbf{x}^m)^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}) \\
&\quad - \mathrm{f}_s(\mathbf{x}^m)^\top \left(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m) \right)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}) \\
&= \mathrm{f}_s(\mathbf{x}^m)^\top \left(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m) \right)^{-1} \left\{ \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m) - \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) \right\} \\
&\quad \cdot \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}) \\
&= \mathrm{f}_s(\mathbf{x}^m)^\top \left(\mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}^m) + \mathbf{\Sigma}_\epsilon(\mathbf{x}^m) \right)^{-1} \mathbf{\Sigma}_M(\mathbf{x}^m, \mathbf{x}) \\
&= \mathrm{f}_s(\mathbf{x}) - \eta_s(\mathbf{x}).
\end{aligned} \tag{16}
$$

Therefore, we can rewrite (16) as

$$
\sum_{j=1}^m \frac{n}{\sigma_j^2} \eta_s(\mathbf{x}_j) \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}) + \eta_s(\mathbf{x}) - \mathrm{f}_s(\mathbf{x}) = 0, \tag{17}
$$

for any $\mathbf{x} \in \mathcal{X}$ and any $s \in \{1, \dots, q\}$.

We proceed with (17) in two ways. On one hand, we can take the $\mathbb{H}$-norm of f_s in (17). Since $\eta_s \in \mathbb{H}$ and it has the expansion in (11), we can derive from (17) that

$$
\begin{aligned}
\mathrm{f}_s(\mathbf{x}) &= \sum_{j=1}^m \frac{n}{\sigma_j^2} \sum_{a=1}^\infty \delta_{sa} \phi_a(\mathbf{x}_j) \sum_{b=1}^\infty \mu_b \phi_b(\mathbf{x}_j) \phi_b(\mathbf{x}) + \sum_{b=1}^\infty \delta_{sb} \phi_b(\mathbf{x}) \\
&= \sum_{b=1}^\infty \left\{ \mu_b \sum_{j=1}^m \frac{n}{\sigma_j^2} \sum_{a=1}^\infty \delta_{sa} \phi_a(\mathbf{x}_j) \phi_b(\mathbf{x}_j) + \delta_{sb} \right\} \phi_b(\mathbf{x}), \\
\|\mathrm{f}_s\|_{\mathbb{H}}^2 &= \sum_{b=1}^\infty \frac{1}{\mu_b} \left\{ \mu_b \sum_{j=1}^m \frac{n}{\sigma_j^2} \sum_{a=1}^\infty \delta_{sa} \phi_a(\mathbf{x}_j) \phi_b(\mathbf{x}_j) + \delta_{sb} \right\}^2 \\
&= \sum_{b=1}^\infty \frac{\delta_{sb}^2}{\mu_b} + 2 \sum_{b=1}^\infty \sum_{j=1}^m \frac{n}{\sigma_j^2} \sum_{a=1}^\infty \delta_{sa} \delta_{sb} \phi_a(\mathbf{x}_j) \phi_b(\mathbf{x}_j) + \sum_{b=1}^\infty \mu_b \left\{ \sum_{j=1}^m \frac{n}{\sigma_j^2} \sum_{a=1}^\infty \delta_{sa} \phi_a(\mathbf{x}_j) \phi_b(\mathbf{x}_j) \right\}^2 \\
&= \|\eta_s\|_{\mathbb{H}}^2 + 2 \sum_{j=1}^m \frac{n}{\sigma_j^2} \left\{ \sum_{a=1}^\infty \delta_{sa} \phi_a(\mathbf{x}_j) \right\}^2 + \sum_{b=1}^\infty \mu_b \left\{ \sum_{j=1}^m \frac{n}{\sigma_j^2} \sum_{a=1}^\infty \delta_{sa} \phi_a(\mathbf{x}_j) \phi_b(\mathbf{x}_j) \right\}^2
\end{aligned}
$$

$$
\begin{aligned}
&\geq \|\eta_s\|_{\mathbb{H}}^2, \\
\Longrightarrow\ & \|\eta_s\|_{\mathbb{H}} \leq \|\mathrm{f}_s\|_{\mathbb{H}}.
\end{aligned} \tag{18}
$$

On the other hand, we take $\mathbb{H}$-inner product of the left-hand-side of (17) with $\phi_l(\mathbf{x})$ for any fixed l with $\mu_l > 0$, and obtain that

$$
\begin{aligned}
0 &= \sum_{j=1}^{m} \frac{n}{\sigma_j^2} \eta_s(\mathbf{x}_j) \langle \mathbf{\Sigma}_M(\mathbf{x}_j, \mathbf{x}), \phi_l(\mathbf{x}) \rangle_{\mathbb{H}} + \langle \eta_s(\mathbf{x}), \phi_l(\mathbf{x}) \rangle_{\mathbb{H}} - \langle \mathrm{f}_s(\mathbf{x}), \phi_l(\mathbf{x}) \rangle_{\mathbb{H}}, \\
&= \sum_{j=1}^{m} \frac{n}{\sigma_j^2} \eta_s(\mathbf{x}_j) \phi_l(\mathbf{x}_j) + \frac{\delta_{sl}}{\mu_l} - \frac{\theta_{sl}}{\mu_l}, \\
&= \sum_{j=1}^{m} \frac{n}{\sigma_j^2} \sum_{a=1}^{\infty} \delta_{sa} \phi_a(\mathbf{x}_j) \phi_l(\mathbf{x}_j) + \frac{\delta_{sl}}{\mu_l} - \frac{\theta_{sl}}{\mu_l}, \\
&= \sum_{j=1}^{m} \sum_{a=1}^{\zeta} \frac{n}{\sigma_j^2} \delta_{sa} \phi_a(\mathbf{x}_j) \phi_l(\mathbf{x}_j) + \sum_{j=1}^{m} \frac{n}{\sigma_j^2} v_{sj} \phi_l(\mathbf{x}_j) + \frac{\delta_{sl}}{\mu_l} - \frac{\theta_{sl}}{\mu_l}
\end{aligned} \tag{19}
$$

where we have used the reproducing property for the function $\phi_l \in \mathbb{H}$. We can then stack (19) in a column for $l = 1, \ldots, \zeta$ for some $\zeta \in \mathbb{N}$ with $\mu_\zeta > 0$, and obtain that

$$
\begin{aligned}
& \mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Phi} \delta_s^{\downarrow} + \mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{v}_s + \mathbf{M}^{-1} \delta_s^{\downarrow} - \mathbf{M}^{-1} \theta_s^{\downarrow} = 0, \\
\Longrightarrow\ & \delta_s^{\downarrow} = \left(\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Phi} + \mathbf{M}^{-1} \right)^{-1} \left(\mathbf{M}^{-1} \theta_s^{\downarrow} - \mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{v}_s \right), \\
& \quad = \left(\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Phi} + \mathbf{M}^{-1} \right)^{-1} \mathbf{Q} \left(\mathbf{Q}^{-1} \mathbf{M}^{-1} \theta_s^{\downarrow} - \mathbf{Q}^{-1} \mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{v}_s \right),
\end{aligned} \tag{20}
$$

where $\mathbf{Q} = \left(\mathbf{I}_\zeta + \frac{\overline{\sigma}_0^2}{mn} \mathbf{M}^{-1} \right)^{1/2}$ as defined in Lemma 3. Therefore,

$$
\left\| \delta_s^{\downarrow} \right\| \leq \left\| \left(\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Phi} + \mathbf{M}^{-1} \right)^{-1} \mathbf{Q} \right\| \left(\left\| \mathbf{Q}^{-1} \mathbf{M}^{-1} \theta_s^{\downarrow} \right\| + \left\| \mathbf{Q}^{-1} \mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{v}_s \right\| \right). \tag{21}
$$

By Assumption A.1, we have that $\mathbf{\Sigma}_\epsilon(\mathbf{x}^m) \preceq \frac{\overline{\sigma}_0^2}{n} \mathbf{I}_m$. Therefore, $\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Phi} + \mathbf{M}^{-1} \succeq \frac{n}{\overline{\sigma}_0^2} \mathbf{\Phi}^\top \mathbf{\Phi} + \mathbf{M}^{-1} \succ 0$. This implies that

$$
0 \prec \mathbf{Q}^{1/2} \left(\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Phi} + \mathbf{M}^{-1} \right)^{-1} \mathbf{Q}^{1/2} \preceq \mathbf{Q}^{1/2} \left(\frac{n}{\overline{\sigma}_0^2} \mathbf{\Phi}^\top \mathbf{\Phi} + \mathbf{M}^{-1} \right)^{-1} \mathbf{Q}^{1/2}. \tag{22}
$$

Note that the matrices $\left(\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Phi} + \mathbf{M}^{-1} \right)^{-1}$, $\left(\frac{n}{\overline{\sigma}_0^2} \mathbf{\Phi}^\top \mathbf{\Phi} + \mathbf{M}^{-1} \right)^{-1}$, and $\mathbf{Q}$ are all symmetric and positive definite matrices. Furthermore, $\left(\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Phi} + \mathbf{M}^{-1} \right)^{-1} \mathbf{Q}$ is similar to the symmetric positive definite matrix $\mathbf{Q}^{1/2} \left(\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Phi} + \mathbf{M}^{-1} \right)^{-1} \mathbf{Q}^{1/2}$. Therefore,

$$
\lambda_{\max} \left\{ \left(\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Phi} + \mathbf{M}^{-1} \right)^{-1} \mathbf{Q} \right\} = \lambda_{\max} \left\{ \mathbf{Q}^{1/2} \left(\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \mathbf{\Phi} + \mathbf{M}^{-1} \right)^{-1} \mathbf{Q}^{1/2} \right\}, \tag{23}
$$

and similarly

$$\lambda_{\max}\left\{\left(\frac{n}{\overline{\sigma}_0^2}\mathbf{\Phi}^\top\mathbf{\Phi}+\mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right\}=\lambda_{\max}\left\{\mathbf{Q}^{1/2}\left(\frac{n}{\overline{\sigma}_0^2}\mathbf{\Phi}^\top\mathbf{\Phi}+\mathbf{M}^{-1}\right)^{-1}\mathbf{Q}^{1/2}\right\}. \tag{24}$$

(22), (23), and (24) imply that

$$\begin{aligned}
&\left|\!\left|\!\left|\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi}+\mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right|\!\right|\!\right|=\lambda_{\max}\left\{\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi}+\mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right\}\\
&\le\lambda_{\max}\left\{\left(\frac{n}{\overline{\sigma}_0^2}\mathbf{\Phi}^\top\mathbf{\Phi}+\mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right\}=\left|\!\left|\!\left|\left(\frac{n}{\overline{\sigma}_0^2}\mathbf{\Phi}^\top\mathbf{\Phi}+\mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right|\!\right|\!\right|\\
&=\frac{\overline{\sigma}_0^2}{mn}\left|\!\left|\!\left|\left\{\left(\mathbf{I}_\zeta+\frac{\overline{\sigma}_0^2}{mn}\mathbf{M}^{-1}\right)+\left(\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi}-\mathbf{I}_\zeta\right)\right\}^{-1}\mathbf{Q}\right|\!\right|\!\right|\\
&=\frac{\overline{\sigma}_0^2}{mn}\left|\!\left|\!\left|\mathbf{Q}^{-1}\left\{\mathbf{I}_\zeta+\mathbf{Q}^{-1}\left(\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi}-\mathbf{I}_\zeta\right)\mathbf{Q}^{-1}\right\}^{-1}\right|\!\right|\!\right|,\\
&\le\frac{\overline{\sigma}_0^2}{mn}|\!|\!|\mathbf{Q}^{-1}|\!|\!|\left|\!\left|\!\left|\left\{\mathbf{I}_\zeta+\mathbf{Q}^{-1}\left(\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi}-\mathbf{I}_\zeta\right)\mathbf{Q}^{-1}\right\}^{-1}\right|\!\right|\!\right|
\end{aligned} \tag{25}$$

We consider the event defined in Lemma 3 with $\delta=1/2$, i.e.

$$\mathcal{E}_3=\left\{\left|\!\left|\!\left|\mathbf{Q}^{-1}\left(\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi}-\mathbf{I}_\zeta\right)\mathbf{Q}^{-1}\right|\!\right|\!\right|\le\frac{1}{2}\right\}.$$

Then on the event $\mathcal{E}_3$, $\mathbf{I}_\zeta+\mathbf{Q}^{-1}\left(\frac{1}{m}\mathbf{\Phi}^\top\mathbf{\Phi}-\mathbf{I}_\zeta\right)\mathbf{Q}^{-1}\succeq(1-1/2)\mathbf{I}_\zeta=(1/2)\mathbf{I}_\zeta$. Moreover, $0\prec\mathbf{Q}^{-1}\prec\mathbf{I}_\zeta$. Therefore, (25) implies that

$$\left|\!\left|\!\left|\left(\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi}+\mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right|\!\right|\!\right|\le\frac{2\overline{\sigma}_0^2}{mn}|\!|\!|\mathbf{Q}^{-1}|\!|\!|\le\frac{2\overline{\sigma}_0^2}{mn}. \tag{26}$$

In (21), the term $\left\|\mathbf{Q}^{-1}\mathbf{M}^{-1}\theta_s^{\downarrow}\right\|$ can be bounded as

$$\begin{aligned}
&\left\|\mathbf{Q}^{-1}\mathbf{M}^{-1}\theta_s^{\downarrow}\right\|=\sqrt{\left(\theta_s^{\downarrow}\right)^\top\mathbf{M}^{-1}\mathbf{Q}^{-2}\mathbf{M}^{-1}\theta_s^{\downarrow}}=\sqrt{\left(\theta_s^{\downarrow}\right)^\top\left(\mathbf{M}^2+\frac{\overline{\sigma}_0^2}{mn}\mathbf{M}\right)^{-1}\theta_s^{\downarrow}}\\
&\le\sqrt{\left(\theta_s^{\downarrow}\right)^\top\left(\frac{\overline{\sigma}_0^2}{mn}\mathbf{M}\right)^{-1}\theta_s^{\downarrow}}=\sqrt{\frac{mn}{\overline{\sigma}_0^2}}\sqrt{\sum_{l=1}^{\zeta}\frac{\theta_{sl}^2}{\mu_l^2}}\le\sqrt{\frac{mn}{\overline{\sigma}_0^2}}\left\|\mathrm{f}_s\right\|_{\mathbb{H}}.
\end{aligned} \tag{27}$$

For the term $\left\|\mathbf{Q}^{-1}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\|$ in (21), we first have that

$$\begin{aligned}
&\left\|\mathbf{Q}^{-1}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\|=\left\|\left(\mathbf{M}+\frac{\overline{\sigma}_0^2}{mn}\mathbf{I}_\zeta\right)^{-1/2}\mathbf{M}^{1/2}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\|\\
&\le\left|\!\left|\!\left|\left(\mathbf{M}+\frac{\overline{\sigma}_0^2}{mn}\mathbf{I}_\zeta\right)^{-1/2}\right|\!\right|\!\right|\cdot\left\|\mathbf{M}^{1/2}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\|=\frac{1}{\sqrt{\mu_\zeta+\frac{\overline{\sigma}_0^2}{mn}}}\left\|\mathbf{M}^{1/2}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\|
\end{aligned}$$

$$
\begin{aligned}
&\le \sqrt{\frac{mn}{\overline{\sigma}_0^2}} \left\|\mathbf{M}^{1/2}\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\| = \sqrt{\frac{mn}{\overline{\sigma}_0^2}} \sqrt{\sum_{l=1}^{\zeta} \mu_l \left(\boldsymbol{\phi}_l^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right)^2} \\
&\overset{(i)}{\le} \sqrt{\frac{mn}{\overline{\sigma}_0^2}} \left\{\sum_{l=1}^{\zeta} \mu_l \left(\boldsymbol{\phi}_l^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\boldsymbol{\phi}_l\right)\left(\mathbf{v}_s^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right)\right\}^{1/2} \overset{(ii)}{\le} \frac{n}{\underline{\sigma}_0^2}\sqrt{\frac{mn}{\overline{\sigma}_0^2}} \sqrt{\sum_{l=1}^{\zeta} \mu_l \left\|\boldsymbol{\phi}_l\right\|^2 \left\|\mathbf{v}_s\right\|^2},
\end{aligned}
\tag{28}
$$

where (i) follows from the Cauchy-Schwarz inequality, and (ii) follows from Assumption A.1 that $\sigma_j^2 \ge \underline{\sigma}_0^2$ for all $j = 1, \dots, m$ and hence $\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1} \preceq \frac{n}{\underline{\sigma}_0^2}\mathbf{I}_\zeta$.

We can combine (21), (26), (27), (28), and apply the inequality $(a+b)^2 \le 2a^2 + 2b^2$ to obtain that

$$
\begin{aligned}
&\left\|\delta_s^{\downarrow}\right\|^2 \le 2\left|\!\left|\!\left|\left(\mathbf{\Phi}^\top \mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{\Phi} + \mathbf{M}^{-1}\right)^{-1}\mathbf{Q}\right|\!\right|\!\right|^2 \left(\left\|\mathbf{Q}^{-1}\mathbf{M}^{-1}\theta_s^{\downarrow}\right\|^2 + \left\|\mathbf{Q}^{-1}\mathbf{\Phi}^\top\mathbf{\Sigma}_\epsilon(\mathbf{x}^m)^{-1}\mathbf{v}_s\right\|^2\right) \\
&\le 2\left(\frac{2\overline{\sigma}_0^2}{mn}\right)^2 \left\{\frac{mn}{\overline{\sigma}_0^2}\left\|\mathrm{f}_s\right\|_{\mathbb{H}}^2 + \left(\frac{n}{\underline{\sigma}_0^2}\right)^2 \frac{mn}{\overline{\sigma}_0^2}\sum_{l=1}^{d} \mu_l \left\|\boldsymbol{\phi}_l\right\|^2 \left\|\mathbf{v}_s\right\|^2\right\} \\
&= 8\left\{\frac{\overline{\sigma}_0^2}{mn}\left\|\mathrm{f}_s\right\|_{\mathbb{H}}^2 + \frac{n\overline{\sigma}_0^2}{m\underline{\sigma}_0^4}\sum_{l=1}^{\zeta} \mu_l \left\|\boldsymbol{\phi}_l\right\|^2 \left\|\mathbf{v}_s\right\|^2\right\}.
\end{aligned}
\tag{29}
$$

Now we evaluate the expectation $\mathrm{E}_{\mathbf{X}^m}\|\delta_s^{\downarrow}\|^2$. From (29), it suffices to control $\mathrm{E}_{\mathbf{X}^m}\left(\|\boldsymbol{\phi}_l\|^2\|\mathbf{v}_s\|^2\right)$ for $l = 1, \dots, d$. By the Cauchy-Schwarz inquality,

$$
\mathrm{E}_{\mathbf{X}^m}\left(\|\boldsymbol{\phi}_l\|^2\|\mathbf{v}_s\|^2\right) \le \sqrt{\mathrm{E}_{\mathbf{X}^m}\left(\|\boldsymbol{\phi}_l\|^4\right)}\sqrt{\mathrm{E}_{\mathbf{X}^m}\left(\|\mathbf{v}_s\|^4\right)}. \tag{30}
$$

By Assumption A.3, $\mathrm{E}_{\mathbf{P}_\mathbf{X}}\left\{\phi_l^{2r_*}(\mathbf{X})\right\} \le \rho_*^{2r_*}$ for some $r_* \ge 2$. By Jensen's inequality, for all $l = 1, 2, \dots$,

$$
\mathrm{E}_{\mathbf{P}_\mathbf{X}}\left\{\phi_l^4(\mathbf{X})\right\} \le \left[\mathrm{E}_{\mathbf{P}_\mathbf{X}}\left\{\phi_l^{2r_*}(\mathbf{X})\right\}\right]^{2/r_*} \le \rho_*^{2r_*\cdot 2/r_*} = \rho_*^4.
$$

