Online Supplement for "Sparse Pseudo-input Local Kriging for Large Spatial Datasets with Exogenous Variables" by Babak Farmanesh and Arash Pourhabib

Appendix A Solving optimization problem (11)

Due to the convex objective function and affine constraints of optimization problem (11), the duality gap between the primal and dual problems of (11) is zero by Lagrange duality principle (Bazaraa et al., 2013). This allows us to transform the optimization problem (11) to an unconstrained optimization problem and maximize the Lagrangian of (11) instead,

$$\begin{aligned}\max_{\mathbf{u}_s(\mathbf{x}_*),\boldsymbol{\lambda}_s(\mathbf{x}_*)} \mathcal{L}(\mathbf{u}_s(\mathbf{x}_*),\boldsymbol{\lambda}_s(\mathbf{x}_*)) = \mathbf{u}_s(\mathbf{x}_*)^T(\tilde{\mathbf{K}}^s_{\mathbf{X}_s\mathbf{X}_s} + \text{diag}(\mathbf{K}_{\mathbf{X}_s\mathbf{X}_s} - \tilde{\mathbf{K}}^s_{\mathbf{X}_s\mathbf{X}_s}) + \sigma_s^2\mathbf{I}_s)\mathbf{u}_s(\mathbf{x}_*) \\ -2\mathbf{u}_s(\mathbf{x}_*)^T\tilde{\mathbf{k}}^s_{\mathbf{X}_s\mathbf{x}_*} - \sum_{i=1:|\mathbf{B}_s|} \lambda_{is}(\mathbf{x}_*)(\mathbf{u}_s(\mathbf{b}_i)^T\mathbf{y}_s - \mathcal{R}(\mathbf{b}_i)),\end{aligned} \tag{23}$$

where $|\mathbf{B}_s|$ is the number of all the control points located on the boundaries of subdomain Ω_s, and $\boldsymbol{\lambda}_s(\mathbf{x}_*) = [\lambda_{1s}(\mathbf{x}_*), \ldots, \lambda_{|\mathbf{B}_s|s}(\mathbf{x}_*)]^T$ is the vector of the Lagrange multipliers.

Assuming $\mathbf{u}_s(\mathbf{x}_*)$ depends on the covariance between $\mathbf{x}_*$ and $\mathbf{X}_s$, and $\lambda_{is}(\mathbf{x}_*)$ depends on the covariance of $\mathbf{b}_i$ and $\mathbf{x}_*$, we write $\mathbf{u}_s(\mathbf{x}_*) = \mathbf{H}_s\tilde{\mathbf{k}}^s_{\mathbf{X}_s\mathbf{x}_*}$ and $\lambda_{is}(\mathbf{x}^*) = \beta_{is}\tilde{k}^s_{\mathbf{b}_i\mathbf{x}_*}$ as suggested in (Park et al., 2011), where $\mathbf{H}_j$ is a squared matrix with size equal to the number of data points in Ω_s, and β_{is} is the Lagrange parameter associated with λ_{is} that does not depend on $\mathbf{x}_*$. Consequently, we rewrite Lagrangian (23) as

$$\begin{aligned}\max_{\mathbf{H}_s,\boldsymbol{\beta}_s} \mathcal{L}(\mathbf{H}_s,\boldsymbol{\beta}_s) = \tilde{\mathbf{k}}^s_{\mathbf{x}_*\mathbf{X}_s}\mathbf{H}_s^T(\tilde{\mathbf{K}}^s_{\mathbf{X}_s\mathbf{X}_s} + \text{diag}(\mathbf{K}_{\mathbf{X}_s\mathbf{X}_s} - \tilde{\mathbf{K}}^s_{\mathbf{X}_s\mathbf{X}_s}) + \sigma_s^2\mathbf{I}_s)\mathbf{H}_s\tilde{\mathbf{k}}^s_{\mathbf{X}_s\mathbf{x}_*} \\ -2\tilde{\mathbf{k}}^s_{\mathbf{x}_*\mathbf{X}_s}\mathbf{H}_s^T\tilde{\mathbf{k}}^s_{\mathbf{X}_s\mathbf{x}_*} - \tilde{\mathbf{k}}^s_{\mathbf{x}_*\mathbf{B}_s}\boldsymbol{\beta}_s(\tilde{\mathbf{K}}^s_{\mathbf{B}_s\mathbf{X}_s}\mathbf{H}_s^T\mathbf{y}_s - \mathbf{r}_s),\end{aligned} \tag{24}$$

where $\boldsymbol{\beta}_s$ is a diagonal matrix with diagonal elements $\beta_{1s}, \ldots, \beta_{|\mathbf{B}_s|s}$, and $\mathbf{r}_s = [\mathcal{R}(\mathbf{b}_1), \ldots, \mathcal{R}(\mathbf{b}_{|\mathbf{B}_s|})]^T$ is the vectors of boundary values of Ω_s.

