Online Supplement for the Paper Titled "Gaussian Process Modeling of a Functional Output with Information from Boundary and Initial Conditions and Analytical Approximations"

Appendix A: Proof that maximum likelihood estimator of $\boldsymbol{\Psi}$ does not exists

We shall show that there does not exist a positive definite $\boldsymbol{\Psi}$ that minimizes $\mathrm{trace}[\boldsymbol{\Psi}^{-1}(\boldsymbol{Z}_* - \widehat{\boldsymbol{\alpha}}\boldsymbol{1}_n^T)\boldsymbol{R}_2^{-1}(\boldsymbol{Z}_* - \widehat{\boldsymbol{\alpha}}\boldsymbol{1}_n^T)^T] + n\log|\boldsymbol{\Psi}|$. Let the eigendecomposition of $(\boldsymbol{Z}_* - \widehat{\boldsymbol{\alpha}}\boldsymbol{1}_n^T)\boldsymbol{R}_2^{-1}(\boldsymbol{Z}_* - \widehat{\boldsymbol{\alpha}}\boldsymbol{1}_n^T)^T$ be denoted by $(\boldsymbol{Z}_* - \widehat{\boldsymbol{\alpha}}\boldsymbol{1}_n^T)\boldsymbol{R}_2^{-1}(\boldsymbol{Z}_* - \widehat{\boldsymbol{\alpha}}\boldsymbol{1}_n^T)^T = \boldsymbol{E}\boldsymbol{L}\boldsymbol{E}^T$, where $\boldsymbol{E}$ is a matrix with orthonormal eigenvectors as columns and $\boldsymbol{L} = \mathrm{diag}\{L_1, \dots, L_N\}$ is a diagonal matrix of eigenvalues ordered such that $L_1 \ge L_2 \ge \cdots \ge L_N$. Since $N > n$, $L_{m+1} = \cdots = L_N = 0$ and $L_m > 0$ for some $m \le n$. Let us replace $L_{m+1}, \dots, L_N$ in $\boldsymbol{L}$ with $l > 0$, which gives the diagonal matrix $\tilde{\boldsymbol{L}} = \mathrm{diag}\{L_1, \dots, L_m, l, \dots, l\}$. Now, let $\boldsymbol{\Psi} = \boldsymbol{E}\tilde{\boldsymbol{L}}\boldsymbol{E}^T$. Then, since $\boldsymbol{\Psi}^{-1} = \boldsymbol{E}\tilde{\boldsymbol{L}}^{-1}\boldsymbol{E}^T$ and $|\boldsymbol{\Psi}| = |\tilde{\boldsymbol{L}}| = l^{N-m}\prod_{i=1}^m L_i$, we have

$$
\begin{aligned}
&\mathrm{trace}[\boldsymbol{\Psi}^{-1}(\boldsymbol{Z}_* - \widehat{\boldsymbol{\alpha}}\boldsymbol{1}_n^T)\boldsymbol{R}_2^{-1}(\boldsymbol{Z}_* - \widehat{\boldsymbol{\alpha}}\boldsymbol{1}_n^T)^T] + n\log|\boldsymbol{\Psi}| \\
&= \mathrm{trace}\left(\boldsymbol{E}\tilde{\boldsymbol{L}}^{-1}\boldsymbol{E}^T\boldsymbol{E}\boldsymbol{L}\boldsymbol{E}^T\right) + n\log\left(l^{N-m}\prod_{i=1}^m L_i\right) \\
&= \mathrm{trace}\left(\boldsymbol{E}\tilde{\boldsymbol{L}}^{-1}\boldsymbol{L}\boldsymbol{E}^T\right) + n(N-m)\log(l) + n\sum_{i=1}^m \log(L_i) \\
&= \mathrm{trace}(\mathrm{diag}\{1,\dots,1,0,\dots,0\}\boldsymbol{E}^T\boldsymbol{E}) + n(N-m)\log(l) + n\sum_{i=1}^m \log(L_i) \\
&= \mathrm{trace}(\mathrm{diag}\{1,\dots,1,0,\dots,0\}) + n(N-m)\log(l) + n\sum_{i=1}^m \log(L_i) \\
&= m + n(N-m)\log(l) + n\sum_{i=1}^m \log(L_i). \qquad \text{(A1)}
\end{aligned}
$$

Observe that we can make (A1) arbitrarily small by taking l arbitrarily small. Thus, there is no maximizer of the likelihood function in the set of positive definite matrices.

Appendix B: Condition for separability of covariance function

As given by (13), $\mathrm{cov}[\boldsymbol{Y}(\boldsymbol{x}), \boldsymbol{Y}(\boldsymbol{x}')] = \sum_{i=1}^N \lambda_i \mathcal{R}_i(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\theta}_i)\boldsymbol{\Phi}_i\boldsymbol{\Phi}_i^T$. Suppose that the covariance function also satisfies $\mathrm{cov}[\boldsymbol{Y}(\boldsymbol{x}), \boldsymbol{Y}(\boldsymbol{x}')] = g(\boldsymbol{x}, \boldsymbol{x}')\boldsymbol{G}$ for all $\boldsymbol{x}$ and $\boldsymbol{x}'$. Then, by setting $\boldsymbol{x}' = \boldsymbol{x}$, we get $\boldsymbol{G} = a\sum_{i=1}^N \lambda_i \boldsymbol{\Phi}_i\boldsymbol{\Phi}_i^T$ and $g(\boldsymbol{x}, \boldsymbol{x}) = 1/a$. We can choose $a = 1$ without loss of generality. Thus, we obtain

$$\sum_{i=1}^{N} \lambda_i \mathcal{R}_i(\boldsymbol{x},\boldsymbol{x}'|\boldsymbol{\theta}_i)\boldsymbol{\Phi}_i\boldsymbol{\Phi}_i^T = g(\boldsymbol{x},\boldsymbol{x}')\boldsymbol{G} = g(\boldsymbol{x},\boldsymbol{x}')\sum_{i=1}^{N} \lambda_i \boldsymbol{\Phi}_i\boldsymbol{\Phi}_i^T. \tag{B1}$$