Since $\mathbf{X}_1, \dots, \mathbf{X}_m$ are i.i.d. distributed as $\mathbf{P}_\mathbf{X}$ and $\mathrm{E}_{\mathbf{P}_\mathbf{X}}\left\{\phi_l^4(\mathbf{X})\right\} \le \rho_*^4$ for all l, we have that

$$
\begin{aligned}
\mathrm{E}_{\mathbf{X}^m}\left(\|\boldsymbol{\phi}_l\|^4\right) &= \mathrm{E}_{\mathbf{X}^m}\left\{\left(\sum_{j=1}^{m}\phi_l^2(\mathbf{X}_j)\right)^2\right\} \\
&\le \mathrm{E}_{\mathbf{X}^m}\left(m\sum_{j=1}^{m}\phi_l^4(\mathbf{X}_j)\right) \le m^2\mathrm{E}_{\mathbf{X}^m}\left(\phi_l^4(\mathbf{X}_1)\right) \le m^2\rho_*^4.
\end{aligned}
\tag{31}
$$

On the other hand, by applying the Cauchy-Schwarz inequality, we have

$$
\mathrm{E}_{\mathbf{X}^m}\left(\|\mathbf{v}_s\|^4\right) = \mathrm{E}_{\mathbf{X}^m}\left\{\left(\sum_{j=1}^{m} v_{sj}^2\right)^2\right\} \le m\mathrm{E}_{\mathbf{X}^m}\left(\sum_{j=1}^{m} v_{sj}^4\right) = m^2\mathrm{E}_{\mathbf{X}^m}\left(v_{s1}^4\right)
$$

$$= m^2 \mathrm{E}_{\mathbf{X}^m}\left\{\left(\sum_{l=\zeta+1}^{\infty}\delta_{sl}\phi_l(\mathbf{X}_1)\right)^4\right\} \le m^2 \mathrm{E}_{\mathbf{X}^m}\left[\left\{\sum_{l=\zeta+1}^{\infty}\frac{\delta_{sl}^2}{\mu_l}\cdot\sum_{l=\zeta+1}^{\infty}\mu_l\phi_l^2(\mathbf{X}_1)\right\}^2\right]. \tag{32}$$

From (18), we can get an upper bound $\sum_{l=\zeta+1}^{\infty}\frac{\delta_{sl}^2}{\mu_l} \le \sum_{l=1}^{\infty}\frac{\delta_{sl}^2}{\mu_l} = \|\eta_s\|_{\mathbb{H}}^2 \le \|\mathrm{f}_s\|_{\mathbb{H}}^2$. Therefore, (32) further implies that

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}^m}\left(\|\mathbf{v}_s\|^4\right) &\le m^2\|\mathrm{f}_s\|_{\mathbb{H}}^4\cdot\mathrm{E}_{\mathbf{X}^m}\left[\left\{\sum_{l=\zeta+1}^{\infty}\mu_l\phi_l^2(\mathbf{X}_1)\right\}^2\right]\\
&= m^2\|\mathrm{f}_s\|_{\mathbb{H}}^4\cdot\mathrm{E}_{\mathbf{X}^m}\left\{\sum_{a=\zeta+1}^{\infty}\sum_{b=\zeta+1}^{\infty}\mu_a\mu_b\phi_a^2(\mathbf{X}_1)\phi_b^2(\mathbf{X}_1)\right\}\\
&\overset{(i)}{\le} m^2\|\mathrm{f}_s\|_{\mathbb{H}}^4\cdot\left\{\sum_{a=\zeta+1}^{\infty}\sum_{b=\zeta+1}^{\infty}\mu_a\mu_b\sqrt{\mathrm{E}_{\mathbf{X}^m}\phi_a^4(\mathbf{X}_1)\cdot\mathrm{E}_{\mathbf{X}^m}\phi_b^4(\mathbf{X}_1)}\right\}\\
&\overset{(ii)}{\le} m^2\rho_*^4\|\mathrm{f}_s\|_{\mathbb{H}}^4\sum_{a=\zeta+1}^{\infty}\sum_{b=\zeta+1}^{\infty}\mu_a\mu_b = m^2\rho_*^4\|\mathrm{f}_s\|_{\mathbb{H}}^4\left\{\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\right\}^2,
\end{aligned} \tag{33}$$

where (i) follows from the Cauchy-Schwarz inequality and the monotone convergence theorem, and (ii) follows from Assumption A.3.

We combine (29), (30), (31), and (33), and to obtain that

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}^m}\left(\left\|\delta_s^{\downarrow}\right\|^2 \;\middle|\; \mathcal{E}_3\right) &\le 8\left\{\frac{\overline{\sigma}_0^2}{mn}\|\mathrm{f}_s\|_{\mathbb{H}}^2 + \frac{n\overline{\sigma}_0^2}{m\underline{\sigma}_0^4}\sum_{l=1}^{\zeta}\mu_l\cdot m^2\rho_*^4\|\mathrm{f}_s\|_{\mathbb{H}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\right\}\\
&\le 8\|\mathrm{f}_s\|_{\mathbb{H}}^2\left\{\frac{\overline{\sigma}_0^2}{mn} + \frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\right\}.
\end{aligned} \tag{34}$$

We also have the coarse upper bound for $\mathrm{E}_{\mathbf{X}^m}\|\delta_s^{\downarrow}\|^2$ using (18):

$$\mathrm{E}_{\mathbf{X}^m}\left\|\delta_s^{\downarrow}\right\|^2 = \sum_{l=1}^{\zeta}\delta_{sl}^2 \le \sum_{l=1}^{\infty}\delta_{sl}^2 \le \mu_1\sum_{l=1}^{\infty}\frac{\delta_{sl}^2}{\mu_l} = \mu_1\|\eta_s\|_{\mathbb{H}}^2 \le \mu_1\|\mathrm{f}_s\|_{\mathbb{H}}^2 \le \|\mathrm{f}_s\|_{\mathbb{H}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right). \tag{35}$$

This together with the upper bound for $\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_3^c)$ in Lemma 3 (with $\delta = 1/2$) implies that

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}^m}\left\|\delta_s^{\downarrow}\right\|^2 &= \mathrm{E}_{\mathbf{X}^m}\left\{\left\|\delta_s^{\downarrow}\right\|^2\mathbb{1}(\mathcal{E}_3)\right\} + \mathrm{E}_{\mathbf{X}^m}\left\{\left\|\delta_s^{\downarrow}\right\|^2\mathbb{1}(\mathcal{E}_3^c)\right\}\\
&\le \mathrm{E}_{\mathbf{X}^m}\left(\left\|\delta_s^{\downarrow}\right\|^2 \;\middle|\; \mathcal{E}_3\right)\cdot\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_3) + \mathrm{E}_{\mathbf{X}^m}\left\|\delta_s^{\downarrow}\right\|^2\cdot\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_3^c)\\
&\le \mathrm{E}_{\mathbf{X}^m}\left(\left\|\delta_s^{\downarrow}\right\|^2 \;\middle|\; \mathcal{E}_3\right) + \mathrm{E}_{\mathbf{X}^m}\left\|\delta_s^{\downarrow}\right\|^2\cdot\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_3^c)\\
&\le 8\|\mathrm{f}_s\|_{\mathbb{H}}^2\left\{\frac{\overline{\sigma}_0^2}{mn} + \frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\mathrm{tr}\left(\mathbf{\Sigma}_M^{(\zeta)}\right)\right\}\\
&\quad + \|\mathrm{f}_s\|_{\mathbb{H}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_M\right)\left\{200\rho_*^2\frac{b(m,\zeta,r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}.
\end{aligned} \tag{36}$$

On the other hand, from (18), we have that

$$\begin{aligned}\left\|\delta_s^{\uparrow}\right\|^2 &= \sum_{l=\zeta+1}^{\infty} \delta_{sl}^2 \le \mu_{\zeta+1} \sum_{l=\zeta+1}^{\infty} \frac{\delta_{sl}^2}{\mu_l} \le \mu_{\zeta+1} \sum_{l=1}^{\infty} \frac{\delta_{sl}^2}{\mu_l} \\ &= \mu_{\zeta+1}\|\eta_s\|_{\mathbb{H}}^2 \le \mu_{\zeta+1}\|\mathrm{f}_s\|_{\mathbb{H}}^2 \le \operatorname{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right)\|\mathrm{f}_s\|_{\mathbb{H}}^2. \end{aligned} \tag{37}$$

Therefore, (36) and (37) together imply that

$$\begin{aligned}\mathrm{E}_{\mathbf{X}^m}\|\eta_s\|_2^2 = \mathrm{E}_{\mathbf{X}^m}\|\delta_s\|^2 &= \mathrm{E}_{\mathbf{X}^m}\left\|\delta_s^{\downarrow}\right\|^2 + \mathrm{E}_{\mathbf{X}^m}\left\|\delta_s^{\uparrow}\right\|^2 \\ &\le 8\|\mathrm{f}_s\|_{\mathbb{H}}^2 \frac{\overline{\sigma}_0^2}{mn} + 8\|\mathrm{f}_s\|_{\mathbb{H}}^2 \frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4 \operatorname{tr}(\boldsymbol{\Sigma}_M)\operatorname{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right) + \|\mathrm{f}_s\|_{\mathbb{H}}^2 \operatorname{tr}\left(\boldsymbol{\Sigma}_M^{(\zeta)}\right) \\ &+ \|\mathrm{f}_s\|_{\mathbb{H}}^2 \operatorname{tr}(\boldsymbol{\Sigma}_M)\left\{200\rho_*^2 \frac{b(m,\zeta,r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}. \end{aligned} \tag{38}$$

Taking the infimum with respect to ζ leads to the result. □

LEMMA 2. *Under Assumptions A.1-A.4, for any $\xi \in (0,1)$, there exists a large integer $m_0 \in \mathbb{N}$ that depends on ξ, $\boldsymbol{\Sigma}_M$, $\mathbf{f}$, n, $\overline{\sigma}_0^2$ in Assumption A.1, and ρ_* in Assumption A.3, such that for all $m > m_0$,*

$$\mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left\{\mathbf{F}^\top(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m))^{-1}\mathbf{F}\right\} \ge \frac{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}{8\operatorname{tr}(\boldsymbol{\Sigma}_M)}\right) \ge 1-\xi.$$

Proof of Lemma 2:

$$\begin{aligned}&\lambda_{\min}\left\{\mathbf{F}^\top(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m))^{-1}\mathbf{F}\right\} = \min_{\|a\|=1} a^\top\mathbf{F}^\top(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m))^{-1}\mathbf{F}a \\ &\ge \lambda_{\min}\left\{m(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m))^{-1}\right\}\cdot \min_{\|a\|=1} a^\top\left(\frac{1}{m}\mathbf{F}^\top\mathbf{F}\right)a \\ &= \lambda_{\min}\left\{m(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m))^{-1}\right\}\cdot\lambda_{\min}\left(\frac{1}{m}\mathbf{F}^\top\mathbf{F}\right).\end{aligned}$$

Therefore, for any constants $c_1, c_2 > 0$,

$$\begin{aligned}&\mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left\{\mathbf{F}^\top(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m))^{-1}\mathbf{F}\right\} < c_1c_2\right) \\ &\le \mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left(\frac{1}{m}\mathbf{F}^\top\mathbf{F}\right) < c_1\right) + \mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left\{m(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m))^{-1}\right\} < c_2\right) \\ &= \mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left(\frac{1}{m}\mathbf{F}^\top\mathbf{F}\right) < c_1\right) + \mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\max}(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)+\boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)) > m/c_2\right). \end{aligned} \tag{39}$$

We choose the values of c_1 and c_2 and bound the two terms separately. Since $\frac{1}{m}\mathbf{F}^\top\mathbf{F} = \frac{1}{m}\sum_{j=1}^m \mathbf{f}(\mathbf{X}_j)\mathbf{f}(\mathbf{X}_j)^\top$, by the strong law of large numbers, $\frac{1}{m}\mathbf{F}^\top\mathbf{F} \xrightarrow{a.s.} \mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]$ as $m\to\infty$, where $\xrightarrow{a.s.}$ means the almost sure convergence. Since $\lambda_{\min}(\cdot)$ is a continuous function, by the continuous mapping theorem, $\lambda_{\min}\left(\frac{1}{m}\mathbf{F}^\top\mathbf{F}\right) \xrightarrow{a.s.} \lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)$ as $m\to\infty$. Therefore, we can set $c_1 = \frac{1}{2}\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)$, and for any given constant $\xi\in(0,1)$, there exists a large integer $m_1 = m_1(\xi)\in\mathbb{N}$, such that for all $m\geq m_1$,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X}^m}\left(\left|\lambda_{\min}\left(\frac{1}{m}\mathbf{F}^\top\mathbf{F}\right) - \lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)\right| > \frac{1}{2}\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)\right) < \frac{\xi}{2} \\
\Longrightarrow\, &\mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left(\frac{1}{m}\mathbf{F}^\top\mathbf{F}\right) < \frac{1}{2}\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)\right) < \frac{\xi}{2}. \qquad (40)
\end{aligned}$$

On the other hand, we know that by Assumption A.1, $\mathbf{\Sigma}_\epsilon(\mathbf{X}^m) \preceq \frac{\overline{\sigma}_0^2}{n}\mathbf{I}_m$. Moreover, using the monotone convergence theorem, the expectation and the variance of $\mathrm{tr}(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m))$ can be controlled as follows:

$$\begin{aligned}
&\mathrm{E}_{\mathbf{X}^m}\,\mathrm{tr}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right) = \mathrm{E}_{\mathbf{X}^m}\left\{\sum_{j=1}^m\sum_{l=1}^\infty \mu_l\phi_l^2(\mathbf{X}_j)\right\} = \sum_{j=1}^m\sum_{l=1}^\infty \mu_l \mathrm{E}_{\mathbf{X}^m}\left\{\phi_l^2(\mathbf{X}_j)\right\} \\
&= m\sum_{l=1}^\infty \mu_l = m\,\mathrm{tr}(\mathbf{\Sigma}_M), \qquad (41) \\
&\mathrm{Var}_{\mathbf{X}^m}\left\{\mathrm{tr}\left(\mathbf{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)\right\} = \mathrm{Var}_{\mathbf{X}^m}\left\{\sum_{j=1}^m\sum_{l=1}^\infty \mu_l\phi_l^2(\mathbf{X}_j)\right\} \\
&\overset{(i)}{=} \sum_{j=1}^m \mathrm{Var}_{\mathbf{X}^m}\left\{\sum_{l=1}^\infty \mu_l\phi_l^2(\mathbf{X}_j)\right\} \overset{(ii)}{\leq} \sum_{j=1}^m \mathrm{E}_{\mathbf{X}^m}\left\{\sum_{l=1}^\infty \mu_l\phi_l^2(\mathbf{X}_j)\right\}^2 \\
&\overset{(iii)}{=} \sum_{j=1}^m\sum_{a=1}^\infty\sum_{a=1}^\infty \mu_a\mu_b\mathrm{E}_{\mathbf{X}^m}\left\{\phi_a^2(\mathbf{X}_j)\phi_b^2(\mathbf{X}_j)\right\} \\
&\overset{(iv)}{\leq} \sum_{j=1}^m\sum_{a=1}^\infty\sum_{b=1}^\infty \mu_a\mu_b\sqrt{\mathrm{E}_{\mathbf{X}^m}\phi_a^4(\mathbf{X}_j)\cdot\mathrm{E}_{\mathbf{X}^m}\phi_b^4(\mathbf{X}_j)} \\
&\overset{(v)}{\leq} \sum_{j=1}^m\sum_{a=1}^\infty\sum_{b=1}^\infty \mu_a\mu_b\rho_*^4 = m\rho_*^4\left\{\mathrm{tr}(\mathbf{\Sigma}_M)\right\}^2, \qquad (42)
\end{aligned}$$

where (i) follows from the independence between $\mathbf{X}_1,\ldots,\mathbf{X}_m$, (ii) follows from the inequality $\mathrm{Var}(Z)\leq\mathrm{E}(Z^2)$ for any random variable Z, (iii) follows from the monotone convergence theorem, (iv) follows from the Cauchy-Schwarz inequality, and (v) follows from

Assumption A.3. Now we set $c_2 = 1/\left[4\operatorname{tr}(\boldsymbol{\Sigma}_M)\right]$, and $m_2 = m_2(\xi) \equiv \max\left\{2c_2\frac{\overline{\sigma}_0^2}{n}, 2\rho_*^4/\xi\right\} = \max\left\{\frac{\overline{\sigma}_0^2}{2n\operatorname{tr}(\boldsymbol{\Sigma}_M)}, 2\rho_*^4/\xi\right\}$. Then for all $m > m_2$, we have that

$$
\begin{aligned}
&\mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\max}\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m) + \boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right) > m/c_2\right) \le \mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\max}\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right) + \frac{\overline{\sigma}_0^2}{n} > m/c_2\right)\\
&\overset{(i)}{\le} \mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\max}\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right) > \frac{m}{2c_2}\right) \le \mathbb{P}_{\mathbf{X}^m}\left(\operatorname{tr}\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right) > \frac{m}{2c_2}\right)\\
&\le \mathbb{P}_{\mathbf{X}^m}\left(\left|\operatorname{tr}\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right) - \operatorname{E}\operatorname{tr}\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)\right| > \frac{m}{2c_2} - \operatorname{E}\operatorname{tr}\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)\right)\\
&\overset{(ii)}{=} \mathbb{P}_{\mathbf{X}^m}\left(\left|\operatorname{tr}\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right) - \operatorname{E}\operatorname{tr}\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)\right| > m\operatorname{tr}\left(\boldsymbol{\Sigma}_M\right)\right)\\
&\overset{(iii)}{\le} \frac{\operatorname{Var}\left\{\operatorname{tr}\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m)\right)\right\}}{m^2\left\{\operatorname{tr}\left(\boldsymbol{\Sigma}_M\right)\right\}^2} \overset{(iv)}{\le} \frac{\rho_*^4}{m} \overset{(v)}{<} \xi,
\end{aligned}
\tag{43}
$$

where (i) follows from the choice of m_2, (ii) follows from (41) and the choice of c_2, (iii) follows from the Chebyshev's inequality, (iv) follows from (42), and (v) follows from the choice of m_2 again.

We combine (39), (40), and (43) to obtain that for any given $\xi \in (0,1)$, for all $m > m_0 = m_0(\xi) \equiv \max\{m_1(\xi), m_2(\xi)\}$,

$$
\mathbb{P}_{\mathbf{X}^m}\left(\lambda_{\min}\left\{\mathbf{F}^\top\left(\boldsymbol{\Sigma}_M(\mathbf{X}^m,\mathbf{X}^m) + \boldsymbol{\Sigma}_\epsilon(\mathbf{X}^m)\right)^{-1}\mathbf{F}\right\} < \frac{\lambda_{\min}\left(\operatorname{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}{8\operatorname{tr}(\boldsymbol{\Sigma}_M)}\right) < \frac{\xi}{2} + \frac{\xi}{2} = \xi.
$$

Taking the probability of the complement leads to the conclusion. □

LEMMA 3. *(Zhang et al. (2015) Lemma 10) Let* $\boldsymbol{\phi}_l = [\phi_l(\mathbf{X}_1), \ldots, \phi_l(\mathbf{X}_m)]^\top$*, for* $l = 1, 2, \ldots$*. For a given* $\zeta \in \mathbb{N}$*, let* $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_\zeta]$*. Let* $\mathbf{M} = diag(\mu_1, \ldots, \mu_\zeta)$ *and let* $\mathbf{Q} = \left(\mathbf{I}_\zeta + \frac{\overline{\sigma}_0^2}{mn}\mathbf{M}^{-1}\right)^{1/2}$ *be the symmetric positive definite square root of* $\mathbf{I}_\zeta + \frac{\overline{\sigma}_0^2}{mn}\mathbf{M}^{-1}$*. For any given* $\delta > 0$*, define the event*

$$
\mathcal{E} = \left\{\left|\left|\left|\mathbf{Q}^{-1}\left(\frac{1}{m}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} - \mathbf{I}_\zeta\right)\mathbf{Q}^{-1}\right|\right|\right| \le \delta\right\}.
$$

Then under Assumptions A.1-A.3,

$$
\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}^c) \le \left\{100\rho_*^2\frac{b(m,\zeta,r_*)\gamma(\frac{\overline{\sigma}_0^2}{mn})}{\delta\sqrt{m}}\right\}^{r_*},
$$

where $b(m,\zeta,r_*)$ *and* $\gamma(\cdot)$ *are defined in Theorem 1.*

Proof of Theorem 3:

We use Theorems 1 and 2 to prove the results for three different types of kernels. The results of Theorems 1 and 2 will be applied to each of k individual design first and then combined.