Due to convexity of function (24) we can calculate the optimal values of $\mathbf{H}_s$ and $\boldsymbol{\beta}_s$ analytically

by writing out the first order necessary conditions,

$$\frac{d\mathcal{L}(\mathbf{H}_s, \boldsymbol{\beta}_s)}{d\mathbf{H}_s} = 2(\mathbf{G}_s\mathbf{H}_s - \mathbf{I}_s)\tilde{\mathbf{k}}^s_{\mathbf{X}_s\mathbf{x}_*}\tilde{\mathbf{k}}^s_{\mathbf{x}_*\mathbf{X}_s} - \mathbf{y}_s\tilde{\mathbf{k}}^s_{\mathbf{x}_*\mathbf{B}_s}\boldsymbol{\beta}_s\tilde{\mathbf{K}}^s_{\mathbf{B}_s\mathbf{X}_s} = 0, \tag{25}$$

$$\frac{d\mathcal{L}(\mathbf{H}_s, \boldsymbol{\beta}_s)}{d\beta_{is}} = \tilde{\mathbf{k}}_{\mathbf{b}_i\mathbf{X}_s}\mathbf{H}_s^T\mathbf{y}_j - r_{is} = 0 \quad \forall i \in [|\mathbf{B}_s|], \tag{26}$$

where $\mathbf{G}_s = (\tilde{\mathbf{K}}^s_{\mathbf{X}_s\mathbf{X}_s} + \text{diag}(\mathbf{K}_{\mathbf{X}_s\mathbf{X}_s} - \tilde{\mathbf{K}}^s_{\mathbf{X}_s\mathbf{X}_s}) + \sigma_s^2\mathbf{I}_s)$, and r_{is} is the i^{th} element of the vector $\mathbf{r}_s$. Reordering equation (25),

$$(\tilde{\mathbf{k}}^s_{\mathbf{x}_*\mathbf{X}_s} + 0.5(\tilde{\mathbf{k}}^j_{\mathbf{x}_*\mathbf{X}_s}\tilde{\mathbf{k}}^j_{\mathbf{X}_s\mathbf{x}_*})^{-1}\tilde{\mathbf{k}}^j_{\mathbf{x}_*\mathbf{X}_s}\tilde{\mathbf{K}}^j_{\mathbf{X}_s\mathbf{B}_s}\boldsymbol{\beta}_s\tilde{\mathbf{k}}^j_{\mathbf{B}_s\mathbf{x}_*}\mathbf{y}_s^T)\mathbf{G}_s^{-1}\mathbf{y}_s = \tilde{\mathbf{k}}^s_{\mathbf{x}_*\mathbf{X}_s}\mathbf{H}_s^T\mathbf{y}_s, \tag{27}$$

and evaluating it at the boundary locations gives the system of equations with $|\mathbf{B}_s|$ equations and Lagrangian parameters,

$$(\tilde{\mathbf{k}}^s_{\mathbf{b}_i\mathbf{X}_s} + 0.5(\tilde{\mathbf{k}}^s_{\mathbf{b}_i\mathbf{X}_s}\tilde{\mathbf{k}}^s_{\mathbf{X}_s\mathbf{b}_i})^{-1}\tilde{\mathbf{k}}^s_{\mathbf{b}_i\mathbf{X}_s}\tilde{\mathbf{K}}^s_{\mathbf{X}_s\mathbf{B}_s}\boldsymbol{\beta}_s\tilde{\mathbf{k}}^s_{\mathbf{B}_s\mathbf{b}_i}\mathbf{y}_s^T)\mathbf{G}_s^{-1}\mathbf{y}_s = r_{is} \ \ \forall i \in [|\mathbf{B}_s|]. \tag{28}$$

After some simple matrix algebra, we obtain the solution to the system of linear equations (28),

$$\boldsymbol{\beta}_s = \frac{\mathbf{I}_s(\mathbf{r}_s - \tilde{\mathbf{K}}^s_{\mathbf{B}_s\mathbf{X}_s}\mathbf{G}_s^{-1}\mathbf{y}_s)\{[(\text{diag}(\tilde{\mathbf{K}}^s_{\mathbf{B}_s\mathbf{X}_s}\tilde{\mathbf{K}}^s_{\mathbf{X}_s\mathbf{B}_s}))^{-1}(\tilde{\mathbf{K}}^s_{\mathbf{B}_s\mathbf{X}_s}\tilde{\mathbf{K}}^s_{\mathbf{X}_s\mathbf{B}_s})] \circ \mathbf{K}^s_{\mathbf{B}_s\mathbf{B}_s}\}^{-1}}{0.5\mathbf{y}_s^T\mathbf{G}_s^{-1}\mathbf{y}_s}. \tag{29}$$

Using the values of $\boldsymbol{\beta}_s$ from (29), we can easily obtain the solution to $\mathbf{u}(\mathbf{x}_*)$ from (25),

$$\mathbf{u}_s^*(\mathbf{x}_*) = \mathbf{H}_s\tilde{\mathbf{k}}^s_{\mathbf{X}_s\mathbf{x}_*} = \mathbf{G}_s^{-1}(\tilde{\mathbf{k}}^s_{\mathbf{X}_s\mathbf{x}_*} + \mathbf{w}_s), \tag{30}$$

where $\mathbf{w}_s = 0.5(\tilde{\mathbf{k}}^s_{\mathbf{x}_*\mathbf{X}_s}\tilde{\mathbf{k}}^s_{\mathbf{X}_s\mathbf{x}_*})^{-1}\mathbf{y}_s\tilde{\mathbf{k}}^s_{\mathbf{x}_*\mathbf{B}_s}\boldsymbol{\beta}_s\tilde{\mathbf{K}}^s_{\mathbf{B}_s\mathbf{X}_s}\tilde{\mathbf{k}}^s_{\mathbf{X}_s\mathbf{x}_*}$.