Suppose $g(\boldsymbol{x},\boldsymbol{x}') = 0$, which gives $\sum_{i=1}^{N} \boldsymbol{\Phi}_i \lambda_i \mathcal{R}_i(\boldsymbol{x},\boldsymbol{x}'|\boldsymbol{\theta}_i)\boldsymbol{\Phi}_i^T = \boldsymbol{0}$. This implies that $\mathcal{R}_1(\boldsymbol{x},\boldsymbol{x}'|\boldsymbol{\theta}_i) = \cdots = \mathcal{R}_N(\boldsymbol{x},\boldsymbol{x}'|\boldsymbol{\theta}_i) = 0$ because $\boldsymbol{\Phi}_1,\dots,\boldsymbol{\Phi}_N$ are linearly independent, each $\lambda_i > 0$, and there is at least one nonzero element in each $\boldsymbol{\Phi}_i$. Suppose $g(\boldsymbol{x},\boldsymbol{x}') \neq 0$. Then

$$\sum_{i=1}^{N} \boldsymbol{\Phi}_i \lambda_i [\mathcal{R}_i(\boldsymbol{x},\boldsymbol{x}'|\boldsymbol{\theta}_i)/g(\boldsymbol{x},\boldsymbol{x}')]\boldsymbol{\Phi}_i^T = \sum_{i=1}^{N} \lambda_i \boldsymbol{\Phi}_i\boldsymbol{\Phi}_i^T. \tag{B2}$$

Note that the right hand side of (B2) is independent of $(\boldsymbol{x},\boldsymbol{x}')$. Since $\lambda_1 \geq \cdots \geq \lambda_N > 0$ and $\boldsymbol{\Phi}_1,\dots,\boldsymbol{\Phi}_N$ are linearly independent, we must have $\mathcal{R}_i(\boldsymbol{x},\boldsymbol{x}'|\boldsymbol{\theta}_i)/g(\boldsymbol{x},\boldsymbol{x}') = 1$ for all $i = 1,\dots,N$. Thus, $\mathcal{R}_i(\boldsymbol{x},\boldsymbol{x}'|\boldsymbol{\theta}_i) = g(\boldsymbol{x},\boldsymbol{x}')$ for all $i = 1,\dots,N$ and all $(\boldsymbol{x},\boldsymbol{x}')$.

Appendix C: Proof of interpolation property

Let $\hat{\lambda}_1,\dots,\hat{\lambda}_J$ be the nonzero eigenvalues of $\boldsymbol{S}$ and $\widehat{\boldsymbol{\Phi}}_1,\dots,\widehat{\boldsymbol{\Phi}}_J$ be the corresponding eigenvectors. Then, we should choose $\{\hat{\xi}_i(\boldsymbol{x}_j), i = 1,\dots,J, j = 1,\dots,n\}$ so that

$$\left(\widehat{\boldsymbol{\Phi}}_1,\dots,\widehat{\boldsymbol{\Phi}}_J\right)\mathrm{diag}\{\hat{\lambda}_1^{1/2},\dots,\hat{\lambda}_J^{1/2}\}\begin{pmatrix} \hat{\xi}_1(\boldsymbol{x}_1) & \cdots & \hat{\xi}_1(\boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ \hat{\xi}_J(\boldsymbol{x}_1) & \cdots & \hat{\xi}_J(\boldsymbol{x}_n) \end{pmatrix} = \boldsymbol{Y}_* - \overline{\boldsymbol{Y}}\boldsymbol{1}^T. \tag{C1}$$

We shall show that (C1) has a unique solution $\{\hat{\xi}_i(\boldsymbol{x}_j), i = 1,\dots,J, j = 1,\dots,n\}$ below.

Note that $\boldsymbol{Y}_* - \overline{\boldsymbol{Y}}\boldsymbol{1}^T$ has column rank $J \leq \min\{N,n\}$ since $\boldsymbol{S}$ has rank J. Thus, $\boldsymbol{Y}_* - \overline{\boldsymbol{Y}}\boldsymbol{1}^T = \boldsymbol{V}\boldsymbol{E}$, where $\boldsymbol{V}$ is a $N{\times}J$ matrix with J linearly independent columns of $\boldsymbol{Y}_* - \overline{\boldsymbol{Y}}\boldsymbol{1}^T$, $\boldsymbol{E}$ is a $J{\times}n$ matrix of rank J. This gives

$$\boldsymbol{S} = \frac{1}{n}\boldsymbol{V}\boldsymbol{E}\boldsymbol{E}^T\boldsymbol{V}^T. \tag{C2}$$

Note that $\boldsymbol{E}\boldsymbol{E}^T$ is a $J{\times}J$ matrix of rank J. Consequently, $\boldsymbol{E}\boldsymbol{E}^T\boldsymbol{V}^T$ has rank J. We also have

$$\boldsymbol{S} = \sum_{i=1}^{J} \widehat{\boldsymbol{\Phi}}_i \hat{\lambda}_i \widehat{\boldsymbol{\Phi}}_i^T = \left(\widehat{\boldsymbol{\Phi}}_1,\dots,\widehat{\boldsymbol{\Phi}}_J\right)\begin{pmatrix} \hat{\lambda}_1\widehat{\boldsymbol{\Phi}}_1^T \\ \vdots \\ \hat{\lambda}_J\widehat{\boldsymbol{\Phi}}_J^T \end{pmatrix}. \tag{C3}$$