First note that by the Markov's inequality, Theorem 1 implies that for the ith design $(i=1,\ldots,k)$,

$$\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{2\overline{\sigma}_0^2}{mn}\gamma_i\left(\frac{\overline{\sigma}_0^2}{mn}\right) + \inf_{\zeta\in\mathbb{N}}\left[\left\{\frac{3mn}{\overline{\sigma}_0^2}\operatorname{tr}(\mathbf{\Sigma}_{M,i})+1\right\}\operatorname{tr}\left(\mathbf{\Sigma}_{M,i}^{(\zeta)}\right)+\operatorname{tr}(\mathbf{\Sigma}_{M,i})\left\{300\rho_*^2\frac{b(m,\zeta,r_*)\gamma_i(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}\right], \quad (44)$$

where $\gamma_i(a)=\sum_{l=1}^{\infty}\mu_{i,l}/(\mu_{i,l}+a)$ for any $a>0$. And similarly, Theorem 2 implies that for the ith design $(i=1,\ldots,k)$,

$$\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right]\lesssim_{\mathbb{P}_{\mathbf{X}^m}}\frac{8q\operatorname{tr}(\mathbf{\Sigma}_{M,i})}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\left\{8C_{\mathrm{f}}^2\frac{\overline{\sigma}_0^2}{mn} + \inf_{\zeta\in\mathbb{N}}\left[8C_{\mathrm{f}}^2\frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\operatorname{tr}(\mathbf{\Sigma}_{M,i})\operatorname{tr}\left(\mathbf{\Sigma}_{M,i}^{(\zeta)}\right) + C_{\mathrm{f}}^2\operatorname{tr}\left(\mathbf{\Sigma}_{M,i}^{(\zeta)}\right)+C_{\mathrm{f}}^2\operatorname{tr}(\mathbf{\Sigma}_{M,i})\left\{200\rho_*^2\frac{b(m,\zeta,r_*)\gamma_i(\frac{\overline{\sigma}_0^2}{mn})}{\sqrt{m}}\right\}^{r_*}\right]\right\}, \quad (45)$$

The subsequent proofs are based on evaluating the right-hand-sides of (44) and (45). To obtain the upper bound for the maximum IMSE over $i=1,\ldots,k$, we notice that if for every $i=1,\ldots,k$, $\mathrm{E}_{\mathbf{X}_0}[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)]\lesssim_{\mathbb{P}_{\mathbf{X}^m}}a(m,n)$ for some sequence of $a(m,n)$ that does not depend on i, then $\max_{i\in\{1,\ldots,k\}}\mathrm{E}_{\mathbf{X}_0}[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)]\lesssim_{\mathbb{P}_{\mathbf{X}^m}}a(m,n)$.

Since we only care about the asymptotic orders of $\mathrm{E}_{\mathbf{X}_0}[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)]$ in terms of m and n, in the analysis below, we will use $C_1,C_2,\ldots$ to denote the constants whose values may vary from case to case but do not depend on m and n.

(i) If the ith covariance kernel $\Sigma_{M,i}$ has finite rank l_{*i}, then for the $\inf_{\zeta\in\mathbb{N}}$ terms in both (44) and (45), we let $l_*=\max_{i\in\{1,\ldots,k\}}l_{*i}$ and choose $\zeta=l_*$, which leads to $\operatorname{tr}\left(\mathbf{\Sigma}_{M,i}^{(\zeta)}\right)=0$ for all $i=1,\ldots,k$. Furthermore, since $r_*\geq 2$ in Assumption A.3, with $\zeta=l_*$,

$$b(m,\zeta,r_*)=\max\left(\sqrt{\max(r_*,\log\zeta)},\frac{\max(r_*,\log\zeta)}{m^{1/2-1/r_*}}\right)\leq\max(r_*,\log l_*),$$

$$\gamma_i\left(\frac{\overline{\sigma}_0^2}{mn}\right)=\sum_{l=1}^{l_*}\frac{\mu_{i,l}}{\mu_{i,l}+\frac{\overline{\sigma}_0^2}{mn}}\leq l_*.$$

(44) and (45) imply that for every $i = 1, \ldots, k$,

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] &\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{2l_*\overline{\sigma}_0^2}{mn} + \mathrm{tr}(\mathbf{\Sigma}_{M,i})\left\{300\rho_*^2 \frac{l_* \max(r_*, \log l_*)}{\sqrt{m}}\right\}^{r_*} \\
&\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{C_1}{mn} + \frac{C_2}{m^{r_*/2}}, \\
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] &\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{8q\,\mathrm{tr}(\mathbf{\Sigma}_{M,i})}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\left[8C_{\mathrm{f}}^2 \frac{\overline{\sigma}_0^2}{mn}\right. \\
&\qquad \left. + C_{\mathrm{f}}^2\,\mathrm{tr}\left(\mathbf{\Sigma}_{M,i}\right)\left\{200\rho_*^2 \frac{l_* \max(r_*, \log l_*)}{\sqrt{m}}\right\}^{r_*}\right] \\
&\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{C_3}{mn} + \frac{C_4}{m^{r_*/2}},
\end{aligned}$$

where C_1, C_2, C_3, C_4 are constants (note that $\max_{i\in\{1,\ldots,k\}} \mathrm{tr}\left(\mathbf{\Sigma}_{M,i}\right)$ is also a finite constant by Assumption A.2). Therefore,

$$\begin{aligned}
\max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] &\le \max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] + \max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] \\
&\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{C_5}{mn} + \frac{C_6}{m^{r_*/2}} \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \max\left(\frac{1}{mn}, \frac{1}{m^{r_*/2}}\right).
\end{aligned}$$

(ii) If the ith covariance kernel $\Sigma_{M,i}$ satisfies $\mu_{i,l} \le c_{1i}\exp\left(-c_{2i}l^{\kappa_i/d}\right)$ for all $l \in \mathbb{N}$, then for the $\inf_{\zeta\in\mathbb{N}}$ terms in both (44) and (45), we can choose $\zeta = (mn)^2$. Let $c_{1*} = \max_{i\in\{1,\ldots,k\}} c_{1i}$, $c_{2*} = \min_{i\in\{1,\ldots,k\}} c_{2i}$, and $\kappa_* = \min_{i\in\{1,\ldots,k\}} \kappa_i$. This definition implies that for any $z \ge 1$, $c_{1i}\exp\left(-c_{2i}z^{\kappa_i/d}\right) \le c_{1*}\exp\left(-c_{2*}z^{\kappa_*/d}\right)$. Then for sufficiently large m,

$$\begin{aligned}
b(m,\zeta,r_*) &= \max\left\{\sqrt{\max(r_*, \log\zeta)}, \frac{\max(r_*, \log\zeta)}{m^{1/2-1/r_*}}\right\} \\
&= \max\left\{\sqrt{\max(r_*, 2\log(mn))}, \frac{\max(r_*, 2\log(mn))}{m^{1/2-1/r_*}}\right\} \le 2\log(mn), \\
\mathrm{tr}\left(\mathbf{\Sigma}_{M,i}^{(\zeta)}\right) &= \sum_{l=(mn)^2+1}^{\infty} \mu_{i,l} \le \sum_{l=(mn)^2+1}^{\infty} c_{1i}\exp\left(-c_{2i}l^{\kappa_i/d}\right) \\
&\le \int_{(mn)^2}^{\infty} c_{1i}\exp\left(-c_{2i}z^{\kappa_i/d}\right)dz \le \int_{(mn)^2}^{\infty} c_{1*}\exp\left(-c_{2*}z^{\kappa_*/d}\right)dz \\
&\overset{(i)}{\le} \frac{c_{1*}d}{\kappa_*}\int_{(mn)^{2\kappa_*}}^{\infty} t^{\frac{d}{\kappa_*}-1}\exp\left(-c_{2*}t\right)dt,
\end{aligned}$$

where in (i), we use the change of variable $t = z^{\kappa_*/d}$. If $\kappa_*/d \ge 1$, then since $t \ge (mn)^{2\kappa_*/d} \ge 1$, we have $t^{\frac{d}{\kappa_*}-1} \le 1$. If $0 < \kappa_*/d < 1$, then there exists a large $m_0 \in \mathbb{N}$ that depends on only

c_{2*}, κ_*, d, such that for all $m \geq m_0$ and $t \geq (mn)^{2\kappa_*/d} \geq m^{2\kappa_*/d}$, we have $t^{\frac{d}{\kappa_*}-1} \leq \exp(c_{2*}t/2)$. Therefore, in all cases,

$$\operatorname{tr}\left(\mathbf{\Sigma}_{M,i}^{(\zeta)}\right) \leq \frac{c_{1*}d}{\kappa}\int_{(mn)^{2\kappa_*/d}}^{\infty} \exp\left(-c_{2*}t/2\right)dt = \frac{2c_{1*}d}{c_{2*}\kappa_*}\exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\}. \tag{46}$$

Let $l_1 = \left\{\frac{2}{c_{2*}}\log(mn)\right\}^{d/\kappa_*}$. For sufficiently large m and every $i = 1,\ldots,k$, $\gamma_i\left(\frac{\overline{\sigma}_0^2}{mn}\right)$ can be bounded by

$$\begin{aligned}
\gamma_i\left(\frac{\overline{\sigma}_0^2}{mn}\right) &= \sum_{l=1}^{\infty}\frac{\mu_{i,l}}{\mu_{i,l}+\frac{\overline{\sigma}_0^2}{mn}} = \sum_{l=1}^{\lfloor l_1\rfloor+1}\frac{\mu_{i,l}}{\mu_{i,l}+\frac{\overline{\sigma}_0^2}{mn}} + \sum_{l=\lfloor l_1\rfloor+2}^{\infty}\frac{\mu_{i,l}}{\mu_{i,l}+\frac{\overline{\sigma}_0^2}{mn}} \\
&\leq l_1 + 1 + \frac{mn}{\overline{\sigma}_0^2}\sum_{l=\lfloor l_1\rfloor+1}^{\infty} c_{1i}\exp\left(-c_{2i}l^{\kappa_i/d}\right) \\
&\leq l_1 + 1 + \frac{mn}{\overline{\sigma}_0^2}\int_{l_1}^{\infty} c_{1*}\exp\left(-c_{2*}z^{\kappa_*/d}\right)dz \\
&= l_1 + 1 + \frac{mnc_{1*}d}{\kappa_*\overline{\sigma}_0^2}\int_{l_1^{\kappa_*/d}}^{\infty} t^{\frac{d}{\kappa_*}-1}\exp\left(-c_{2*}t\right)dt \\
&\leq l_1 + 1 + \frac{mnc_{1*}d}{\kappa_*\overline{\sigma}_0^2}\int_{l_1^{\kappa_*/d}}^{\infty}\exp\left(-c_{2*}t/2\right)dt \\
&= l_1 + 1 + \frac{mnc_{1*}d}{2c_{2*}\kappa_*\overline{\sigma}_0^2}\exp\left(-c_{2*}l_1^{\kappa_*/d}/2\right) \\
&= l_1 + 1 + \frac{c_{1*}d}{2c_{2*}\kappa\overline{\sigma}_0^2} \leq C_1\log^{\frac{d}{\kappa_*}}(mn),
\end{aligned}$$

for some constant $C_1 > 0$ that does not depend on i. Therefore, (44) and (45) imply that for every $i = 1,\ldots,k$,

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] &\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{2C_1\overline{\sigma}_0^2\log^{\frac{d}{\kappa_*}}(mn)}{mn} \\
&\quad + \left\{\frac{3mn}{\overline{\sigma}_0^2}\operatorname{tr}(\mathbf{\Sigma}_{M,i}) + 1\right\}\frac{2c_{1*}d}{c_{2*}\kappa_*}\exp\left\{-\frac{c_{2*}}{2}(mn)^{2\kappa_*/d}\right\} \\
&\quad + \operatorname{tr}(\mathbf{\Sigma}_{M,i})\left\{300\rho_*^2\frac{2\log(mn)\cdot C_1\log^{\frac{d}{\kappa_*}}(mn)}{\sqrt{m}}\right\}^{r_*} \\
&\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{C_2\log^{\frac{d}{\kappa_*}}(mn)}{mn} + C_3 mn\exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\} + C_4\frac{\log^{\frac{r_*(\kappa_*+d)}{\kappa_*}}(mn)}{m^{r_*/2}}, \\
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] &\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \frac{8q\operatorname{tr}(\mathbf{\Sigma}_{M,i})}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\left[8C_{\mathrm{f}}^2\frac{\overline{\sigma}_0^2}{mn}\right. \\
&\quad + 8C_{\mathrm{f}}^2\frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4\operatorname{tr}\left(\mathbf{\Sigma}_{M,i}\right)\frac{c_{1*}}{c_{2*}}\exp\left\{-c_{2*}(mn)^2\right\}
\end{aligned}$$

$$
\begin{aligned}
&+ C_{\mathrm{f}}^2 \frac{2c_{1*}d}{c_{2*}\kappa_*} \exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\} \\
&+ C_{\mathrm{f}}^2 \operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}\right) \left\{200\rho_*^2 \frac{2\log(mn)\cdot C_1 \log^{\frac{d}{\kappa_*}}(mn)}{\sqrt{m}}\right\}^{r_*} \Bigg] \\
\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \; & \frac{C_5}{mn} + C_6 mn \exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\} + C_7 \frac{\log^{\frac{r_*(\kappa_*+d)}{\kappa_*}}(mn)}{m^{r_*/2}},
\end{aligned}
$$

for some positive constants $C_2, C_3, C_4, C_5, C_6, C_7$. Therefore,

$$
\begin{aligned}
&\max_{i\in\{1,\dots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \\
\le & \max_{i\in\{1,\dots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] + \max_{i\in\{1,\dots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] \\
\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \; & \frac{C_2 \log^{\frac{d}{\kappa_*}}(mn)}{mn} + C_3 mn \exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\} + C_4 \frac{\log^{\frac{r_*(\kappa_*+d)}{\kappa_*}}(mn)}{m^{r_*/2}} \\
& + \frac{C_5}{mn} + C_6 mn \exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\} + C_7 \frac{\log^{\frac{r_*(\kappa_*+d)}{\kappa_*}}(mn)}{m^{r_*/2}} \\
\lesssim_{\mathbb{P}_{\mathbf{X}^m}} \; & \max\left\{\frac{\log^{\frac{d}{\kappa_*}}(mn)}{mn}, \frac{\log^{\frac{r_*(\kappa_*+d)}{\kappa_*}}(mn)}{m^{r_*/2}}\right\},
\end{aligned}
$$

where the last inequality follows because $\frac{mn\exp\left\{-c_{2*}(mn)^{2\kappa_*/d}/2\right\}}{\log^{\frac{d}{\kappa_*}}(mn)/(mn)} \to 0$ and $\frac{1/(mn)}{\log^{\frac{d}{\kappa_*}}(mn)/(mn)} \to 0$ as $mn \to \infty$.

(iii) If the ith covariance kernel $\Sigma_{M,i}$ satisfies $\mu_{i,l} \le c_i l^{-2\nu_i/d-1}$ for all $l \in \mathbb{N}$, then $\mu_{i,l} \le c_* l^{-2\nu_*/d-1}$ for all $l \in \mathbb{N}$ and all $i = 1,\dots,k$, where $c_* = \max_{i\in\{1,\dots,k\}} c_i$ and $\nu_* = \min_{i\in\{1,\dots,k\}} \nu_i$. For the $\inf_{\zeta\in\mathbb{N}}$ terms in both (44) and (45), we choose $\zeta = \lfloor (mn)^{3d/(2\nu_*)} \rfloor$. Then for sufficiently large m,

$$
\begin{aligned}
b(m,\zeta,r_*) &= \max\left\{\sqrt{\max(r_*,\log\zeta)}, \frac{\max(r_*,\log\zeta)}{m^{1/2-1/r_*}}\right\} \\
&= \max\left\{\sqrt{\max\left(r_*, \frac{3d}{2\nu_*}\log(mn)\right)}, \frac{\max\left(r_*, \frac{3d}{2\nu_*}\log(mn)\right)}{m^{1/2-1/r_*}}\right\} \le \frac{3d}{2\nu_*}\log(mn),
\end{aligned}
$$

$$
\operatorname{tr}\left(\boldsymbol{\Sigma}_{M,i}^{(\zeta)}\right) = \sum_{l=\zeta+1}^{\infty} \mu_{i,l} \le \sum_{l=\zeta+1}^{\infty} c_* l^{-2\nu_*/d-1} \le \int_{\zeta}^{\infty} c_* z^{-2\nu_*/d-1} dz = \frac{c_* d}{2\nu_*}\zeta^{-2\nu_*/d} \le \frac{c_* d}{2\nu_*}(mn)^{-3},
$$

$$
\begin{aligned}
\gamma_i\left(\frac{\overline{\sigma}_0^2}{mn}\right) &= \sum_{l=1}^{\infty} \frac{1}{1+\frac{\overline{\sigma}_0^2}{mn\mu_{i,l}}} = \sum_{l=1}^{\infty} \frac{1}{1+\frac{\overline{\sigma}_0^2 l^{2\nu_*/d+1}}{c_* mn}} \le (mn)^{d/(2\nu_*+d)} + 1 + \sum_{l=\lfloor (mn)^{d/(2\nu_*+d)}\rfloor+2} \frac{c_* mn}{\overline{\sigma}_0^2 l^{2\nu_*/d+1}} \\
&\le (mn)^{d/(2\nu_*+d)} + 1 + \frac{c_* mn}{\overline{\sigma}_0^2} \int_{(mn)^{d/(2\nu_*+d)}}^{\infty} \frac{1}{z^{2\nu_*/d+1}} dz
\end{aligned}
$$

$$= (mn)^{d/(2\nu_*+d)} + 1 + \frac{c_* dmn}{2\nu_* \overline{\sigma}_0^2}(mn)^{-\frac{2\nu_*}{2\nu_*+d}} \le C_1 (mn)^{d/(2\nu_*+d)},$$

for some large constant $C_1 > 0$ that does not depend on i. Therefore, (44) and (45) imply that for every $i = 1, \dots, k$,

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} & \frac{2C_1 \overline{\sigma}_0^2 (mn)^{d/(2\nu_*+d)}}{mn} + \left\{\frac{3mn}{\overline{\sigma}_0^2}\operatorname{tr}(\mathbf{\Sigma}_{M,i}) + 1\right\}\frac{c_* d}{2\nu_*}(mn)^{-3} \\
& + \operatorname{tr}(\mathbf{\Sigma}_{M,i})\left\{300\rho_*^{2}\frac{\frac{3d}{2\nu_*}\log(mn)\cdot C_1 (mn)^{d/(2\nu_*+d)}}{\sqrt{m}}\right\}^{r_*} \\
\lesssim_{\mathbb{P}_{\mathbf{X}^m}} & C_2 (mn)^{-\frac{2\nu_*}{2\nu_*+d}} + C_3 (mn)^{-2} + C_4 \frac{n^{\frac{dr_*}{2\nu_*+d}}\log^{r_*}(mn)}{m^{\frac{r_*(2\nu_*-d)}{2(2\nu_*+d)}}}, \\
\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} & \frac{8q\operatorname{tr}(\mathbf{\Sigma}_{M,i})}{\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right)}\left[8C_{\mathrm{f}}^2 \frac{\overline{\sigma}_0^2}{mn}\right. \\
& + 8C_{\mathrm{f}}^2 \frac{mn\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\rho_*^4 \operatorname{tr}(\mathbf{\Sigma}_{M,i})\frac{c_* d}{2\nu_*}(mn)^{-3} + C_{\mathrm{f}}^2 \frac{c_* d}{2\nu_*}(mn)^{-3} \\
& \left. + C_{\mathrm{f}}^2 \operatorname{tr}(\mathbf{\Sigma}_{M,i})\left\{200\rho_*^{2}\frac{\frac{3d}{2\nu_*}\log(mn)\cdot C_1 (mn)^{d/(2\nu_*+d)}}{\sqrt{m}}\right\}^{r_*}\right] \\
\lesssim_{\mathbb{P}_{\mathbf{X}^m}} & \frac{C_5}{mn} + C_6 (mn)^{-2} + C_7 \frac{n^{\frac{dr_*}{2\nu_*+d}}\log^{r_*}(mn)}{m^{\frac{r_*(2\nu_*-d)}{2\nu_*+d}}},
\end{aligned}$$

for some positive constants $C_2, C_3, C_4, C_5, C_6, C_7$. Therefore,

$$\begin{aligned}
\max_{i\in\{1,\dots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \le & \max_{i\in\{1,\dots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] + \max_{i\in\{1,\dots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] \\
\lesssim_{\mathbb{P}_{\mathbf{X}^m}} & C_2 (mn)^{-\frac{2\nu_*}{2\nu_*+d}} + C_3 (mn)^{-2} + C_4 \frac{n^{\frac{dr_*}{2\nu_*+d}}\log^{r_*}(mn)}{m^{\frac{r_*(2\nu_*-d)}{2\nu_*+d}}} \\
& + \frac{C_5}{mn} + C_6 (mn)^{-2} + C_7 \frac{n^{\frac{dr_*}{2\nu+d}}\log^{r_*}(mn)}{m^{\frac{r_*(2\nu_*-d)}{2\nu_*+d}}} \\
\lesssim_{\mathbb{P}_{\mathbf{X}^m}} & \max\left\{\frac{1}{(mn)^{\frac{2\nu_*}{2\nu_*+d}}}, \frac{n^{\frac{dr_*}{2\nu_*+d}}\log^{r_*}(mn)}{m^{\frac{r_*(2\nu_*-d)}{2\nu_*+d}}}\right\},
\end{aligned}$$

where the last inequality follows because $\frac{1/(mn)}{(mn)^{-\frac{2\nu_*}{2\nu_*+d}}} = (mn)^{-d/(2\nu_*+d)} \to 0$ and $\frac{1/(mn)^2}{(mn)^{-\frac{2\nu_*}{2\nu_*+d}}} = (mn)^{-d/(2\nu_*+d)-1} \to 0$ as $mn \to \infty$. □

Proof of Theorem 4:

For $i=1,\dots,k$, let $\widetilde{\mathbf{X}}_{i,0}=\left(\sqrt{b_i},\mathbf{X}_0^\top\right)^\top$ be the $\mathbb{R}^{d+1}$ random vector version of $\widetilde{\mathbf{x}}_{i,0}$ with $\mathbf{X}_0$ following the distribution $\mathbb{P}_{\mathbf{X}}$. For the covariance kernel $\mathbf{\Sigma}_{M,i}(\mathbf{x},\mathbf{x}')=a_i\left(\mathbf{x}^\top\mathbf{x}'+b_i\right)$, using the definition of $\widetilde{\mathbf{x}}_{i,0}$ and $\mathbf{Z}_i$ in Theorem 4, we have that

$$\begin{aligned}
\mathbf{\Sigma}_{M,i}(\mathbf{X}^m,\mathbf{X}_0)&=\left(\mathbf{\Sigma}_{M,i}(\mathbf{X}_1,\mathbf{X}_0),\dots,\mathbf{\Sigma}_{M,i}(\mathbf{X}_m,\mathbf{X}_0)\right)^\top\\
&=\left(a_i(\mathbf{X}_1^\top\mathbf{X}_0+b_i),\dots,a_i(\mathbf{X}_m^\top\mathbf{X}_0+b_i)\right)^\top=a_i\mathbf{Z}_i\widetilde{\mathbf{X}}_0,\\
\mathbf{\Sigma}_{M,i}(\mathbf{X}^m,\mathbf{X}^m)&=\begin{pmatrix}\mathbf{\Sigma}_{M,i}(\mathbf{X}_1,\mathbf{X}_1)&\dots&\mathbf{\Sigma}_{M,i}(\mathbf{X}_1,\mathbf{X}_m)\\ &\dots&\\ \mathbf{\Sigma}_{M,i}(\mathbf{X}_m,\mathbf{X}_1)&\dots&\mathbf{\Sigma}_{M,i}(\mathbf{X}_m,\mathbf{X}_m)\end{pmatrix}\\
&=\begin{pmatrix}a_i\left(\mathbf{X}_1^\top\mathbf{X}_1+b_i\right)&\dots&a_i\left(\mathbf{X}_1^\top\mathbf{X}_m+b_i\right)\\ &\dots&\\ a_i\left(\mathbf{X}_m^\top\mathbf{X}_1+b_i\right)&\dots&a_i\left(\mathbf{X}_m^\top\mathbf{X}_m+b_i\right)\end{pmatrix}\\
&=a_i\mathbf{Z}_i\mathbf{Z}_i^\top.
\end{aligned}$$

Therefore, we plug in $\mathbf{f}_i(\mathbf{X})\equiv 0$ to (2) of the manuscript and obtain that

$$\begin{aligned}
\widehat{y}_i(\mathbf{X}_0)&=\mathbf{\Sigma}_{M,i}(\mathbf{X}^m,\mathbf{X}_0)^\top\left[\mathbf{\Sigma}_{M,i}(\mathbf{X}^m,\mathbf{X}^m)+\frac{\sigma^2}{n}\mathbf{I}_m\right]^{-1}\overline{Y}_i\\
&=a_i\widetilde{\mathbf{X}}_{i,0}^\top\mathbf{Z}_i^\top\left(a_i\mathbf{Z}_i\mathbf{Z}_i^\top+\frac{\sigma^2}{n}\mathbf{I}_m\right)^{-1}\overline{Y}_i.
\end{aligned}$$

Similarly we obtain from (3) of the manuscript that

$$\begin{aligned}
\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)&=\mathbf{\Sigma}_{M,i}(\mathbf{X}_0,\mathbf{X}_0)-\mathbf{\Sigma}_{M,i}^\top(\mathbf{X}^m,\mathbf{X}_0)\left[\mathbf{\Sigma}_{M,i}(\mathbf{X}^m,\mathbf{X}^m)+\frac{\sigma^2}{n}\mathbf{I}_m\right]^{-1}\mathbf{\Sigma}_{M,i}(\mathbf{X}^m,\mathbf{X}_0)\\
&=a_i\widetilde{\mathbf{X}}_{i,0}^\top\widetilde{\mathbf{X}}_{i,0}-a_i\widetilde{\mathbf{X}}_{i,0}^\top\mathbf{Z}_i^\top\left(a_i\mathbf{Z}_i\mathbf{Z}_i^\top+\frac{\sigma^2}{n}\mathbf{I}_m\right)^{-1}a_i\mathbf{Z}_i\widetilde{\mathbf{X}}_{i,0}\\
&=a_i\widetilde{\mathbf{X}}_{i,0}^\top\left[\mathbf{I}_{d+1}-\mathbf{Z}_i^\top\left(\mathbf{Z}_i\mathbf{Z}_i^\top+\frac{\sigma^2}{a_in}\mathbf{I}_m\right)^{-1}\mathbf{Z}_i\right]\widetilde{\mathbf{X}}_{i,0}\\
&\overset{(i)}{=}a_i\widetilde{\mathbf{X}}_{i,0}^\top\left(\mathbf{I}_{d+1}+\frac{a_in}{\sigma^2}\mathbf{Z}_i^\top\mathbf{Z}_i\right)^{-1}\widetilde{\mathbf{X}}_{i,0},
\end{aligned}$$

where we have applied the Woodbury matrix inversion formula (Rasmussen and Williams 2006, Appendix A.3) in the step (i). This has proved (12) of the main text.

Now we turn to (13) of the manuscript. Note that

$$\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right]=\mathrm{E}_{\mathbf{X}_0}\left[a_i\widetilde{\mathbf{X}}_{i,0}^\top\left(\mathbf{I}_{d+1}+\frac{a_in}{\sigma^2}\mathbf{Z}_i^\top\mathbf{Z}_i\right)^{-1}\widetilde{\mathbf{X}}_{i,0}\right]$$

$$= \operatorname{tr}\left\{\left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2}\mathbf{Z}_i^\top \mathbf{Z}_i\right)^{-1} \cdot a_i \mathrm{E}_{\mathbf{X}_0}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right)\right\}. \tag{47}$$

According to the definition of $\mathbf{Z}_i$ and the fact that $\mathbf{X}^m, \mathbf{X}_0$ are i.i.d. draws from $\mathbb{P}_{\mathbf{X}}$, by the strong law of large numbers, as $m \to \infty$, almost surely in $\mathbb{P}_{\mathbf{X}^m}$,

$$\begin{aligned} \frac{1}{m}\mathbf{Z}_i^\top \mathbf{Z}_i &= \begin{pmatrix} b_i & \frac{\sqrt{b_i}}{m}\sum_{j=1}^m \mathbf{X}_j^\top \\ \frac{\sqrt{b_i}}{m}\sum_{j=1}^m \mathbf{X}_j & \frac{1}{m}\sum_{j=1}^m \mathbf{X}_j\mathbf{X}_j^\top \end{pmatrix} \\ &\to \begin{pmatrix} b_i & \sqrt{b_i}\mathrm{E}_{\mathbf{X}_1}(\mathbf{X}_1^\top) \\ \sqrt{b_i}\mathrm{E}_{\mathbf{X}_1}(\mathbf{X}_1) & \mathrm{E}_{\mathbf{X}_1}(\mathbf{X}_1\mathbf{X}_1^\top) \end{pmatrix} = \mathrm{E}_{\mathbf{X}_0}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right). \end{aligned} \tag{48}$$

Therefore, (47) and (48) together imply that for each $i = 1, \ldots, k$, as $m \to \infty$, almost surely in $\mathbb{P}_{\mathbf{X}^m}$,

$$\begin{aligned} mn \cdot \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] &= \operatorname{tr}\left\{\left(\frac{1}{mn}\mathbf{I}_{d+1} + \frac{a_i}{\sigma^2}\cdot\frac{1}{m}\mathbf{Z}_i^\top\mathbf{Z}_i\right)^{-1} \cdot a_i \mathrm{E}_{\mathbf{X}_0}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right)\right\} \\ &\to \operatorname{tr}\left\{\left[\frac{a_i}{\sigma^2}\cdot \mathrm{E}_{\mathbf{X}_0}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right)\right]^{-1} \cdot a_i \mathrm{E}_{\mathbf{X}_0}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right)\right\} \\ &= \operatorname{tr}\left(\sigma^2 \mathbf{I}_{d+1}\right) = (d+1)\sigma^2. \end{aligned} \tag{49}$$

Define the event $\mathcal{A}_i = \{\text{The convergence in (49) happens as } n \to \infty\}$ for $i = 1, \ldots, k$. Then the almost sure convergence in (49) implies $\mathbb{P}_{\mathbf{X}^m}(\mathcal{A}_i) = 1$ for every $i = 1, \ldots, k$. This further implies that

$$\mathbb{P}_{\mathbf{X}^m}\left(\cap_{i=1}^k \mathcal{A}_i\right) = 1 - \mathbb{P}_{\mathbf{X}^m}\left(\cup_{i=1}^k \mathcal{A}_i^c\right) \geq 1 - \sum_{i=1}^k \mathbb{P}_{\mathbf{X}^m}\left(\mathcal{A}_i^c\right) = 1 - \sum_{i=1}^k 0 = 1,$$

which implies that $\mathbb{P}_{\mathbf{X}^m}\left(\cap_{i=1}^k \mathcal{A}_i\right) = 1$, i.e. the convergence in (49) happens jointly over $i = 1, \ldots, k$ as $m \to \infty$ almost surely in $\mathbb{P}_{\mathbf{X}^m}$. Therefore, on the event $\cap_{i=1}^k \mathcal{A}_i$, (49) implies that $mn \cdot \max_{i \in \{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \to (d+1)\sigma^2$ as $m \to \infty$. This has proved (13) of the main text. □

Proof of Theorem 5:

We first derive a natural bound for $\mathrm{PFS}(\mathbf{x}_0)$. For any $\mathbf{x}_0 \in \mathcal{X}$, any $i, i' \in \{1, \ldots, k\}$, define the random variable $W_{i,i'}(\mathbf{x}_0) = [\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)] - [\widehat{y}_{i'}(\mathbf{x}_0) - y_{i'}(\mathbf{x}_0)]$ (so $W_{i,i}(\mathbf{x}_0) = 0$). According to our definition in (4) of the manuscript, $y^\circ(\mathbf{x}_0) = y_{i^\circ(\mathbf{x}_0)}(\mathbf{x}_0)$. Therefore,

$$\mathrm{PFS}(\mathbf{x}_0) = \mathbb{P}_\epsilon\left(y_{\widehat{i^\circ}(\mathbf{x}_0)}(\mathbf{x}_0) - y^\circ(\mathbf{x}_0) \geq \delta_0\right)$$

$$
\begin{aligned}
&= \mathbb{P}_\epsilon \left\{ \left[\widehat{y}_{i^\circ(\mathbf{x}_0)}(\mathbf{x}_0) - y_{i^\circ(\mathbf{x}_0)}(\mathbf{x}_0)\right] - \left[\widehat{y}^\circ(\mathbf{x}_0) - y_{\widehat{i}^\circ(\mathbf{x}_0)}(\mathbf{x}_0)\right] \geq \delta_0 + \left[\widehat{y}_{i^\circ(\mathbf{x}_0)}(\mathbf{x}_0) - \widehat{y}^\circ(\mathbf{x}_0)\right] \right\} \\
&\overset{(i)}{\leq} \mathbb{P}_\epsilon \left(W_{i^\circ(\mathbf{x}_0),\widehat{i}^\circ(\mathbf{x}_0)}(\mathbf{x}_0) \geq \delta_0 \right) \leq \mathbb{P}_\epsilon \left(\max_{1\leq i,i'\leq k} W_{i,i'}(\mathbf{x}_0) \geq \delta_0 \right) \\
&\leq \sum_{1\leq i,i'\leq k} \mathbb{P}_\epsilon \left(W_{i,i'}(\mathbf{x}_0) \geq \delta_0\right) = \sum_{1\leq i<i'\leq k} \mathbb{P}_\epsilon \left(|W_{i,i'}(\mathbf{x}_0)| \geq \delta_0\right), \qquad (50)
\end{aligned}
$$

where $\widehat{y}^\circ(\mathbf{x}_0) = \min_{i\in\{1,2,\ldots,k\}} \widehat{y}_i(\mathbf{x}_0)$. Inequality (i) holds because $\widehat{i}^\circ(\mathbf{x}_0) = \arg\min_{i\in\{1,\ldots,k\}} \widehat{y}_i(\mathbf{x}_0)$ and hence $\widehat{y}_{i^\circ(\mathbf{x}_0)}(\mathbf{x}_0) \geq \widehat{y}^\circ(\mathbf{x}_0)$. Now since $y_i(\mathbf{x}) = \mathbf{f}_i(\mathbf{x})^\top \boldsymbol{\beta}_i + M_i(\mathbf{x})$ in (1) of the manuscript includes $M_i(\mathbf{x})$, it is clear that $W_{i,i'}$ depends on $M_i(\cdot)$, $M_{i'}(\cdot)$, $\mathbf{X}^m$ and $\mathbf{x}_0$, which are all random. We first remove the randomness from $M_i(\mathbf{x})$'s ($i=1,\ldots,k$) by taking the expectation of PFS($\mathbf{x}_0$) with respect to the joint Gaussian measure $\mathbb{P}_M$ induced by the k independent Gaussian processes with mean zero and covariance function $\boldsymbol{\Sigma}_{M,i}(\cdot,\cdot)$ for $i=1,\ldots,k$. Then from (50) we can obtain that

$$
\begin{aligned}
&\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{x}_0)\right] \leq \mathrm{E}_M \Big[\sum_{1\leq i<i'\leq k} \mathbb{P}_\epsilon \left(|W_{i,i'}(\mathbf{x}_0)| \geq \delta_0\right) \Big] \\
&= \sum_{1\leq i<i'\leq k} \mathrm{E}_M \mathrm{E}_\epsilon \left[\mathbb{1}\left\{|W_{i,i'}(\mathbf{x}_0)| \geq \delta_0\right\}\right] = \sum_{1\leq i<i'\leq k} \mathbb{P}_{M,\epsilon} \left(|W_{i,i'}(\mathbf{x}_0)| \geq \delta_0\right), \qquad (51)
\end{aligned}
$$

where $\mathbb{P}_{M,\epsilon}$ denotes the joint (independent) probability measure of all $M_i(\cdot)$'s from Gaussian processes and the error terms. The inequality of (51) allows us to directly consider all randomness in $W_{i,i'}$'s given fixed $\mathbf{X}^m$ and $\mathbf{X}_0$.