Let $\boldsymbol{A}$ be a $J{\times}J$ matrix with J linearly independent columns of $\begin{pmatrix} \hat{\lambda}_1\widehat{\boldsymbol{\Phi}}_1^T \\ \vdots \\ \hat{\lambda}_J\widehat{\boldsymbol{\Phi}}_J^T \end{pmatrix}$. Let $\boldsymbol{B}$ be a $J{\times}J$ matrix of the corresponding columns (columns with same indices) of $\frac{1}{n}\boldsymbol{E}\boldsymbol{E}^T\boldsymbol{V}^T$. Then, by equation (C2) and (C3), we obtain

$$\boldsymbol{V}\boldsymbol{B} = \left(\widehat{\boldsymbol{\Phi}}_1, \dots, \widehat{\boldsymbol{\Phi}}_J\right)\boldsymbol{A}. \tag{C4}$$

Because $\left(\widehat{\boldsymbol{\Phi}}_1, \dots, \widehat{\boldsymbol{\Phi}}_J\right)\boldsymbol{A}$ has rank J, the columns of $\boldsymbol{B}$ are necessarily linearly independent. Since both $\boldsymbol{A}$ and $\boldsymbol{B}$ are invertible, we conclude that

$$\text{Column Span } (\boldsymbol{Y}_* - \overline{\boldsymbol{Y}}\boldsymbol{1}^T) = \text{Column Span } (\boldsymbol{V}) = \text{Column Span } \left(\widehat{\boldsymbol{\Phi}}_1, \dots, \widehat{\boldsymbol{\Phi}}_J\right). \tag{C5}$$

This implies that (C1) has a unique solution, which is easily seen to be given by

$$\begin{pmatrix} \hat{\xi}_1(\boldsymbol{x}_1) & \cdots & \hat{\xi}_1(\boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ \hat{\xi}_J(\boldsymbol{x}_1) & \cdots & \hat{\xi}_J(\boldsymbol{x}_n) \end{pmatrix} = \text{diag}\{\hat{\lambda}_1^{-1/2}, \dots, \hat{\lambda}_J^{-1/2}\} \begin{pmatrix} \widehat{\boldsymbol{\Phi}}_1^T \\ \vdots \\ \widehat{\boldsymbol{\Phi}}_J^T \end{pmatrix} (\boldsymbol{Y}_* - \overline{\boldsymbol{Y}}\boldsymbol{1}^T). \tag{C6}$$

Appendix D: Proof of Theorem 1

Denote $\Xi_i = \left(\xi_i(\boldsymbol{x}_1), \dots, \xi_i(\boldsymbol{x}_n)\right)^T$. By the model (12),

$$\boldsymbol{Y}_* - \overline{\boldsymbol{Y}}\boldsymbol{1}^T = \boldsymbol{Y}_* - \boldsymbol{Y}_*\boldsymbol{1}\boldsymbol{1}^T/n = \boldsymbol{\mu}_{\mathcal{M}}\boldsymbol{1}^T + \textstyle\sum_{i=1}^N \boldsymbol{\Phi}_i\sqrt{\lambda_i}\Xi_i^T - \left(\boldsymbol{\mu}_{\mathcal{M}}\boldsymbol{1}^T + \sum_{i=1}^N \boldsymbol{\Phi}_i\sqrt{\lambda_i}\Xi_i^T\right)\boldsymbol{1}\boldsymbol{1}^T/n$$

$$= \textstyle\sum_{i=1}^N \boldsymbol{\Phi}_i\sqrt{\lambda_i}\Xi_i^T(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n). \tag{D1}$$

Thus, we have

$$\boldsymbol{S} = n^{-1}\textstyle\sum_{i=1}^n (\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}})^T = n^{-1}(\boldsymbol{Y}_* - \overline{\boldsymbol{Y}}\boldsymbol{1}^T)(\boldsymbol{Y}_* - \overline{\boldsymbol{Y}}\boldsymbol{1}^T)^T$$

$$= n^{-1}\textstyle\sum_{i=1}^N \boldsymbol{\Phi}_i\sqrt{\lambda_i}\Xi_i^T(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n)\left[\sum_{j=1}^N \boldsymbol{\Phi}_j\sqrt{\lambda_j}\Xi_j^T(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n)\right]^T$$

$$= n^{-1}\textstyle\sum_{i=1}^N \boldsymbol{\Phi}_i\sqrt{\lambda_i}\Xi_i^T(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n)\sum_{j=1}^N \sqrt{\lambda_j}(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n)\Xi_j\boldsymbol{\Phi}_j^T$$

$$= n^{-1}\textstyle\sum_{i=1}^N\sum_{j=1}^N \sqrt{\lambda_i}\sqrt{\lambda_j}\boldsymbol{\Phi}_i\Xi_i^T(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n)^2\,\Xi_j\boldsymbol{\Phi}_j^T$$

$$= n^{-1}\textstyle\sum_{i=1}^N\sum_{j=1}^N \sqrt{\lambda_i}\sqrt{\lambda_j}\boldsymbol{\Phi}_i\Xi_i^T(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n)\,\Xi_j\boldsymbol{\Phi}_j^T. \tag{D2}$$

Define the $n\times n$ matrix $\boldsymbol{R}^i$ by $\left(\boldsymbol{R}^i\right)_{kl} = \mathcal{R}_i(\boldsymbol{x}_k, \boldsymbol{x}_l|\boldsymbol{\theta}_i)$. By independence of Ξ_i and Ξ_j for $i \neq j$, we have

$$E(\boldsymbol{S}) = n^{-1}\textstyle\sum_{i=1}^N \lambda_i\boldsymbol{\Phi}_i E[\Xi_i^T(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n)\Xi_i]\boldsymbol{\Phi}_i^T = n^{-1}\sum_{i=1}^N \lambda_i\boldsymbol{\Phi}_i \text{trace}\left[(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T/n)\boldsymbol{R}^i\right]\boldsymbol{\Phi}_i^T$$

$$= \textstyle\sum_{i=1}^N \lambda_i\boldsymbol{\Phi}_i\boldsymbol{\Phi}_i^T\,[1 - n^{-2}\sum_{l=1}^n\sum_{m=1}^n \mathcal{R}_i(\boldsymbol{x}_l, \boldsymbol{x}_m|\boldsymbol{\theta}_i)]. \tag{D3}$$

Define $a_{ij,kl} = \left(\sqrt{\lambda_i}\sqrt{\lambda_j}\boldsymbol{\Phi}_i\boldsymbol{\Phi}_j^T\right)_{kl}$. Then, by (D2), we have

$$\text{var}[(\boldsymbol{S})_{kl}] = n^{-2}\text{var}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij,kl}\Xi_i^T(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\Xi_j\right]$$

$$= n^{-2}\text{var}\left[\sum_{i=1}^{N} a_{ii,kl}\Xi_i^T(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\Xi_i + \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} a_{ij,kl}\Xi_i^T(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\Xi_j\right]$$

$$= n^{-2}\sum_{i=1}^{N} a_{ii,kl}^2\text{var}[\Xi_i^T(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\Xi_i]$$

$$+n^{-2}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\left(a_{ij,kl}^2 + a_{ij,kl}a_{ji,kl}\right)E\left[\Xi_i^T(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\Xi_j\Xi_j^T(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\Xi_i\right]$$

$$= n^{-2}\sum_{i=1}^{N} a_{ii,kl}^2\text{var}[\Xi_i^T(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\Xi_i]$$

$$+n^{-2}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\left(a_{ij,kl}^2 + a_{ij,kl}a_{ji,kl}\right)\text{var}\left[\Xi_i^T(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\Xi_j\right]. \tag{D4}$$

The third equality follows from the fact that $E\left[\Xi_i^T(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\Xi_j\Xi_k^T(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\Xi_l\right] = 0$ whenever there are three or four distinct indices among i, j, k, l.

It is well-known that (Page 109 of Rencher and Schaalje (2008))

$$\text{var}[\Xi_i^T(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\Xi_i] = 2\text{trace}\left[(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\boldsymbol{R}^i(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\boldsymbol{R}^i\right]. \tag{D5}$$

In addition, we have

$$\text{var}\left[\Xi_i^T(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\Xi_j\right] = \text{var}\left[\left(\Xi_i^T,\Xi_j^T\right)\begin{pmatrix}\mathbf{0} & (\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)/2\\(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)/2 & \mathbf{0}\end{pmatrix}\begin{pmatrix}\Xi_i\\\Xi_j\end{pmatrix}\right]$$

$$= 2\text{trace}\left\{\left[\begin{pmatrix}\mathbf{0} & (\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)/2\\(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)/2 & \mathbf{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{R}^i & \mathbf{0}\\\mathbf{0} & \boldsymbol{R}^j\end{pmatrix}\right]^2\right\}$$

$$= \text{trace}\left[(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\boldsymbol{R}^i(\boldsymbol{I}-\mathbf{1}\mathbf{1}^T/n)\boldsymbol{R}^j\right]. \tag{D6}$$

References

Rencher, A. C. and Schaalje, G. B. (2008). *Linear models in statistics (2nd Edition)*. New York: John Wiley & Sons.

Appendix E: Asymptotic rates of convergence of $\boldsymbol{S}$ and $\overline{\boldsymbol{Y}}$

In this appendix, we establish the asymptotic rate of convergence of $\boldsymbol{S}$ and $\overline{\boldsymbol{Y}}$ under increasing domain asymptotics, as given by Theorem 2 and Theorem 3 below.

Theorem 2: For fixed $\mathcal{M}$, each component of $\boldsymbol{S}$ converges to the corresponding component of $\boldsymbol{\Sigma} = \sum_{i=1}^{N}\lambda_i\boldsymbol{\Phi}_i\boldsymbol{\Phi}_i^T$ in mean square at a rate of $O(n^{-1})$.

Theorem 3: For fixed $\mathcal{M}$, $\widehat{\boldsymbol{\mu}}_{\mathcal{M}} = \overline{\boldsymbol{Y}}$ is an unbiased estimator of $\boldsymbol{\mu}_{\mathcal{M}}$ that converges to $\boldsymbol{\mu}_{\mathcal{M}}$ in mean square at a rate of $O(n^{-1})$.

First, we need two assumptions.

Assumption 1: For all i, $0 \le \mathcal{R}_i(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\theta}_i) = O(\|\boldsymbol{x} - \boldsymbol{x}'\|^{-\gamma}), \gamma > d$ as $\|\boldsymbol{x} - \boldsymbol{x}'\| \to \infty$.

Remark: This assumption holds for the product Matèrn correlation in (6) and $\|\cdot\|$ can refer to any norm defined on $\mathbb{R}^d$.