Let $M_i(\mathbf{X}^m) = (M_i(\mathbf{X}_1),\ldots,M_i(\mathbf{X}_m))^\top$ and $\bar{\epsilon}(\mathbf{X}^m) = (\bar{\epsilon}_i(\mathbf{X}_1),\ldots,\bar{\epsilon}_i(\mathbf{X}_m))^\top$, for $i = 1,\ldots,k$. Under the joint measure $\mathbb{P}_{M,\epsilon}$ (with expectation $\mathrm{E}_{M,\epsilon}$), based on (2) of the manuscript, we have that for any given $\mathbf{X}^m$ and $\mathbf{x}_0 \in \mathcal{X}$,

$$
\begin{aligned}
&\mathrm{E}_{M,\epsilon}(\overline{\mathbf{Y}}_i) = \mathrm{E}_{M,\epsilon}\left[\mathbf{F}_i\boldsymbol{\beta}_i + M_i(\mathbf{X}^m) + \bar{\epsilon}(\mathbf{X}^m)\right] = \mathbf{F}_i\boldsymbol{\beta}_i, \\
&\mathrm{E}_{M,\epsilon}\left[\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)\right] \\
&= \mathrm{E}_{M,\epsilon}\left[\mathbf{f}_i(\mathbf{x}_0)^\top \widehat{\boldsymbol{\beta}}_i + \boldsymbol{\Sigma}_{M,i}(\mathbf{X}^m,\mathbf{x}_0)^\top \boldsymbol{\Sigma}_{y,i}^{-1}\left(\overline{\mathbf{Y}}_i - \mathbf{F}_i\widehat{\boldsymbol{\beta}}_i\right) - \mathbf{f}_i(\mathbf{x}_0)^\top\boldsymbol{\beta}_i - M_i(\mathbf{x}_0)\right] \\
&= \mathbf{f}_i(\mathbf{x}_0)^\top \left(\mathbf{F}_i^\top \boldsymbol{\Sigma}_{y,i}^{-1}\mathbf{F}_i\right)^{-1}\mathbf{F}_i^\top \boldsymbol{\Sigma}_{y,i}^{-1}\mathbf{F}_i\boldsymbol{\beta}_i + \boldsymbol{\Sigma}_{M,i}(\mathbf{X}^m,\mathbf{x}_0)^\top\boldsymbol{\Sigma}_{y,i}^{-1}\mathbf{F}_i\boldsymbol{\beta}_i \\
&\quad - \boldsymbol{\Sigma}_{M,i}(\mathbf{X}^m,\mathbf{x}_0)^\top\boldsymbol{\Sigma}_{y,i}^{-1}\mathbf{F}_i\left(\mathbf{F}_i^\top \boldsymbol{\Sigma}_{y,i}^{-1}\mathbf{F}_i\right)^{-1}\mathbf{F}_i^\top \boldsymbol{\Sigma}_{y,i}^{-1}\mathbf{F}_i\boldsymbol{\beta}_i - \mathbf{f}_i(\mathbf{x}_0)^\top\boldsymbol{\beta}_i \\
&= 0.
\end{aligned}
$$

Hence $\mathrm{E}_{M,\epsilon}(W_{i,i'}) = 0$ for all $1 \leq i < i' \leq k$. Furthermore, the variance of $\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)$ is $\mathrm{Var}_{M,\epsilon}[\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)] = \mathrm{E}_{M,\epsilon}[\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)]^2$, which is the MSE of $\widehat{y}_i(\mathbf{x}_0)$ and hence is equal to $\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)$ given in (3) of the manuscript. For $W_{i,i'}$ $(1 \le i < i' \le k)$, the independence between different $M_i(\cdot)$'s and errors implies that

$$\begin{aligned}\mathrm{Var}_{M,\epsilon}(W_{i,i'}) &= \mathrm{Var}_{M,\epsilon}[\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)] + \mathrm{Var}_{M,\epsilon}[\widehat{y}_{i'}(\mathbf{x}_0) - y_{i'}(\mathbf{x}_0)] \\ &= \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) + \mathrm{MSE}_{i',\mathrm{opt}}(\mathbf{x}_0).\end{aligned}$$

From (51), we apply the Markov's inequality and obtain that

$$\begin{aligned}&\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] \le \sum_{1\le i<i'\le k} \mathrm{E}_{\mathbf{X}_0}\left[\mathbb{P}_{M,\epsilon}\left(|W_{i,i'}(\mathbf{X}_0)| \ge \delta_0\right)\right] \\ &\le \sum_{1\le i<i'\le k} \mathrm{E}_{\mathbf{X}_0}\left[\frac{\mathrm{E}_{M,\epsilon}\left|W_{i,i'}(\mathbf{X}_0)\right|^2}{\delta_0^2}\right] = \sum_{1\le i<i'\le k} \mathrm{E}_{\mathbf{X}_0}\left[\frac{\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0) + \mathrm{MSE}_{i',\mathrm{opt}}(\mathbf{X}_0)}{\delta_0^2}\right] \\ &\le \frac{k(k-1)}{\delta_0^2} \max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \qquad (52)\end{aligned}$$

We now prove Part (i) of Theorem 5. Under Assumptions A.1-A.4, Part (i) of Theorem 3 says that $\max_{i\in\{1,2,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} R(m,n)$ as $m \to \infty$. This is to say that for any $\xi \in (0, 1/2)$, there exist $m_0 \ge 1$ and $c_1 > 0$ that depends on ξ, such that for all $m \ge m_0$,

$$\mathbb{P}_{\mathbf{X}^m}\left(\max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \le c_1 R(m,n)\right) \ge 1 - \xi. \qquad (53)$$

(52) and (53) together implies that

$$\begin{aligned}&\mathbb{P}_{\mathbf{X}^m}\left(\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] \le \frac{c_1 k(k-1)}{\delta_0^2} R(m,n)\right) \\ &\ge \mathbb{P}_{\mathbf{X}^m}\left(\max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \le c_1 R(m,n)\right) \ge 1 - \xi. \qquad (54)\end{aligned}$$

This is to say that for any $\xi \in (0, 1/2)$, there exist $m_0 \ge 1$ and $c_1 > 0$ that depends on ξ, such that for all $m \ge m_0$, the relation (54) holds. In other words, we have proved that $\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} R(m,n)$.

Next we prove Part (ii) of Theorem 5 with the additional Assumptions A.5 and A.6. The simulation errors $\epsilon_{il}(\mathbf{x})$'s are all normally distributed by Assumption A.5. Also $M_i(\mathbf{x})$'s are normally distributed due to the Gaussian process model. Hence we know that for given $\mathbf{X}^m$ and $\mathbf{x}_0$, $\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)$ as a linear function of $\overline{\mathbf{Y}}_i$ and $y_i(\mathbf{x}_0)$, is normally distributed as $N(0, \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0))$. The independence of $\widehat{y}_i(\mathbf{x}_0) - y_i(\mathbf{x}_0)$ and $\widehat{y}_{i'}(\mathbf{x}_0) - y_{i'}(\mathbf{x}_0)$ for $1 \le i < i' \le k$ further implies that for given $\mathbf{X}^m$ and $\mathbf{x}_0$, $\mathrm{Var}_{M,\epsilon}(W_{i,i'}) = \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) + \mathrm{MSE}_{i',\mathrm{opt}}(\mathbf{x}_0)$ and thus $W_{i,i'} \sim N(0, \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) + \mathrm{MSE}_{i',\mathrm{opt}}(\mathbf{x}_0))$. We can apply the tail probability bound of normal distributions ($\mathbb{P}(|Z| > z) \leq \exp(-z^2/2)$ if $Z \sim N(0,1)$ and $z > 0$) and obtain that

$$\mathbb{P}_{M,\epsilon}\left(|W_{i,i'}(\mathbf{x}_0)| \geq \delta_0\right) \leq \exp\left(-\frac{\delta_0^2}{2\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) + \mathrm{MSE}_{i',\mathrm{opt}}(\mathbf{x}_0)\right]}\right). \tag{55}$$

(51) and (55) together imply that

$$\begin{aligned}
\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{x}_0)\right] &\leq \sum_{1 \leq i < i' \leq k} \exp\left(-\frac{\delta_0^2}{2\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) + \mathrm{MSE}_{i',\mathrm{opt}}(\mathbf{x}_0)\right]}\right) \\
&\leq \frac{k(k-1)}{2} \exp\left(-\frac{\delta_0^2}{4 \max_{i \in \{1,\ldots,k\}} \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)}\right).
\end{aligned} \tag{56}$$

For abbreviation, we let $V = \max_{i \in \{1,\ldots,k\}} \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)$. Assumption A.6 says that for any given $\xi \in (0, 1/2)$, there exist constants $w_1 > 0, w_2 > 0, m_0 \geq 1$ that depend on ξ, such that for $m \geq m_0$, for any $t > 0$, we have $\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_4) \geq 1 - \xi$, where $\mathcal{E}_4$ is defined as

$$\mathcal{E}_4 = \left\{\mathbb{P}_{\mathbf{X}_0}\left(V \geq tR(m,n)\right) \leq w_1 \exp\left(-w_2 t\right)\right\}.$$

Conditional on the event $\mathcal{E}_4$, from (56), we can derive that

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] &\leq \mathrm{E}_{\mathbf{X}_0}\left[\frac{k(k-1)}{2}\exp\left(-\frac{\delta_0^2}{4V}\right)\right] \\
&\overset{(i)}{=} \frac{k(k-1)}{2}\int_0^{+\infty} \mathbb{P}_{\mathbf{X}_0}\left\{\exp\left(-\frac{\delta_0^2}{4V}\right) > u\right\} \mathrm{d}u \\
&= \frac{k(k-1)}{2}\int_0^{+\infty} \mathbb{P}_{\mathbf{X}_0}\left\{V > \frac{\delta_0^2}{-4\log u}\right\} \mathrm{d}u \\
&\overset{(ii)}{\leq} \frac{k(k-1)}{2}\int_0^{+\infty} w_1 \exp\left\{-w_2\left(\frac{\delta_0^2}{-4R(m,n)\log u}\right)\right\} \mathrm{d}u \\
&\overset{(iii)}{\leq} \frac{w_1 k(k-1)}{2}\int_0^{+\infty} \exp\left\{-v - \left(\frac{w_2\delta_0^2}{4R(m,n)}\right)\frac{1}{v}\right\} \mathrm{d}v, \\
&\overset{(iv)}{=} \frac{w_1 k(k-1)}{2} \cdot \sqrt{\frac{w_2\delta_0^2}{R(m,n)}} \cdot K_1\left(\sqrt{\frac{w_2\delta_0^2}{R(m,n)}}\right),
\end{aligned} \tag{57}$$

where (i) uses the relation $\mathrm{E}(Z) = \int_0^\infty P(Z > t)\mathrm{d}t$ for any nonnegative random variable Z, (ii) follows from Assumption A.6 and the relation on the event $\mathcal{E}_4$, and (iii) uses a change of variable $v = -\log u$ in the integral. (iv) follows because the integral in (57) can be recognized as the density of a generalized inverse Gaussian distribution without normalizing constant, and here $K_1(\cdot)$ is the modified Bessel function of the second kind with parameter 1.

Theorem 2.13 of Kreh (2012) has shown that

$$\lim_{x\to+\infty}\frac{K_1(x)}{\sqrt{\frac{\pi}{2x}}e^{-x}}=1,$$

which implies that there exists a constant $x_0>0$, such that for all $x>x_0$, $K_1(x)<2\sqrt{\frac{\pi}{2x}}e^{-x}=\sqrt{\frac{2\pi}{x}}e^{-x}$. Since $R(m,n)\to 0$ for fixed n as $m\to\infty$, we can take $m\geq m_1$ for some large integer $m_1\geq m_0$ such that $\sqrt{\frac{w_2\delta_0^2}{R(m,n)}}>x_0$ and meanwhile

$$[R(m,n)]^{-1/4}\leq \exp\left\{\frac{1}{2}w_2^{1/2}\delta_0\,[R(m,n)]^{-1/2}\right\}.$$

As a result, we can derive from (57) that on the event $\mathcal{E}_4$, for all $m>m_1$,

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] &\leq \frac{w_1k(k-1)}{2}\cdot\sqrt{\frac{w_2\delta_0^2}{R(m,n)}}\cdot\sqrt{\frac{2\pi}{\sqrt{\frac{w_2\delta_0^2}{R(m,n)}}}}\exp\left\{-\sqrt{\frac{w_2\delta_0^2}{R(m,n)}}\right\}\\
&\leq \sqrt{\frac{\pi}{2}}w_1w_2^{1/4}k(k-1)\delta_0^{1/2}\,[R(m,n)]^{-1/4}\exp\left\{-w_2^{1/2}\delta_0\,[R(m,n)]^{-1/2}\right\}\\
&\leq \sqrt{\frac{\pi}{2}}w_1w_2^{1/4}k(k-1)\delta_0^{1/2}\exp\left\{-\frac{1}{2}w_2^{1/2}\delta_0\,[R(m,n)]^{-1/2}\right\}. \qquad (58)
\end{aligned}$$

Thus, $\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right]\lesssim_{\mathbb{P}_{\mathbf{X}^m}}\exp\left\{-\frac{1}{2}w_2^{1/2}\delta_0\,[R(m,n)]^{-1/2}\right\}$ with probability at least $1-\xi$ for all $m\geq m_1$, which has proved Part (ii) of Theorem 5.

Finally, we prove Part (iii) of Theorem 5 with the additional Assumptions A.5 and A.7. Similar to the derivation of Part (ii), we define the quantity $\widetilde{V}=\max_{i\in\{1,\dots,k\}}\sup_{\mathbf{x}_0\in\mathcal{X}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)$ for abbreviation. Then Assumption A.7 says that for any given $\xi\in(0,1/2)$, there exist constants $w_3>0, m_0\geq 1$ that depend on ξ, such that for $m\geq m_0$, for any $t>0$, we have $\mathbb{P}_{\mathbf{X}^m}(\mathcal{E}_5)\geq 1-\xi$, where $\mathcal{E}_5$ is defined as $\mathcal{E}_5=\left\{\widetilde{V}\leq w_3R(m,n)\right\}$. Therefore, from (56), we can derive that on the event $\mathcal{E}_5$, for all $m\geq m_0$,

$$\begin{aligned}
\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right] &\leq \frac{k(k-1)}{2}\mathrm{E}_{\mathbf{X}_0}\exp\left(-\frac{\delta_0^2}{4\max_{i\in\{1,\dots,k\}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)}\right)\\
&\leq \frac{k(k-1)}{2}\sup_{\mathbf{x}_0\in\mathcal{X}}\exp\left(-\frac{\delta_0^2}{4\max_{i\in\{1,\dots,k\}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)}\right)\\
&= \frac{k(k-1)}{2}\exp\left(-\frac{\delta_0^2}{4\max_{i\in\{1,\dots,k\}}\sup_{\mathbf{x}_0\in\mathcal{X}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)}\right)\\
&= \frac{k(k-1)}{2}\exp\left(-\frac{\delta_0^2}{4\widetilde{V}}\right)\leq\frac{k(k-1)}{2}\exp\left(-\frac{\delta_0^2}{4w_3R(m,n)}\right).
\end{aligned}$$

Thus, $\mathrm{E}_{\mathbf{X}_0}\mathrm{E}_M\left[\mathrm{PFS}(\mathbf{X}_0)\right]\lesssim_{\mathbb{P}_{\mathbf{X}^m}}\exp\left\{-\frac{\delta_0^2}{4w_3}\left[R(m,n)\right]^{-1}\right\}$ with probability at least $1-\xi$ for all $m\ge m_0$, which has proved Part (iii) of Theorem 5. □

Now we discuss the restrictiveness of Assumptions A.6 and A.7 in the main text. We present Theorem 6 below to illustrate that A.6 and A.7 can hold, by using the finite-rank kernel example as described in Remark 2 and Theorem 4 of the main text.

Theorem 6. *(Exponentially decaying IPFS for finite-rank kernels) For a fixed positive integer k, consider the same model setup in Remark 2 of the main text with k finite-rank kernels $\mathbf{\Sigma}_{M,i}=a_i\left(\mathbf{x}^\top\mathbf{x}'+b_i\right)$ for any $\mathbf{x},\mathbf{x}'\in\mathcal{X}\subseteq\mathbb{R}^d$, where $a_i>0$ and $b_i>0$ are known constants for $i=1,\dots,k$. Let $\mathbb{P}_{\mathbf{X}}$ be any non-degenerate sampling distribution on $\mathcal{X}$ for $\mathbf{X}^m$ and $\mathbf{X}_0$.*

(i) Suppose that there exist constants $c_1>0,c_2>0,t_0>0$, such that $\mathbb{P}_{\mathbf{X}}$ has the tail bound $\mathbb{P}_{\mathbf{X}}(\|\mathbf{X}\|>t)\le c_1\exp(-c_2t^2)$ for all $t>t_0$. Then for the optimal MSE given in (12) of the main text, for any given $\xi\in(0,1/2)$, there exist constants $w_1>0,w_2>0,m_0\ge1$ that depend on ξ, such that for all $m\ge m_0$, for any $t>0$,

$$\mathbb{P}_{\mathbf{X}^m}\left\{\mathbb{P}_{\mathbf{X}_0}\left(mn\cdot\max_{i\in\{1,\dots,k\}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\ge t\right)\le w_1\exp\left(-w_2t\right)\right\}\ge1-\xi. \tag{59}$$

(ii) Suppose that $\mathcal{X}$ is a compact set in $\mathbb{R}^d$. Then for the optimal MSE given in (12) of the main text, for any given $\xi\in(0,1/2)$, there exist constants $w_3>0,m_0\ge1$ that depend on ξ, such that for all $m\ge m_0$,

$$\mathbb{P}_{\mathbf{X}^m}\left\{mn\cdot\max_{i\in\{1,\dots,k\}}\sup_{\mathbf{x}_0\in\mathcal{X}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)\le w_3\right\}\ge1-\xi. \tag{60}$$

Note that the rate $1/(mn)$ here is a tight convergence rate given Theorem 4, in the sense that it cannot be improved to any faster rate. The tail condition in Part (i) of Theorem 6 is satisfied by any d-dimensional multivariate normal distribution by the Hanson-Wright inequality (Hsu et al. 2012). Theorem 6 shows that Assumption A.6 holds for the finite-rank kernel if the sampling distribution of $\mathbf{X}^m$ and $\mathbf{X}_0$ has tail decaying like the Gaussian distribution. Similarly, Assumption A.7 holds when the covariance kernel and the f-functions are continuous with a compact domain.

Proof of Theorem 6:

First we show Part (i). We note that the tail condition $\mathbb{P}_{\mathbf{X}}(\|\mathbf{X}\|>t)\le c_1\exp(-c_2t^2)$ implies the finite second moment for $\mathbb{P}_{\mathbf{X}}$, because,

$$\mathrm{E}_{\mathbf{X}}\left[\|\mathbf{X}\|^2\right]=\int_0^{+\infty}\mathbb{P}_{\mathbf{X}}\left(\|\mathbf{X}\|^2>u\right)\mathrm{d}u\le\int_0^{+\infty}c_1\exp(-c_2u)\mathrm{d}u=\frac{c_1}{c_2}<+\infty.$$

Furthermore, since $\mathbb{P}_{\mathbf{X}}$ is a non-degenerate sampling distribution on $\mathbb{R}$, the covariance matrix $\mathbf{V}_{\mathbf{X}}\equiv\mathrm{E}_{\mathbf{X}_0}\left\{[\mathbf{X}_0-\mathrm{E}_{\mathbf{X}_0}(\mathbf{X}_0)][\mathbf{X}_0-\mathrm{E}_{\mathbf{X}_0}(\mathbf{X}_0)]^\top\right\}$ must be positive definite. This is because otherwise, there exists a vector $\mathbf{a}\in\mathbb{R}^d$, such that

$$0=\mathbf{a}^\top\mathrm{E}_{\mathbf{X}_0}\left\{[\mathbf{X}_0-\mathrm{E}_{\mathbf{X}_0}(\mathbf{X}_0)][\mathbf{X}_0-\mathrm{E}_{\mathbf{X}_0}(\mathbf{X}_0)]^\top\right\}\mathbf{a}=\mathrm{E}_{\mathbf{X}_0}\left\{\mathbf{a}^\top[\mathbf{X}_0-\mathrm{E}_{\mathbf{X}_0}(\mathbf{X}_0)]\right\}^2,$$

which implies that $\mathbf{a}^\top\mathbf{X}_0$ is almost surely a constant, contradicting the assumption that $\mathbb{P}_{\mathbf{X}}$ is not degenerate.