Assumption 2: Let $\mathcal{D}_\infty = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots\} \subset \mathbb{R}^d$ be the sequence of design points. For each compact ball $\mathcal{C}$ in $\mathbb{R}^d$, i.e., $\mathcal{C} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}_\mathcal{C}\| \le r_\mathcal{C}\}$, where $\boldsymbol{x}_\mathcal{C}$ is a fixed point in $\mathbb{R}^d$, the set $\mathcal{C} \cap \mathcal{D}_\infty$ is finite and $|\mathcal{C} \cap \mathcal{D}_\infty| \le \mathcal{W} r_\mathcal{C}^d$ for fixed $\mathcal{W} > 0$, where $|\mathcal{C} \cap \mathcal{D}_\infty|$ is the cardinality of $\mathcal{C} \cap \mathcal{D}_\infty$.

Remark: This assumption holds if the design points are constrained to be on a grid, such as $\{\dots, -0.02, -0.01, 0, 0.01, 0.02, \dots\}^d$. Note that other authors such as Mardia and Marshall (1984) have also used increasing domain asymptotics to study GP model parameter estimates.

In order to prove Theorem 2 and Theorem 3, we need the following lemma:

Lemma 1:

1. $n^{-2} \sum_{l=1}^n \sum_{m=1}^n \mathcal{R}_i(\boldsymbol{x}_l, \boldsymbol{x}_m|\boldsymbol{\theta}_i) = O(n^{-1})$.
2. $n^{-2} \sum_{l=1}^n \sum_{m=1}^n \left[\mathcal{R}_i(\boldsymbol{x}_l, \boldsymbol{x}_m|\boldsymbol{\theta}_i)\mathcal{R}_j\left(\boldsymbol{x}_l, \boldsymbol{x}_m|\boldsymbol{\theta}_j\right)\right] = O(n^{-1})$
3. $n^{-2} \sum_{p=1}^n \sum_{m=1}^n \left[n^{-2} \sum_{l=1}^n \mathcal{R}_i(\boldsymbol{x}_l, \boldsymbol{x}_m|\boldsymbol{\theta}_i) \sum_{l=1}^n \mathcal{R}_j\left(\boldsymbol{x}_l, \boldsymbol{x}_p|\boldsymbol{\theta}_j\right)\right] = O(n^{-2})$
4. $n^{-2} \sum_{m=1}^n n^{-1} [\sum_{l=1}^n \mathcal{R}_i(\boldsymbol{x}_m, \boldsymbol{x}_l|\boldsymbol{\theta}_i)]\left[\sum_{l=1}^n \mathcal{R}_j\left(\boldsymbol{x}_m, \boldsymbol{x}_l|\boldsymbol{\theta}_j\right)\right] = O(n^{-2})$

Proof of 1:

By Assumption 1, there exists τ_0 such that $\mathcal{R}_i(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\theta}_i) \le C/\tau^\gamma$ for some $C > 0$ whenever $\|\boldsymbol{x} - \boldsymbol{x}'\| = \tau \ge \tau_0$. Suppose that $n = \mathcal{W}\tau_\infty^d, \tau_\infty > \tau_0$, and $\tau_1, \dots, \tau_{q-1}$ are numbers such that $\mathcal{W}\tau_1^d = n_1 = \lceil \mathcal{W}\tau_0^d \rceil, \mathcal{W}\tau_2^d = n_2 = n_1 + 1, \dots, \mathcal{W}\tau_q^d = n_q = n$, where $\tau_q = \tau_\infty$. Then, for fixed $l \in \{1, \dots, n\}$, we have

$n^{-1} \sum_{m=1}^n \mathcal{R}_i(\boldsymbol{x}_l, \boldsymbol{x}_m|\boldsymbol{\theta}_i) \le n^{-1}\left[\mathcal{W}\tau_0^d(1) + \left(\mathcal{W}\tau_1^d - \lfloor \mathcal{W}\tau_0^d \rfloor\right)C/\tau_1^\gamma + \dots + \left(\mathcal{W}\tau_q^d - \mathcal{W}\tau_{q-1}^d\right)C/\tau_q^\gamma\right] = O\left\{n^{-1}\left[\mathcal{W}\tau_0^d + \sum_{i=1}^q C/\tau_i^\gamma\right]\right\} = O\left\{n^{-1}\left[\mathcal{W}\tau_0^d + \sum_{i=1}^q \mathcal{W}^{\gamma/d} C/n_i^{\gamma/d}\right]\right\} =$

$O\{n^{-1}[\mathcal{W}\tau_0^d + \sum_{i=1}^{\infty} \mathcal{W}^{\gamma/d} C/i^{\gamma/d}]\} = O(n^{-1}) \forall l \in \{1, \dots, n\}$.

Proof of 2:

By Assumption 1, there exists τ_0 such that $\mathcal{R}_i(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\theta}_i) \leq C/\tau^\gamma$ and $\mathcal{R}_j(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\theta}_j) \leq C/\tau^\gamma$ for some $C > 0$ whenever $\|\boldsymbol{x} - \boldsymbol{x}'\| = \tau \geq \tau_0$. Suppose that $n = \mathcal{W}\tau_\infty^d, \tau_\infty > \tau_0$, and $\tau_1, \dots, \tau_{q-1}$ are numbers such that $\mathcal{W}\tau_1^d = n_1 = \lceil \mathcal{W}\tau_0^d \rceil, \mathcal{W}\tau_2^d = n_2 = n_1 + 1, \dots, \mathcal{W}\tau_q^d = n_q = n$, where $\tau_q = \tau_\infty$. Then, for fixed $l \in \{1, \dots, n\}$,

$n^{-1} \sum_{m=1}^{n} [\mathcal{R}_i(\boldsymbol{x}_l, \boldsymbol{x}_m|\boldsymbol{\theta}_i)\mathcal{R}_j(\boldsymbol{x}_l, \boldsymbol{x}_m|\boldsymbol{\theta}_j)] \leq n^{-1} \left[\mathcal{W}\tau_0^d(1)^2 + (\mathcal{W}\tau_1^d - \lceil \mathcal{W}\tau_0^d \rceil)(C/\tau_1^\gamma)^2 + \right.$

$\left. \dots + (\mathcal{W}\tau_q^d - \mathcal{W}\tau_{q-1}^d)(C/\tau_q^\gamma)^2\right] = O\{n^{-1}[\mathcal{W}\tau_0^d + \sum_{i=1}^{q} C^2/\tau_i^{2\gamma}]\} = O\{n^{-1}[\mathcal{W}\tau_0^d +$

$\sum_{i=1}^{q} \mathcal{W}^{2\gamma/d} C^2/n_i^{2\gamma/d}]\} = O\{n^{-1}[\mathcal{W}\tau_0^d + \sum_{i=1}^{\infty} \mathcal{W}^{2\gamma/d} C^2/i^{2\gamma/d}]\} = O(n^{-1}) \forall l \in \{1, \dots, n\}$.

Proof of 3 and 4:

This follows from the fact that $n^{-1} \sum_{m=1}^{n} \mathcal{R}_i(\boldsymbol{x}_l, \boldsymbol{x}_m|\boldsymbol{\theta}_i) = O(n^{-1})$.

Proof of Theorem 2

By Lemma 1 and (D3)-(D6), we obtain

$E\left\{\left[(\boldsymbol{S})_{kl} - \sum_{i=1}^{N} a_{ii,kl}\right]^2\right\} = \left\{E[(\boldsymbol{S})_{kl}] - \sum_{i=1}^{N} a_{ii,kl}\right\}^2 + \text{var}[(\boldsymbol{S})_{kl}]$

$= \left\{\sum_{i=1}^{N} a_{ii,kl} [1 - O(n^{-1})] - \sum_{i=1}^{N} a_{ii,kl}\right\}^2 + \sum_{i=1}^{N} a_{ii,kl}^2 O(n^{-1}) + \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (a_{ij,kl}^2 +$

$a_{ij,kl}a_{ji,kl})O(n^{-1}) = O(n^{-1})$. (E1)

Proof of Theorem 3

Since $\overline{\boldsymbol{Y}} = (\boldsymbol{\mu}_{\mathcal{M}}\mathbf{1}^T + \sum_{i=1}^{N} \boldsymbol{\Phi}_i\sqrt{\lambda_i}\boldsymbol{\Xi}_i^T)\mathbf{1}/n$, we immediately see that $E(\overline{\boldsymbol{Y}}) = \boldsymbol{\mu}_{\mathcal{M}}$.

Moreover, due to independence of $\boldsymbol{\Xi}_i$ and $\boldsymbol{\Xi}_j$ for $i \neq j$, we have

$\text{cov}(\overline{\boldsymbol{Y}}) = \sum_{i=1}^{N} \lambda_i \boldsymbol{\Phi}_i \boldsymbol{\Phi}_i^T n^{-2} \text{var}(\boldsymbol{\Xi}_i^T \mathbf{1}) = \sum_{i=1}^{N} \lambda_i \boldsymbol{\Phi}_i \boldsymbol{\Phi}_i^T [n^{-2} \sum_{l=1}^{n} \sum_{m=1}^{n} \mathcal{R}_i(\boldsymbol{x}_l, \boldsymbol{x}_m|\boldsymbol{\theta}_i)] =$

$\sum_{i=1}^{N} \lambda_i \boldsymbol{\Phi}_i \boldsymbol{\Phi}_i^T O(n^{-1})$. (E2)

References

Mardia, K. V. and Marshall, R. J. (1984). "Maximum likelihood estimation of models for residual covariance in spatial regression," *Biometrika*, 71(1), 135-146.

Appendix F: Example with Dirichlet boundary and initial conditions

We consider the PDE system given by (1)-(4). We want to model the temperature of a plate with plane geometry $[-L, L]^2 \backslash [-0.25L, 0.25L]^2$ and uniform thermal conductivity K. To illustrate how a single spatial mesh can be used when the spatial domain changes size, we transform the PDE so that $(s_1, s_2) \in \mathcal{S} = [-1,1]^2 \backslash [-0.25,0.25]^2$. The transformed PDE is obtained by replacing $\nabla \cdot \big(K \nabla T(s_1, s_2, s_3)\big) H$ in (1) with $\nabla \cdot \big(K \nabla T(s_1, s_2, s_3)\big) H L^{-2}$. We let $\partial \mathcal{S}_1 = \partial \mathcal{S}$, and $\partial \mathcal{S}_2 = \{\}$. The initial condition is $T_0(s_1, s_2) = x_1 + x_2[(\|(s_1, s_2)\|_\infty - 0.25)/0.75]^2$, and the Dirichlet boundary condition is $G_1(s_1, s_2) = T_0(s_1, s_2)$. The other inputs are $x_3 = L$, $x_4 = H$, $x_5 = h_c$, $x_6 = T_a$, $x_7 = K$, $x_8 = \rho C_p$. Ranges of the inputs are given in Table F1. The PDE is solved using the Matlab PDE toolbox. The mesh used is shown in Figure F1a, and the output values on the six-point time grid $\{0,18, \dots, 90\}$ are observed. Note that the time grid is not the actual time grid used by Matlab to solve the PDE, which is unknown. It is simply the time points at which an output is requested from Matlab. For illustration, we plot the initial condition and solution at time 90 for two different x_2 values with the other inputs held fixed. For $x_2 = -100$, we obtain the initial condition shown in Figure F1b and solution at time 90 shown in Figure F1c. For $x_2 = 0$, we obtain the initial condition shown in Figure F1d and solution at time 90 shown in Figure F1e. It is seen that the functional output can have very different shapes over the experiment region.