For every $i=1,\ldots,k$, we define

$$\widetilde{\mathbf{x}}_{i,0}=\left(\sqrt{b_i},\ \mathbf{x}_0^\top\right)^\top\in\mathbb{R}^{d+1},\quad \mathbf{Z}_i=\begin{pmatrix}\sqrt{b_i}&\ldots&\sqrt{b_i}\\ \mathbf{x}_1&\ldots&\mathbf{x}_m\end{pmatrix}^\top\in\mathbb{R}^{(d+1)\times m},$$

and $\widetilde{\mathbf{X}}_{i,0}$ is the $\mathbb{R}^{d+1}$ random vector version of $\widetilde{\mathbf{x}}_{i,0}$ with $\mathbf{X}_0$ following the distribution $\mathbb{P}_{\mathbf{X}}$. Define $\widetilde{\mathbf{V}}_i=\mathrm{E}_{\mathbf{X}_0}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right)$ for $i=1,\ldots,k$. Then we can write that

$$\begin{aligned}\widetilde{\mathbf{V}}_i=\mathrm{E}_{\mathbf{X}_{i,0}}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right)&=\begin{pmatrix}b_i&\sqrt{b_i}\mathrm{E}_{\mathbf{X}_{i,0}}(\mathbf{X}_{i,0}^\top)\\ \sqrt{b_i}\mathrm{E}_{\mathbf{X}_{i,0}}(\mathbf{X}_{i,0})&\mathrm{E}_{\mathbf{X}_{i,0}}(\mathbf{X}_{i,0}\mathbf{X}_{i,0}^\top)\end{pmatrix}\\ &=\begin{pmatrix}\sqrt{b_i}&0\\ \mathrm{E}_{\mathbf{X}_{i,0}}(\mathbf{X}_{i,0})&\mathbf{I}_d\end{pmatrix}\begin{pmatrix}1&0\\ 0&\mathbf{V}_{\mathbf{X}}\end{pmatrix}\begin{pmatrix}\sqrt{b_i}&\mathrm{E}_{\mathbf{X}_{i,0}}(\mathbf{X}_{i,0}^\top)\\ 0&\mathbf{I}_d\end{pmatrix}.\end{aligned}$$

From the last expression, we can see that the matrix $\widetilde{\mathbf{V}}_i=\mathrm{E}_{\mathbf{X}_{i,0}}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right)$ must be positive definite since it is congruent to a block diagonal matrix which is positive definite. Let $\underline{\lambda}_{\min}\equiv\min_{i\in\{1,\ldots,k\}}\lambda_{\min}\big(\widetilde{\mathbf{V}}_i\big)$ which is strictly positive.

Similar to the convergence in (48) in the proof of Theorem 4, by the strong law of large numbers, $\frac{1}{m}\mathbf{Z}_i^\top\mathbf{Z}_i$ converges to $\widetilde{\mathbf{V}}_i=\mathrm{E}_{\mathbf{X}_{i,0}}\left(\widetilde{\mathbf{X}}_{i,0}\widetilde{\mathbf{X}}_{i,0}^\top\right)$ entry-wise as $m\to\infty$ for each $i=1,\ldots,k$. Furthermore, for a fixed k, we have that for each $i=1,\ldots,k$, for any given $\xi\in(0,1/2)$, there exists a large integer $m_{i,0}>0$ that depends on ξ, such that for all $m\ge m_{i,0}$,

$$\mathbb{P}_{\mathbf{X}^m}\left(\left|\!\left|\!\left|\frac{1}{m}\mathbf{Z}_i^\top\mathbf{Z}_i-\widetilde{\mathbf{V}}_i\right|\!\right|\!\right|>\frac{1}{2}\lambda_{\min}\left(\widetilde{\mathbf{V}}_i\right)\right)<\frac{\xi}{k}.$$

Taking a union bound over all k designs implies that for all $m \geq m_0 \equiv \max_{i\in\{1,\dots,k\}} m_{i,0}$ implies that

$$\mathbb{P}_{\mathbf{X}^m}\left(\left\|\left|\frac{1}{m}\mathbf{Z}_i^\top\mathbf{Z}_i - \widetilde{\mathbf{V}}_i\right\|\right| > \frac{1}{2}\lambda_{\min}\left(\widetilde{\mathbf{V}}_i\right),\ \text{for all } i = 1,\dots,k\right) < \sum_{i=1}^{k}\frac{\xi}{k} = \xi.$$

This further implies that with $\mathbb{P}_{\mathbf{X}^m}$-probability at least $1-\xi$, for all $m \geq m_0$,

$$\begin{aligned}
\lambda_{\min}\left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2}\mathbf{Z}_i^\top\mathbf{Z}_i\right) &= \lambda_{\min}\left[\mathbf{I}_{d+1} + \frac{a_i mn}{\sigma^2}\left\{\frac{1}{m}\mathbf{Z}_i^\top\mathbf{Z}_i - \widetilde{\mathbf{V}}_i\right\} + \frac{a_i mn}{\sigma^2}\widetilde{\mathbf{V}}_i\right] \\
&\geq \lambda_{\min}\left(\mathbf{I}_{d+1} + \frac{a_i mn}{\sigma^2}\left[\widetilde{\mathbf{V}}_i - \frac{1}{2}\lambda_{\min}\left(\widetilde{\mathbf{V}}_i\right)\mathbf{I}_{d+1}\right]\right) \\
&\geq \lambda_{\min}\left[\mathbf{I}_{d+1} + \frac{a_i mn}{2\sigma^2}\lambda_{\min}\left(\widetilde{\mathbf{V}}_i\right)\mathbf{I}_{d+1}\right] \\
&> \frac{a_i mn}{2\sigma^2}\underline{\lambda}_{\min}.
\end{aligned}$$

Therefore, using the expression of $\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)$ derived in Theorem 4, we have that with $\mathbb{P}_{\mathbf{X}^m}$-probability at least $1-\xi$, for any $t > 0$, for all $m \geq m_0$,

$$\begin{aligned}
&\mathbb{P}_{\mathbf{X}_0}\left(mn\cdot\max_{i\in\{1,\dots,k\}}\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0) \geq t\right) \\
&= \mathbb{P}_{\mathbf{X}_0}\left(mn\cdot a_i\widetilde{\mathbf{X}}_{i,0}^\top\left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2}\mathbf{Z}_i^\top\mathbf{Z}_i\right)^{-1}\widetilde{\mathbf{X}}_{i,0} \geq t\right) \\
&\leq \mathbb{P}_{\mathbf{X}_0}\left(mna_i\cdot\left(\frac{a_i mn}{2\sigma^2}\underline{\lambda}_{\min}\right)^{-1}\widetilde{\mathbf{X}}_{i,0}^\top\widetilde{\mathbf{X}}_{i,0} \geq t\right) \\
&= \mathbb{P}_{\mathbf{X}_0}\left(b_i + \|\mathbf{X}_0\|^2 \geq \frac{\underline{\lambda}_{\min}}{2\sigma^2}t\right)
\end{aligned} \tag{61}$$

Let $\bar{b} = \max_{i\in\{1,\dots,k\}} b_i$. If $t > \max\left(4\sigma^2\underline{\lambda}_{\min}^{-1}\bar{b}, t_0\right)$, then for all $i = 1,\dots,k$,

$$\frac{\underline{\lambda}_{\min}}{2\sigma^2}t - b_i > \frac{\underline{\lambda}_{\min}}{4\sigma^2}t,$$

and from the tail assumption $\mathbb{P}_{\mathbf{X}}(\|\mathbf{X}\| > t) \leq c_1\exp(-c_2 t^2)$ in Theorem 6, we have that

$$\begin{aligned}
\mathbb{P}_{\mathbf{X}_0}\left(b_i + \|\mathbf{X}_0\|^2 \geq \frac{\underline{\lambda}_{\min}}{2\sigma^2}t\right) &\leq \mathbb{P}_{\mathbf{X}_0}\left(\|\mathbf{X}_0\|^2 \geq \frac{\underline{\lambda}_{\min}}{4\sigma^2}t\right) \\
&= \mathbb{P}_{\mathbf{X}_0}\left(\|\mathbf{X}_0\| \geq \frac{\sqrt{\underline{\lambda}_{\min}}}{2\sigma}\sqrt{t}\right) \leq c_1\exp\left(-\frac{c_2\underline{\lambda}_{\min}}{4\sigma^2}t\right).
\end{aligned} \tag{62}$$

If $0 < t \leq \max\left(4\sigma^2\underline{\lambda}_{\min}^{-1}\bar{b}, t_0\right)$, then we use the simple bound

$$\mathbb{P}_{\mathbf{X}_0}\left(b_i + \|\mathbf{X}_0\|^2 \geq \frac{\underline{\lambda}_{\min}}{2\sigma^2}t\right) \leq 1 \leq e^{c_2+1}\cdot\exp\left\{-t/\max\left(4\sigma^2\underline{\lambda}_{\min}^{-1}\bar{b}, t_0\right)\right\}. \tag{63}$$

Now let $w_1 = \max(e^{c_2+1}, c_1)$, $w_2 = \min\left\{c_2\underline{\lambda}_{\min}/(4\sigma^2), \underline{\lambda}_{\min}/(4\sigma^2\overline{b}), 1/t_0\right\}$, then (61), (62), and (63) together imply that with $\mathbb{P}_{\mathbf{X}^m}$-probability at least $1-\xi$, for any $t>0$, for all $m \geq m_0$,

$$
\begin{aligned}
&\mathbb{P}_{\mathbf{X}_0}\left(mn \cdot \max_{i\in\{1,\ldots,k\}} \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0) \geq t\right) \\
&\leq \mathbb{1}\left(0 < t \leq 4\sigma^2\underline{\lambda}_{\min}^{-1}\overline{b}\right) \cdot e^{c_2+1} \cdot \exp\left\{-t/\max\left(4\sigma^2\underline{\lambda}_{\min}^{-1}\overline{b}, t_0\right)\right\} \\
&\quad + \mathbb{1}\left(t > 4\sigma^2\underline{\lambda}_{\min}^{-1}\overline{b}\right) \cdot c_1 \exp\left(-\frac{c_2\underline{\lambda}_{\min}}{4\sigma^2} t\right) \\
&\leq w_1 \exp(-w_2 t).
\end{aligned}
$$

This has proved Part (i) of Theorem 6.

Next we show Part (ii). Let $\underline{a} = \min_{i\in\{1,\ldots,k\}} a_i$ which is strictly positive given a fixed k. From the proof above, with $\mathbb{P}_{\mathbf{X}^m}$-probability at least $1-\xi$, there exists a large integer m_0 such that uniformly for all $i = 1, \ldots, k$ and all $m \geq m_0$,

$$
\lambda_{\min}\left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2}\mathbf{Z}_i^\top \mathbf{Z}_i\right) > \frac{a_i m n}{2\sigma^2}\underline{\lambda}_{\min}.
$$

Since $\mathcal{X}$ is a compact set, there exists a constant $c_3 > 0$ such that $|\mathbf{x}| \leq c_3$ for all $\mathbf{x} \in \mathcal{X}$. Recall that $\widetilde{\mathbf{x}}_{i,0} = (\sqrt{b_i}, \mathbf{x}_0^\top)^\top$ for any $\mathbf{x}_0 \in \mathcal{X}$. Therefore, with $\mathbb{P}_{\mathbf{X}^m}$-probability at least $1-\xi$, for all $m \geq m_0$,

$$
\begin{aligned}
&mn \cdot \max_{i\in\{1,\ldots,k\}} \sup_{\mathbf{x}_0\in\mathcal{X}} \mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) \\
&= mn \cdot \max_{i\in\{1,\ldots,k\}} \sup_{\mathbf{x}_0\in\mathcal{X}} \widetilde{\mathbf{x}}_{i,0}^\top \left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2}\mathbf{Z}_i^\top \mathbf{Z}_i\right)^{-1} \widetilde{\mathbf{x}}_{i,0} \\
&\leq mn \cdot \max_{i\in\{1,\ldots,k\}} \left\{\lambda_{\min}^{-1}\left(\mathbf{I}_{d+1} + \frac{a_i n}{\sigma^2}\mathbf{Z}_i^\top \mathbf{Z}_i\right) \sup_{\mathbf{x}_0\in\mathcal{X}} \widetilde{\mathbf{x}}_{i,0}^\top \widetilde{\mathbf{x}}_{i,0}\right\} \\
&\leq mn \cdot \max_{i\in\{1,\ldots,k\}} \left\{\frac{2\sigma^2}{a_i m n}\underline{\lambda}_{\min}^{-1} \cdot \sup_{\mathbf{x}_0\in\mathcal{X}} \left(b_i + \|\mathbf{x}_0\|^2\right)\right\} \\
&\leq \frac{2\sigma^2(\overline{b} + c_3^2)\underline{\lambda}_{\min}^{-1}}{\underline{a}}.
\end{aligned}
$$

Set $w_3 = 2\sigma^2(\overline{b} + c_3^2)\underline{\lambda}_{\min}^{-1}/\underline{a}$ and then Part (ii) of Theorem 6 is proved. □

3. Estimators of IMSE and IPFS

In this section, we propose simple estimators of IMSE and IPFS based on Monte Carlo draws from the sampling distribution $\mathbb{P}_{\mathbf{X}}$. Suppose that we already have the covariate sample $\mathbf{X}^m = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$. To estimate MSE, we draw another random sample $\widetilde{\mathbf{X}}^{m'} = \{\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_{m'}\}$ from the distribution $\mathbb{P}_{\mathbf{X}}$. The two samples $\mathbf{X}^m$ and $\tilde{\mathbf{X}}^m$ are independent. The sample size m' can be different from m. Then, according the definition of MSE in Equation (3) of the main text, we estimate the IMSE under the ith design ($i = 1, \ldots, k$) as

$$\begin{aligned}
&\widehat{\mathrm{IMSE}}_i = \frac{1}{m'}\sum_{j=1}^{m'} \mathrm{MSE}_{i,\mathrm{opt}}(\widetilde{\mathbf{X}}_j), \qquad \text{where for } \ j = 1, \ldots, m', \\
&\mathrm{MSE}_{i,\mathrm{opt}}(\widetilde{\mathbf{X}}_j) = \mathbf{\Sigma}_{M,i}(\widetilde{\mathbf{X}}_j, \widetilde{\mathbf{X}}_j) - \mathbf{\Sigma}_{M,i}^\top(\mathbf{X}^m, \widetilde{\mathbf{X}}_j)\left[\mathbf{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{X}^m) + \mathbf{\Sigma}_{\epsilon,i}(\mathbf{X}^m)\right]^{-1} \mathbf{\Sigma}_{M,i}(\mathbf{X}^m, \widetilde{\mathbf{X}}_j) \\
&\quad + \eta_i(\widetilde{\mathbf{X}}_j)^\top \left[\mathbf{F}_i^\top \left(\mathbf{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{X}^m) + \mathbf{\Sigma}_{\epsilon,i}(\mathbf{X}^m)\right)^{-1} \mathbf{F}_i\right]^{-1} \eta_i(\widetilde{\mathbf{X}}_j), \\
&\text{and } \ \eta_i(\tilde{\mathbf{X}}_j) = \mathbf{f}_i(\widetilde{\mathbf{X}}_j) - \mathbf{F}_i^\top \left[\mathbf{\Sigma}_{M,i}(\mathbf{X}^m, \mathbf{X}^m) + \mathbf{\Sigma}_{\epsilon,i}(\mathbf{X}^m)\right]^{-1} \mathbf{\Sigma}_{M,i}(\mathbf{X}^m, \widetilde{\mathbf{X}}_j).
\end{aligned} \tag{64}$$

It is straightforward to see that since $\widetilde{\mathbf{X}}^{m'}$ is an i.i.d. sample from $\mathbb{P}_{\mathbf{X}}$ and is independent of the sample $\mathbf{X}^m$, the proposed estimator $\widehat{\mathrm{IMSE}}_i$ in (64) is unbiased for the IMSE defined as $\mathrm{E}_{\mathbf{X}^m}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right]$. The maximal IMSE among the k designs can be then estimated by $\max_{i\in\{1,\ldots,k\}} \widehat{\mathrm{IMSE}}_i$.

For IPFS with an IZ parameter $\delta_0 > 0$, we first need to estimate the PFS at a given covariate point $\mathbf{x}_0$, which can be approximated by the following quantity:

$$\mathrm{APFS}(\mathbf{x}_0) = \sum_{i \neq \widehat{i^\circ}(\mathbf{x}_0)} \mathbb{P}\left(N(0,1) < -\frac{\widehat{y}_i(\mathbf{x}_0) - \widehat{y}_{\widehat{i^\circ}(\mathbf{x}_0)}(\mathbf{x}_0) + \delta_0}{\sqrt{\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0) + \mathrm{MSE}_{\widehat{i^\circ}(\mathbf{x}_0),\mathrm{opt}}(\mathbf{x}_0)}}\right), \tag{65}$$

where $\widehat{i^\circ}(\mathbf{x}_0)$ and $\widehat{y}_i(\mathbf{x}_0)$ are defined in Equation (4) of the main text, $\widehat{y}_i(\mathbf{x}_0)$ is defined in Equation (2) of the main text, and $\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{x}_0)$ is defined in Equation (3) of the main text. Then, based on the random sample $\widetilde{\mathbf{X}}^{m'} = \{\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_{m'}\}$ from the distribution $\mathbb{P}_{\mathbf{X}}$ independent of $\mathbf{X}^m$, we can estimate the IPFS as

$$\widehat{\mathrm{IPFS}} = \frac{1}{m'}\sum_{j=1}^{m'} \mathrm{APFS}(\widetilde{\mathbf{X}}_j), \tag{66}$$

where $\mathrm{APFS}(\cdot)$ is defined in (65). The $\widehat{\mathrm{IPFS}}$ in (66) is a consistent estimator of $\mathrm{IPFS} = \mathrm{E}_M\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{PFS}(\mathbf{X}_0)\right]$.

4. Analysis for the Case of Unequal n_i's

Let n_i be the number of simulation replications allocated to each of the m covariate points with design i, $i=1,\ldots,k$. In this section, we fix the number of covariate points m, allow n_i to be unequal among different designs i, and develop a ranking and selection (R&S) framework for optimizing the simulation budget allocation n_i's in simulation with covariates introduced in the main text.

Suppose that the m covariate points collected are $\mathbf{x}_1,\ldots,\mathbf{x}_m$, and the total simulation budget to be allocated among pairs of covariate points and designs is n_{tot}, i.e., $m\sum_{i=1}^k n_i = n_{tot}$. With the target measures of the maximal IMSE and IPFS, the corresponding R&S problems can be formulated as

$$\begin{aligned} &\min \max_{i\in\{1,2,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \\ &\text{s.t. } m\sum_{i=1}^k n_i = n_{tot}, \text{ and } n_i \geq 0, \text{ for } i=1,\ldots,k, \end{aligned} \tag{67}$$

and

$$\begin{aligned} &\min \mathrm{E}_M\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{PFS}(\mathbf{X}_0)\right] \\ &\text{s.t. } m\sum_{i=1}^k n_i = n_{tot}, \text{ and } n_i \geq 0, \text{ for } i=1,\ldots,k, \end{aligned} \tag{68}$$

However, both optimization problems (67) and (68) cannot be directly solved due to the lack of analytical expressions of the objective functions $\max_{i\in\{1,2,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right]$ and $\mathrm{E}_M\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{PFS}(\mathbf{X}_0)\right]$. Here we propose two methods to approximate them.

Our first proposal is to replace the maximal IMSE and IPFS in (67) and (68) with their Monte Carlo estimators proposed in Section 3. Both $\max_{i\in\{1,\ldots,k\}}\widehat{\mathrm{IMSE}}_i$ defined in (64) and $\widehat{\mathrm{IPFS}}$ defined in (66) have already taken into account the unequal n_i's in the matrix $\mathbf{\Sigma}_{\epsilon,i}(\mathbf{X}^m)$. We can choose the Monte Carlo sample size m' according to the optimization budget. Then (67) and (68) can be solved using numerical optimization methods.

Our second proposal is to approximate them by the analytical upper bounds in our Theorems 1 and 2. Note that analytical approximations are common in solving R&S problems, especially in the OCBA method (Chen et al. 2000, 2008). They make the optimization problem tractable, and can often lead to efficient budget allocation rules.

Let $\{\mu_{i,l} : l=1,2,\ldots\}$ be the eigenvalues of the linear operator $T_{\mathbf{\Sigma}_{M,i}}$ defined in Section 2.1 of the main text. We recall from the second paragraph after Assumptions A.1-A.4 that the constants r_* and ρ_* in Assumption A.3 can be made common for all the k designs. Using the results in Theorems 1, 2 and 5, we can prove the following proposition.

PROPOSITION 1. *Suppose that Assumptions A.1 - A.4 in the main text hold for all the k designs. Let $\varrho_i = mn_i/n_{tot}$. For any $0 \le \varrho \le 1$, define the following quantities for $i = 1, \ldots, k$:*

$$\begin{aligned}R_i(\varrho) = &\frac{2\overline{\sigma}_0^2}{n_{tot}\varrho}\gamma_i\left(\frac{\overline{\sigma}_0^2}{n_{tot}\varrho}\right) + \frac{64C_i^\dagger q\overline{\sigma}_0^2\,\mathrm{tr}(\mathbf{\Sigma}_{M,i})}{n_{tot}\varrho} \\ &+ \inf_{\zeta\in\mathbb{N}}\left[\left\{\frac{64C_i^\dagger q\rho_*^4\overline{\sigma}_0^2}{\underline{\sigma}_0^4}\mathrm{tr}\left(\mathbf{\Sigma}_{M,i}\right)^2 + 8C_i^\dagger q\,\mathrm{tr}\left(\mathbf{\Sigma}_{M,i}\right) + \frac{3}{\underline{\sigma}_0^2}\mathrm{tr}(\mathbf{\Sigma}_{M,i}) + 1\right\}\mathrm{tr}\left(\mathbf{\Sigma}_{M,i}^{(\zeta)}\right)n_{tot}\varrho\right. \\ &\left. + \left[8C_i^\dagger q\,\mathrm{tr}\left(\mathbf{\Sigma}_{M,i}\right)^2 + \mathrm{tr}\left(\mathbf{\Sigma}_{M,i}\right)\right]\left\{300\rho_*^2\frac{b(m,\zeta,r_*)}{\sqrt{m}}\gamma_i\left(\frac{\overline{\sigma}_0^2}{n_{tot}\varrho}\right)\right\}^{r_*}\right],\end{aligned} \tag{69}$$

where A is the universal constant in Theorem 1 and

$$\begin{aligned}&C_i^\dagger = C_{\mathrm{f},i}^2/\lambda_{\min}\left(\mathrm{E}_{\mathbf{X}}[\mathbf{f}(\mathbf{X})\mathbf{f}(\mathbf{X})^\top]\right), \quad C_{\mathrm{f},i} = \max_{1\le s\le q}\|\mathrm{f}_s\|_{\mathbb{H}_i} \\ &\gamma_i(a) = \sum_{l=1}^{\infty}\frac{\mu_{i,l}}{\mu_{i,l}+a} \text{ for any } a > 0, \\ &\mathrm{tr}\left(\mathbf{\Sigma}_{M,i}\right) = \sum_{l=1}^{\infty}\mu_{i,l}, \quad \mathrm{tr}\left(\mathbf{\Sigma}_{M,i}^{(\zeta)}\right) = \sum_{l=\zeta+1}^{\infty}\mu_{i,l} \text{ for any } \zeta\in\mathbb{N}, \\ &b(m,\zeta,r_*) = \max\left(\sqrt{\max(r_*,\log\zeta)},\ \frac{\max(r_*,\log\zeta)}{m^{1/2-1/r_*}}\right).\end{aligned}$$

Then, for the measures of the maximal IMSE and IPFS, we have

$$\max_{i\in\{1,\ldots,k\}}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \max_{i\in\{1,\ldots,k\}} R_i(\varrho_i), \tag{70}$$

$$\mathrm{E}_M\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{PFS}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \max_{i\in\{1,\ldots,k\}} R_i(\varrho_i),. \tag{71}$$

Proof of Proposition 1:

By directly combining the upper bounds in Theorems 1 and 2 together with the MSE decomposition in Equation (6) of the main text, we have that with $\mathbb{P}_{\mathbf{X}^m}-$ probability approaching 1, for each $i = 1, \ldots, k$,

$$\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] = \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(M)}(\mathbf{X}_0)\right] + \mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}^{(\boldsymbol{\beta})}(\mathbf{X}_0)\right] \le R_i(mn_i/n_{tot}) = R_i(\varrho_i),$$

where $R_i(\cdot)$ is defined in (69) above and is slightly larger than the combined upper bounds from Theorems 1 and 2 by adjusting some constants. This implies the following upper bound

$$\max_{i\in\{1,\ldots,k\}}\mathrm{E}_{\mathbf{X}_0}\left[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)\right] \lesssim_{\mathbb{P}_{\mathbf{X}^m}} \max_{i\in\{1,\ldots,k\}} R_i(\varrho_i),$$

which proves (70).

For the IPFS measure, we notice that for each of the three cases in Theorem 5, the upper bound is a monotone increasing function of $R(m,n)$, which is defined as a probablistic upper bound for $\max_{i\in\{1,\dots,k\}} \mathrm{E}_{\mathbf{X}_0}[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)]$ in Theorem 3 of the main text. Therefore, the inequality in (71) holds as well. $\square$

With Proposition (1), we can build an analytical R&S model for both (67) and (68),

$$\begin{aligned} &\min \max_{i\in\{1,\dots,k\}} R_i(\varrho_i) \\ &\text{s.t.} \sum_{i=1}^{k} \varrho_i = 1, \text{ and } \varrho_i \geq 0, \text{ for } i=1,\dots,k. \end{aligned} \tag{72}$$

This is a typical nonlinear optimization problem. Its optimal solution ϱ_i^* gives us an approximately optimal allocation of the simulation budget among pairs of covariate points and designs with $n_i^* = \frac{\varrho_i^* n_{tot}}{m}$, $i = 1, 2, ..., k$.

Problem (72) involves a number of constants that depend on the properties of the covariance kernels used in the k designs. These constants can be made concrete when the covariate space $\mathcal{X}$, the covariance kernels $\mathbf{\Sigma}_{M,i}$, the sampling distribution $\mathbb{P}_{\mathbf{X}}$, and the regression functions $\mathbf{f}_{i1}, \dots, \mathbf{f}_{iq}$ are fully specified in practice. Problem (72) is not necessarily a convex optimization problem. Since it is built based on a different setting (fixed m and unequal n_i's) from that of the main questions in this research, we do not pursue further development of it in this paper. We emphasize that our proposed theoretical analysis and results can be used to formulate and solve R&S type of problems that arise in simulation with covariates.

5. Additional Numerical Results

This section provides additional numerical results to the main text. Section 5.1 plots the two test functions in Section 5.1 of the main text. Section 5.2 compares our static sampling with an adaptive design procedure, under the target measures of the maximal IMSE and IPFS. Section 5.3 provides a procedure that can help the analyst make the design decision for achieving a target precision of the maximal IMSE.

5.1. Plots of the Test Functions used in Section 5.1 of the Main Text

The 1-d De Jong's function and Griewank's function without noise are shown in Figure 1. The 2-d De Jong's function under selected designs without noise is shown in Figure 2.

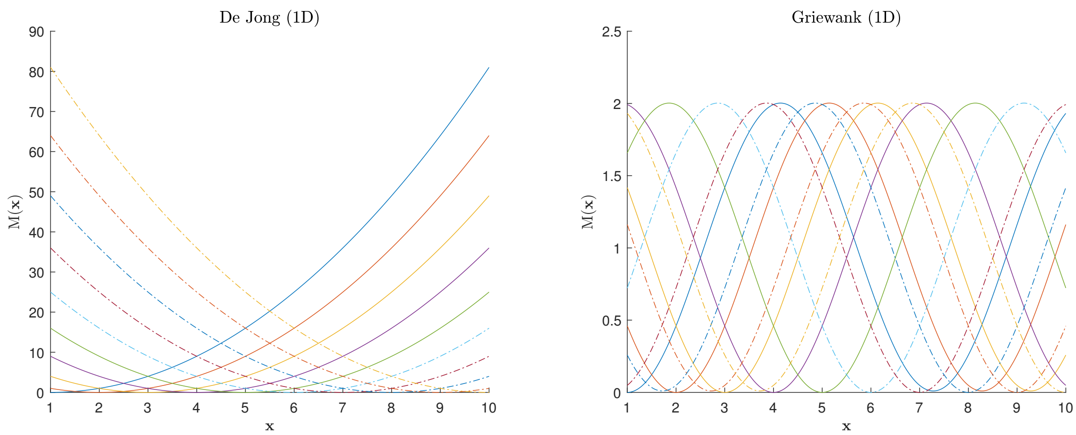

Figure 1 Plots of the two test functions for 1-dimensional x. Ten different curves stand for the ten designs.

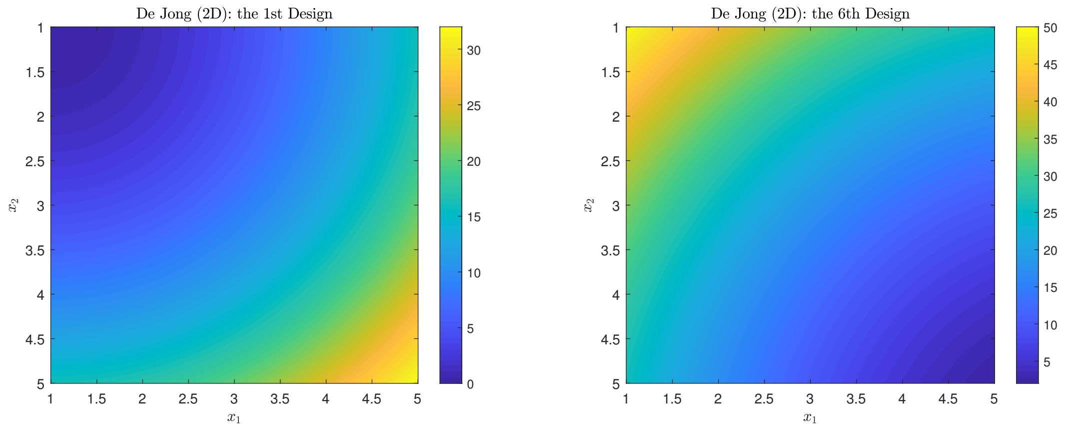

Figure 2 Heatmaps of De Jong's function for 2-dimensional x under the 1st and 6th design ($i = 1$ and $i = 6$).

In the 1-d case (Figure 1), we present ten curves for the two test functions, each corresponding to $M(\mathbf{x})$ at one of the ten designs. It can be observed that no design can dominate the others in the tested functions, and the best design might not be unique for some $\mathbf{x}$. The De Jong's functions are smooth while the Griewank's functions are highly nonlinear with many oscillations, which brings difficulty to SK modeling when the number of covariate points m is small. In the 2-d case (Figure 2), we present the heatmap of the 1st and 6th design ($i = 1$ and $i = 6$) for the De Jong's function. We can see that $M(\mathbf{x})$ varies a lot with $\mathbf{x}$.

5.2. Comparison between Static Sampling and an Adaptive Procedure

In this section, we compare our static sampling (i.e., fixed-distribution sampling; this is the sampling method studied in this research) with an intuitive adaptive design procedure. The adaptive procedure works in a greedy manner and iteratively collects the covariate point that maximizes the largest MSE of the fitted SK models. In this way, it sequentially explores the whole covariate space and reduces the overall MSE of the SK prediction. We call it Adaptive MSE Procedure.

Suppose that m_1 covariate points $\mathbf{x}^{m_1} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{m_1}\}$ have been sampled already. For the illustration purpose, we will use the superscript $[m_1]$ to indicate that the SK estimators are derived from the current simulation samples $\mathbf{x}^{m_1}$. From Equation (3) in the main text, the mean squared error of the current-stage SK predictor of design i at $\mathbf{x}_0$ is

$$\begin{aligned}\mathrm{MSE}_{i,\mathrm{opt}}^{[m_1]}(\mathbf{x}_0) = {} & \boldsymbol{\Sigma}_{M,i}(\mathbf{x}_0, \mathbf{x}_0) - \boldsymbol{\Sigma}_{M,i}^\top(\mathbf{x}^{m_1}, \mathbf{x}_0)\left[\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^{m_1}, \mathbf{x}^{m_1}) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^{m_1})\right]^{-1} \boldsymbol{\Sigma}_{M,i}(\mathbf{x}^{m_1}, \mathbf{x}_0) \\ & + \eta_i^{[m_1]}(\mathbf{x}_0)^\top \left[(\mathbf{F}_i^{[m_1]})^\top \left(\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^{m_1}, \mathbf{x}^{m_1}) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^{m_1})\right)^{-1} \mathbf{F}_i^{[m_1]}\right]^{-1} \eta_i^{[m_1]}(\mathbf{x}_0),\end{aligned} \tag{73}$$

where $\eta_i^{[m_1]}(\mathbf{x}_0) = \mathbf{f}_i(\mathbf{x}_0) - (\mathbf{F}_i^{[m_1]})^\top \left(\boldsymbol{\Sigma}_{M,i}(\mathbf{x}^{m_1}, \mathbf{x}^{m_1}) + \boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^{m_1})\right)^{-1} \boldsymbol{\Sigma}_{M,i}(\mathbf{x}^{m_1}, \mathbf{x}_0)$, $\mathbf{F}_i^{[m_1]} = (\mathbf{f}_i(\mathbf{x}_1), \ldots, \mathbf{f}_i(\mathbf{x}_{m_1}))^\top$, and $\boldsymbol{\Sigma}_{\epsilon,i}(\mathbf{x}^{m_1})$ is the $m_1 \times m_1$ covariance matrix of the averaged simulation errors across m_1 covariate points under design i.

The Adaptive MSE Procedure samples the next covariate point $\mathbf{x}_{m_1+1}$ with the largest maximal $\mathrm{MSE}_{i,\mathrm{opt}}^{[m_1]}(\mathbf{x}_0)$, where "largest" is over the covariate space $\mathcal{X}$ and "maximal" is over the k SK models. That is,

$$\mathbf{x}_{m_1+1} = \arg\max_{\mathbf{x}_0 \in \mathcal{X}} \max_{i \in \{1,\ldots,k\}} \mathrm{MSE}_{i,\mathrm{opt}}^{[m_1]}(\mathbf{x}_0). \tag{74}$$

The formal description of the Adaptive MSE Procedure is given as follows.

Adaptive MSE Procedure

1. Specify the covariate space $\mathcal{X}$ and the total number of covariate points m. Perform n_0 replications for the pair of the center point of the covariate space and design i, $i = 1, \ldots, k$. $m_1 \leftarrow 0$.
2. If $m_1 > m$, stop. Otherwise,
   a. Obtain $\mathbf{x}_{m_1+1}$ by (74).
   b. Perform n_0 replications for the pair of covariate point $\mathbf{x}_{m_1+1}$ and design i, $i = 1, \ldots, k$.
   c. Update the SK model for each design $i = 1, \ldots, k$. $m_1 \leftarrow m_1 + 1$.

We use the De Jong's and Griewank's functions under the same parameter settings as in Section 5.1 of the main text for testing, i.e., the covariate space is $\mathcal{X} = [1, 10]^d$ and there are

$k = 10$ designs. Meanwhile, we vary the domain dimension d and the sampling distribution $\mathbb{P}_{\mathbf{X}}$. Specifically, we test the following examples on our static sampling from $\mathbb{P}_{\mathbf{X}}$ and the Adaptive MSE Procedure:

(i) De Jong's functions, for dimension $d = 1$, $\mathbb{P}_{\mathbf{X}}$ being the truncated $N(5.5, 1^2)$;
(ii) De Jong's functions, for dimension $d = 1$, $\mathbb{P}_{\mathbf{X}}$ being the truncated $N(5.5, 0.25^2)$;
(iii) De Jong's functions, for dimension $d = 2$, $\mathbb{P}_{\mathbf{X}}$ being the truncated $N(5.5, 0.3^2)$ in each dimension;
(iv) De Jong's functions, for dimension $d = 3$, $\mathbb{P}_{\mathbf{X}}$ being the truncated $N(2.5, 0.3^2)$ in each dimension;
(v) Griewank's functions, for dimension $d = 1$, $\mathbb{P}_{\mathbf{X}}$ being the uniform distribution on $[1, 10]$;
(vi) Griewank's functions, for dimension $d = 1$, $\mathbb{P}_{\mathbf{X}}$ being the truncated $N(5.5, 1^2)$;
(vii) Griewank's functions, for dimension $d = 10$, $\mathbb{P}_{\mathbf{X}}$ being the uniform distribution on $[1, 10]^{10}$;
(viii) Griewank's functions, for dimension $d = 10$, $\mathbb{P}_{\mathbf{X}}$ being the truncated $N(2.5, 0.75^2)$ in each dimension.

Figures 3-10 report the comparison results for our static sampling and the Adaptive MSE Procedure under the measures of the maximal IMSE and IPFS. We have the following observations:

- For Case (i), where $d = 1$ and the De Jong's functions are smooth enough, the Adaptive MSE Procedure has smaller maximal IMSE and IPFS than the static sampling from $\mathbb{P}_{\mathbf{X}}$.
- For Cases (ii), (iii), and (iv) with the De Jong's functions in dimension $d = 1, 2, 3$, where the normal variance becomes smaller, i.e., the sampling distribution $\mathbb{P}_{\mathbf{X}}$ becomes more concentrated, the static sampling from $\mathbb{P}_{\mathbf{X}}$ has slightly better performance than the Adaptive MSE Procedure, but overall their performances are similar.
- For Case (v), where $d = 1$, the sampling distribution $\mathbb{P}_{\mathbf{X}}$ is uniform, and the target is the Griewank's functions, we can see from Figure 7 that the Adaptive MSE Procedure has slightly smaller maximal IMSE and IPFS than the static sampling, but overall their performances are similar.
- For Case (vi), where $d = 1$, the sampling distribution $\mathbb{P}_{\mathbf{X}}$ is truncated normal with a moderately large variance, and the target Griewank's functions have strong oscillation, we can see from Figure 8 that the static sampling almost always yields smaller maximal IMSE and IPFS than the Adaptive MSE Procedure.

- For Cases (vii) and (viii), where the dimension is high ($d = 10$) and the target Griewank's functions have strong oscillation, we can see from Figures 9 and 10 that the static sampling always yields much smaller maximal IMSE and IPFS than the Adaptive MSE Procedure, for both the uniform distribution and the truncated normal distribution.

In conclusion, the static sampling from $\mathbb{P}_{\mathbf{X}}$ seems to yield comparable performance to the Adaptive MSE Procedure under the two measures in general. The static sampling tends to perform better than the Adaptive MSE Procedure when the target function has strong oscillation, the dimension becomes higher, and the covariate distribution $\mathbb{P}_{\mathbf{X}}$ becomes more concentrated.

5.3. Achieving a Target Precision of the Maximal IMSE

Based on the linear decreasing trend of the maximal IMSE in Figure 5 of the main text, we propose a simple procedure to determine the sample size m_0 such that the maximal IMSE satisfies $\max_{i\in\{1,\ldots,k\}} \mathrm{E}_{\mathbf{X}_0}[\mathrm{MSE}_{i,\mathrm{opt}}(\mathbf{X}_0)] = c_0$ for a target precision c_0. Suppose that we have already drawn m covariate points $\mathbf{X}^m$ from $\mathbb{P}_{\mathbf{X}}$ and each covariate point has n simulation replications. Then, for an integer $L \geq 3$, we draw $L-1$ subsamples of sizes $m_1 < \ldots < m_{L-1} (< m_L \equiv m)$ from $\mathbf{X}^m$ without replacement. Denote these subsets as $\mathbf{X}^{m_1}, \ldots, \mathbf{X}^{m_{L-1}}$. We then fit $(L-1)k$ SK models based on each dataset of $\mathbf{X}^{m_1}, \ldots, \mathbf{X}^{m_{L-1}}$, and estimate the maximal IMSE for each subset using the Monte Carlo estimator described in Section 3 of the Online Supplement. We repeat this subsampling-fitting-estimating process for multiple times and take the average of the estimated maximal IMSE's at each size $m_1, \ldots, m_{L-1}, m_L$, denoted by $\max_{i\in\{1,\ldots,k\}} \widehat{\mathrm{IMSE}}_i(m_l)$, $l = 1, \ldots, L$. Finally, we fit the linear model $\log(\max_{i\in\{1,\ldots,k\}} \mathrm{IMSE}_i) = c_1 + c_2 \log m + \text{error}$ using the pairs $\left\{\left(\max_{i\in\{1,\ldots,k\}} \widehat{\mathrm{IMSE}}_l(m_l), m_l\right) : l = 1, \ldots, L\right\}$, and predict m_0 by $\widehat{m}_0 = \exp\{(\log c_0 - \widehat{c}_1)/\widehat{c}_2\}$, where $\widehat{c}_1, \widehat{c}_2$ are the fitted linear coefficients.