Table F1: Ranges of the Inputs (SI units are used for all inputs)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Minimum	500	-100	0.2	0.0025	30	270	170	$2.3\text{x}10^6$
Maximum	600	0	0.3	0.005	60	300	230	$2.5\text{x}10^6$

We generate 30 maximin Latin hypercube designs (LHDs) of size 56. For each design, we fit the covariance separable GP, modified KL-GP, and KL-GP models. The modified KL-GP model in Section 4.1 is used to take into account the Dirichlet boundary and initial conditions. The starting value for optimizing h in (32) is 1. All GP's use the product

Matèrn correlation function (6). Each scale parameter in (6) is restricted to be less than or equal to 16.4888. The performance of each of the three alternative methods in predicting the functional output is evaluated on a test set of 50 values of $\boldsymbol{x}$ given by a maximin LHD. The performance criteria are the mean absolute error (MAE), mean squared error (MSE),

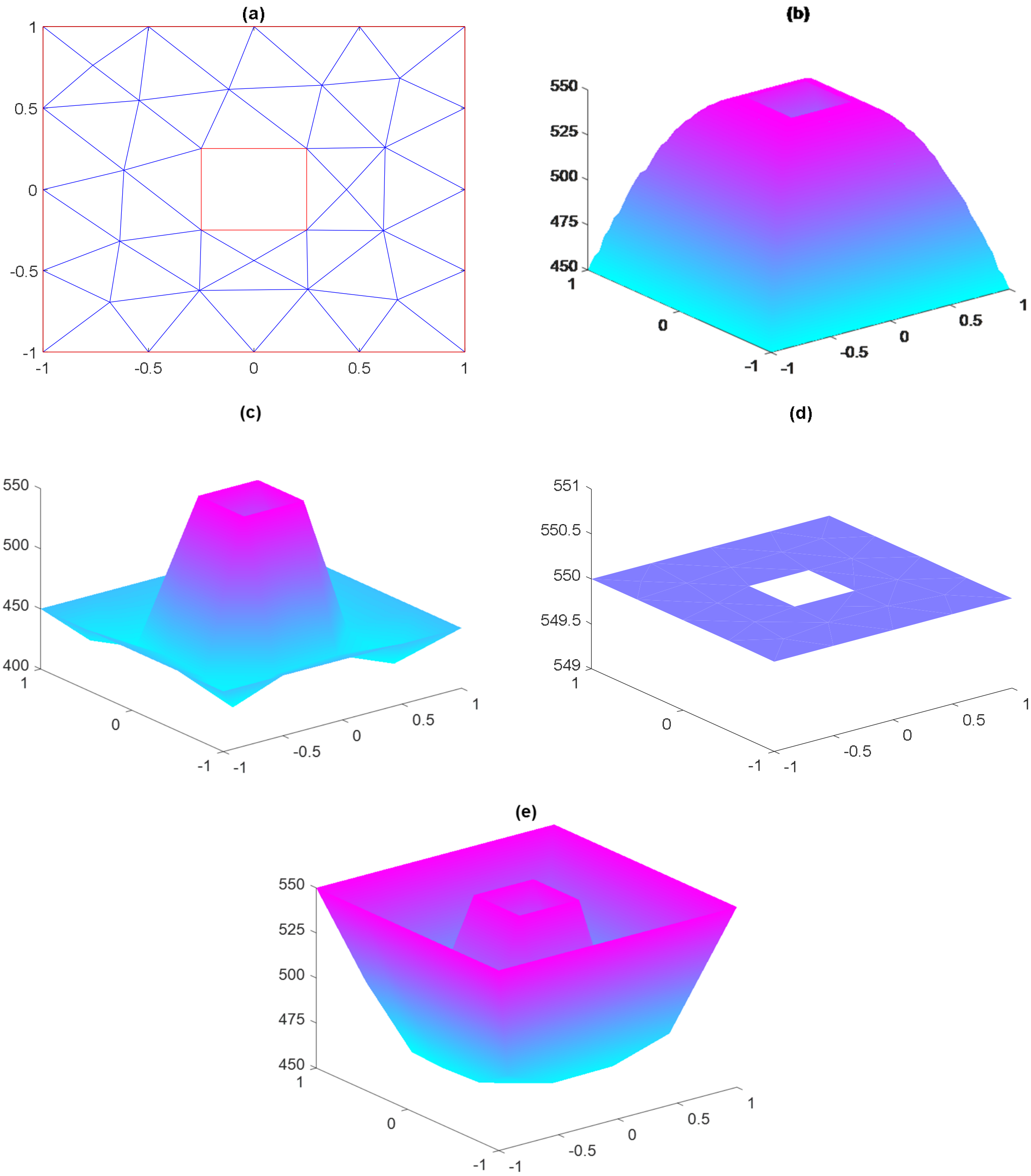

Figure F1: (a) Mesh; (b) Initial condition given by $x_2 = -100$ ($x_1 = 550, x_3 = 0.3, x_4 = 0.005, x_5 = 30, x_6 = 270, x_7 = 170, x_8 = 2.3\times10^6$); (c) Numerical solution at time 90 with initial condition (b); (d) Initial condition given by $x_2 = 0$ ($x_1 = 550, x_3 = 0.3, x_4 = 0.005, x_5 = 30, x_6 = 270, x_7 = 170, x_8 = 2.3\times10^6$); (e) Numerical solution at time 90 with initial condition (d).