Next, we apply this procedure to the M/M/1 queue example. We draw $m = 80$ covariate points from the sampling distribution $\mathbb{P}_{\mathbf{X}}$ with $n = 10$ replications, and estimate the maximal IMSE with subsample sizes $\{10, 15, 23, 35, 53, 80\}$. For example, for the squared exponential kernel and uniform sampling distribution, we obtain the fitted linear regression model $\log(\max_{i\in\{1,\ldots,k\}} \mathrm{IMSE}_i) = -1.03 \log(m) - 4.58$ and the predicted $\widehat{m}_0 \approx 119$ such that $\max_{i\in\{1,\ldots,k\}} \mathrm{IMSE}_i = c_0 = 7.5 \times 10^{-5}$. To numerically verify whether the true maximal IMSE is around 7.5×10^{-5} at sample size $\widehat{m}_0 = 119$, we randomly draw another 39 covariate points

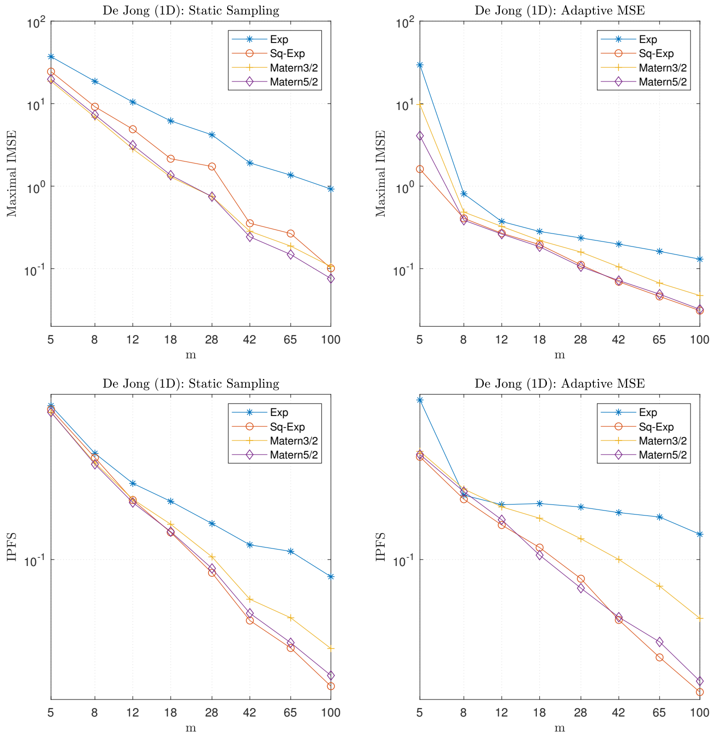

Figure 3 Truncated $N(5.5, 1^2)$ of $d=1$: The maximal IMSE and IPFS for the 1-dimensional De Jong's functions and four covariance kernels.

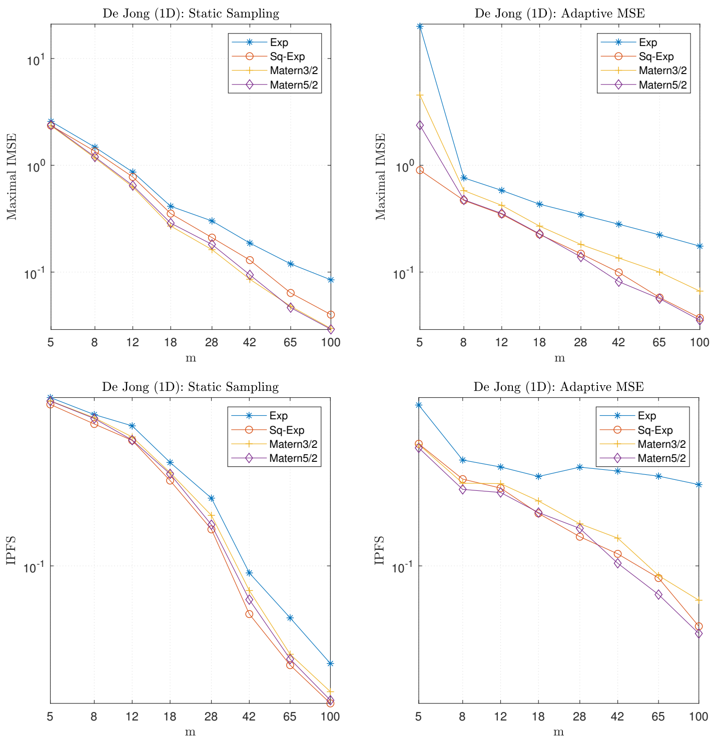

Figure 4 Truncated $N(5.5, 0.25^2)$ of $d=1$: The maximal IMSE and IPFS for the 1-dimensional De Jong's functions and four covariance kernels.

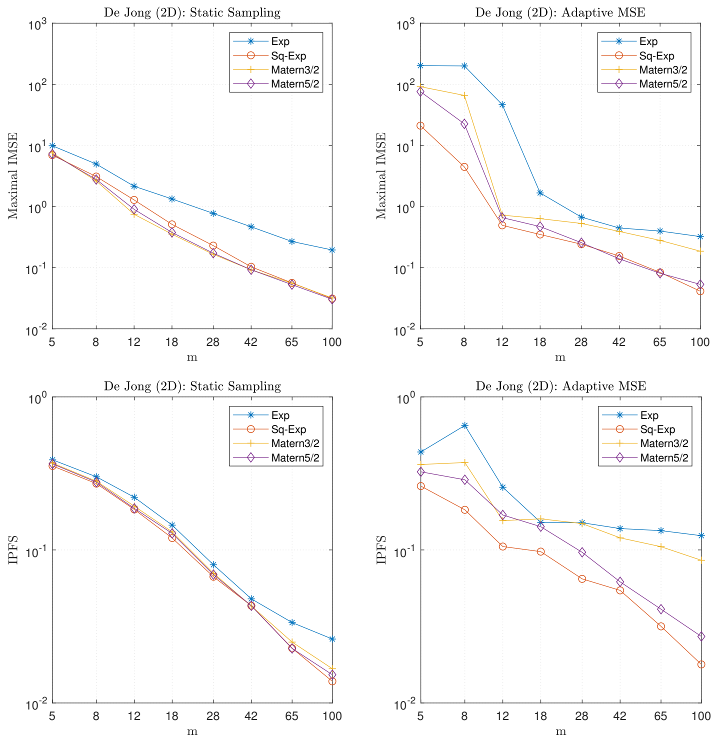

Figure 5 Truncated $N(5.5, 0.3^2)$ on each dimension of $d = 2$: The maximal IMSE and IPFS for the 2-dimensional De Jong's functions and four covariance kernels.

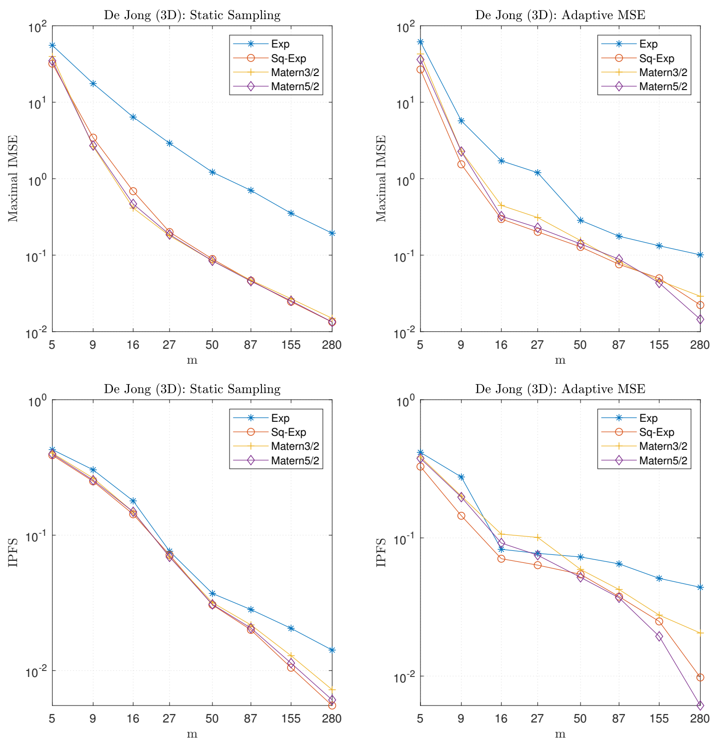

Figure 6 Truncated $N(2.5, 0.3^2)$ on each dimension of $d = 3$: The maximal IMSE and IPFS for the 3-dimensional De Jong's functions and four covariance kernels.

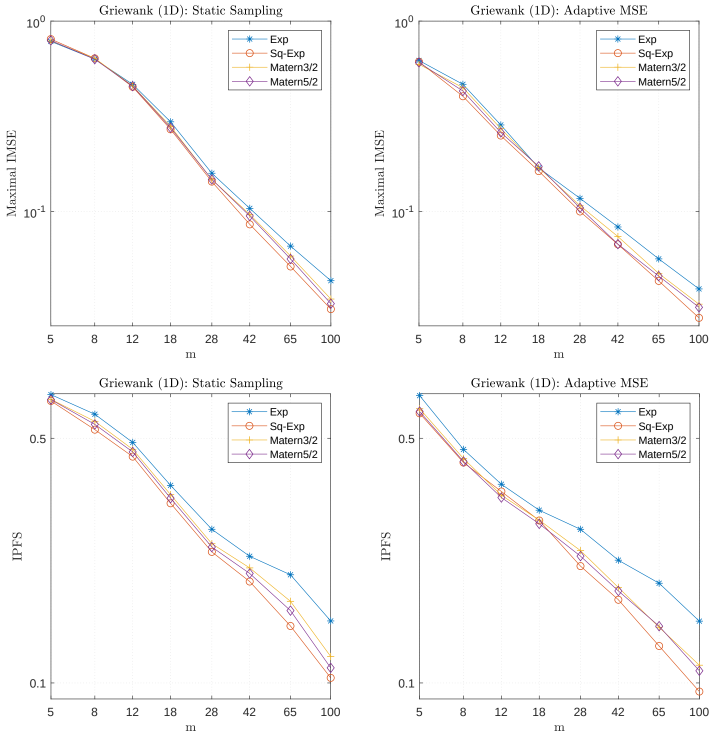

Figure 7 **Uniform distribution of $d = 1$: The maximal IMSE and IPFS for the 1-dimensional Griewank's functions and four covariance kernels.**

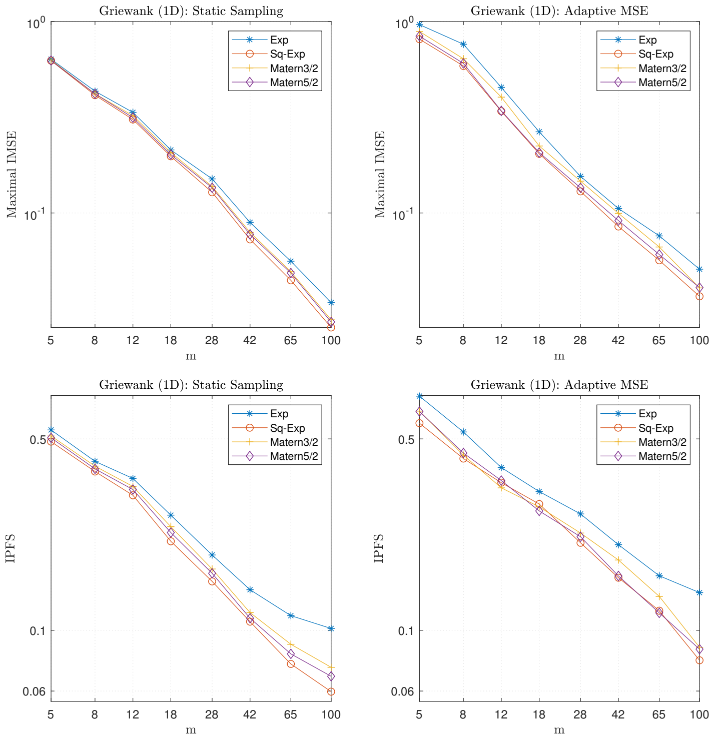

Figure 8 **Truncated $N(5.5, 1^2)$ of $d = 1$: The maximal IMSE and IPFS for the 1-dimensional Griewank's functions and four covariance kernels.**

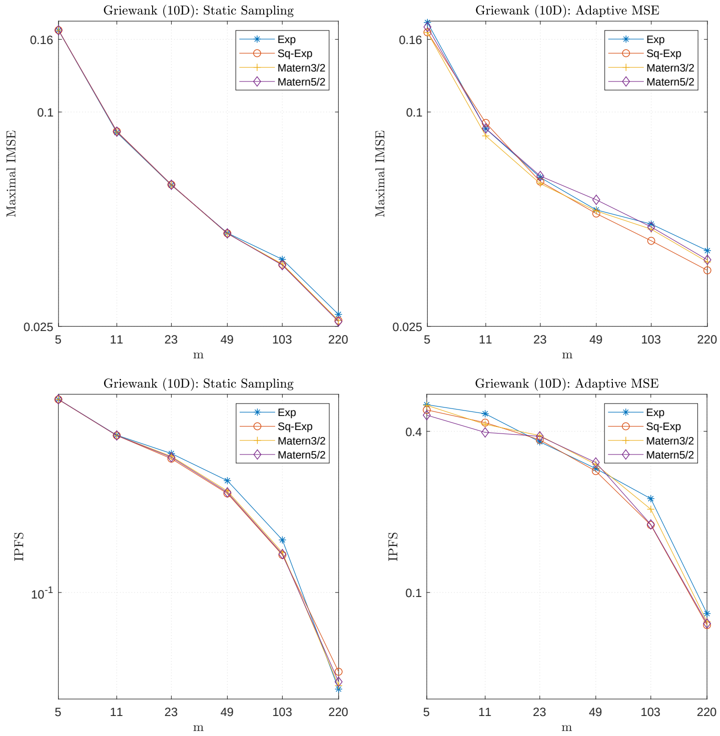

Figure 9 **Uniform distribution of $d = 10$: The maximal IMSE and IPFS for the 10-dimensional Griewank's functions and four covariance kernels.**

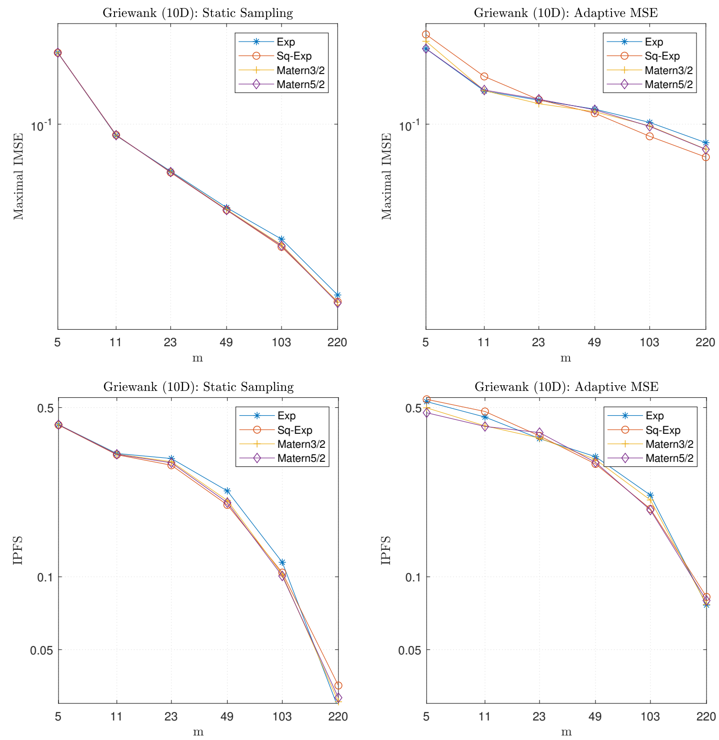

Figure 10 **Truncated $N(2.5, 0.75^2)$ on each dimension of $d = 10$: The maximal IMSE and IPFS for the 10-dimensional Griewank's functions and four covariance kernels.**

from the uniform distribution, establish the SK models based on the union of the 39 new points and the 80 existing points, and compute the maximal IMSE. We repeat this process for 40 macro Monte Carlo replications. We find that the median maximal IMSE over the 40 macro replications is 7.37×10^{-5}. The numerical results for the two tested sampling distributions and four covariance kernels are summarized in Table 2. In almost all cases, the predicted $\widehat{m}_0$ values yield very similar or smaller maximal IMSE's compared to the target values. This demonstrates that our theory can help the decision makers determine the number of additional covariate points needed to achieve a target precision.

Table 2 Prediction of sample size m_0 for a maximal IMSE precision c_0 based on $m = 80$ covariate points.

	Kernels	c_0	$\widehat{c}_1$	$\widehat{c}_2$	$\widehat{m}_0$	Mean	Median
uniform, $n = 10$	`SqExp`	7.5×10^{-5}	-1.03	-4.58	119	7.37×10^{-5}	7.37×10^{-5}
	`Matern 5/2`	7.5×10^{-5}	-1.12	-4.06	130	7.07×10^{-5}	6.95×10^{-5}
	`Matern 3/2`	7.5×10^{-5}	-1.12	-3.92	147	7.19×10^{-5}	6.95×10^{-5}
	`Exp`	2.0×10^{-4}	-0.95	-3.99	118	2.23×10^{-4}	2.09×10^{-4}
truncated normal, $n = 10$	`SqExp`	7.5×10^{-5}	-1.00	-4.70	122	7.44×10^{-5}	7.04×10^{-5}
	`Matern 5/2`	7.5×10^{-5}	-1.06	-4.24	144	6.57×10^{-5}	6.45×10^{-5}
	`Matern 3/2`	7.5×10^{-5}	-1.08	-4.04	160	6.67×10^{-5}	6.69×10^{-5}
	`Exp`	2.0×10^{-4}	-0.95	-4.00	117	2.34×10^{-4}	2.11×10^{-4}

Notes: "Mean" is the sample average of the maximal IMSE over 40 macro Monte Carlo replications. "Median" is the sample median of the maximal IMSE over 40 macro Monte Carlo replications.

References

Chen CH, He D, Fu M, Lee LH (2008) Efficient simulation budget allocation for selecting an optimal subset. *INFORMS J. Comput.* 20(4):579–595.

Chen CH, Lin J, Yücesan E, Chick SE (2000) Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dyn. S.* 10:251–270.

Hsu D, Kakade SM, Zhang T (2012) A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability* 17(52):1–6.

Kreh M (2012) *Bessel Functions. Lecture Notes, Penn State - Göttingen Summer School on Number Theory.*

Rasmussen CE, Williams CK (2006) *Gaussian process for machine learning* (MIT press).

Zhang Y, Duchi JC, Wainwright MJ (2015) Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.* 16:3299–3340.