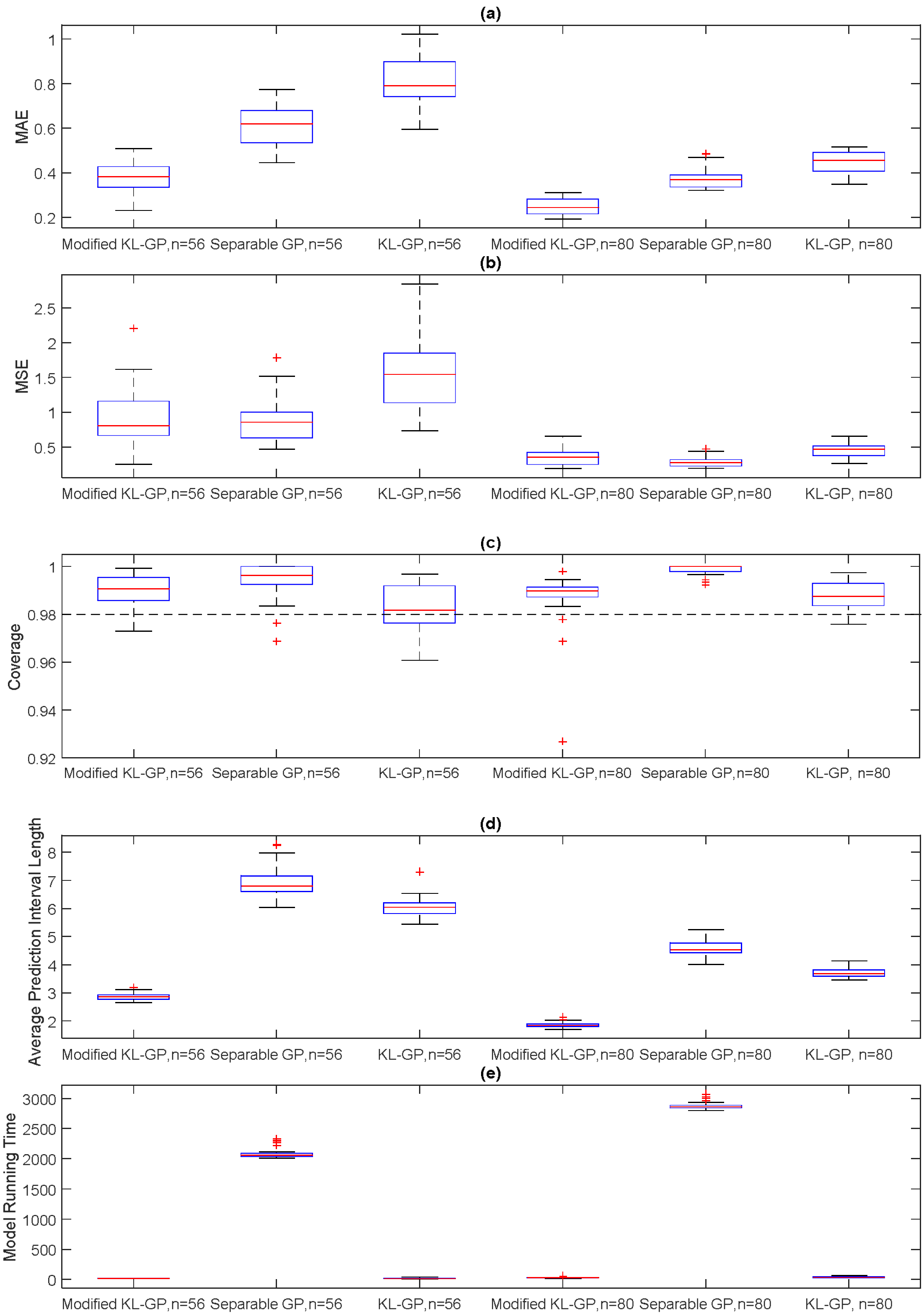

Figure F2: (a) Boxplots of MAE; (b) Boxplots of MSE; (c) Boxplots of coverage; (d) Boxplots of average prediction interval length; (e) Boxplots of model running time. The six boxplots in each figure are for the modified KL-GP, covariance separable GP, and KL-GP models fitted with maximin LHDs of size 56 and 80.

coverage of 98% prediction intervals, average prediction interval length, and model running time, which includes the time to fit the model and the time to compute predictions on the test set on a Lenovo laptop with 4GB RAM and 2.9GHz processor. The above simulation is repeated with 30 maximin LHDs of size 80 as experimental designs.

Results for these simulations are presented in Figures F2a-F2e. We see that the modified KL-GP model performs best with respect to the MAE criterion. However, the modified KL-GP and covariance separable GP models have similar performance with respect to the MSE criterion, with the former having more variation in MSE. It is seen that the KL-GP model has worst performance with respect to both MAE and MSE. The separable GP model and modified KL-GP model tend to have coverage higher than the nominal 0.98, and the KL-GP model tends to have coverage closest to nominal. We see that the modified KL-GP model clearly gives prediction intervals of far shorter length than the separable GP and KL-GP models. The average interval lengths of the latter two models tend to be more than two times the average interval length given by the modified KL-GP model. Despite these shorter intervals given by the modified KL-GP model, its coverage is comparable to the coverage achieved by the covariance separable GP and KL-GP models. Lastly, we see from Figure 2e that the modified KL-GP and KL-GP models take negligible time compared to the covariance separable GP model. Note that the size N of the functional output is 210, which is small for FE simulations since we use a coarse mesh and a coarse time grid. However, the covariance separable GP model requires 2000 to 3000 seconds, which is quite expensive.