Appendix B Derivation of low-rank covariance approximation error

We follow the procedure proposed in (Smola and Schölkopf, 2000) to derive the low-rank covariance approximation error in each subdomain Ω_s. In this derivation, given the covariance function $\phi(\cdot,\cdot) : \Omega_s \times \Omega_s \to \mathbb{R}$ as a symmetric positive semidefinite kernel, we intend to approximate the kernel

$\phi(\mathbf{x}, \cdot) : \Omega_s \to \mathbb{R}^{\Omega_s}$ centered at $\mathbf{z} \in \Omega_s$ as a linear combination of kernels centered at each element of $\mathbf{X}_s$, i.e.,

$$\phi(\mathbf{z}, \cdot) \approx \sum_{i \in [m_s]} c_i \phi(\tilde{\mathbf{x}}_i, \cdot). \tag{31}$$

To this end, let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) that is defined as the space of functions constructed by the span of $\phi(\mathbf{x}, \cdot)$ centered at a finite number of elements of Ω_s, i.e.,

$$\left\{ \sum_{i \in [n]} \alpha_i \phi(\mathbf{x}_i, \cdot) : n \in \mathbb{N}, \mathbf{x}_i \in \Omega_s, c_i \in \mathbf{R} \right\}.$$

$\mathcal{H}$ is also equipped with the inner product

$$\left\langle \sum_{i \in [n_1]} \alpha_i \phi(\mathbf{x}_i, \cdot), \sum_{j \in [n_2]} \beta_j \phi(\mathbf{x}_j, \cdot) \right\rangle_{\mathcal{H}} = \sum_{i \in [n_1]} \sum_{j \in [n_2]} \alpha_i \beta_j \phi(\mathbf{x}_i, \mathbf{x}_j), \tag{32}$$

which, for any function $f \in \mathcal{H}$, induces the norm

$$||f||^2_{\mathcal{H}} =< f, f >_{\mathcal{H}} . \tag{33}$$

Given such $\mathcal{H}$, a natural criterion to find an approximation for the covariance function is to minimize the norm of function $\phi(\mathbf{z}, \cdot) - \sum_{i \in [m_s]} c_i \phi(\tilde{\mathbf{x}}_i, \cdot)$, which belongs to $\mathcal{H}$, that is

$$\min_{\mathbf{c}} \left\| \phi(\mathbf{z}, \cdot) - \sum_{i \in [m_s]} c_i \phi(\tilde{\mathbf{x}}_i, \cdot) \right\|^2_{\mathcal{H}}, \tag{34}$$

where $\mathbf{c} = [c_1, \ldots, c_{m_s}]^T$. Assuming $\phi(\mathbf{z}, \mathbf{z}) = h$, objective function (34) can be expanded after plugging in (32) and (33) as

$$\min_{\mathbf{c}} \; h - 2\mathbf{c}^T \mathbf{k}_{\tilde{\mathbf{X}}_s \mathbf{z}} + \mathbf{c}^T \mathbf{K}_{\tilde{\mathbf{X}}_s \tilde{\mathbf{X}}_s} \mathbf{c},$$

which has the solution $\mathbf{c}^* = \mathbf{K}^{-1}_{\tilde{\mathbf{X}}_s \tilde{\mathbf{X}}_s} \mathbf{k}_{\tilde{\mathbf{X}}_s \mathbf{z}}$. Therfore, the approximation of $\phi(\mathbf{z}, \cdot)$ becomes $\mathbf{k}_{\mathbf{z} \tilde{\mathbf{X}}_s} \mathbf{K}^{-1}_{\tilde{\mathbf{X}}_s \tilde{\mathbf{X}}_s} \mathbf{k}_{\tilde{\mathbf{X}}_s \mathbf{z}}$, and the error of covariance approximation becomes

$$h - \mathbf{k}_{\mathbf{z} \tilde{\mathbf{X}}_s} \mathbf{K}^{-1}_{\tilde{\mathbf{X}}_s \tilde{\mathbf{X}}_s} \mathbf{k}_{\tilde{\mathbf{X}}_s \mathbf{z}}.$$

We finally note that using $\mathbf{z} = \mathbf{x}_i$ for all $\mathbf{x}_i \in \mathbf{X}_s$ in objective function (34) and minimizing the sum over all terms obtains $\mathbf{K}_{\mathbf{X}_s\tilde{\mathbf{X}}_s}\mathbf{K}^{-1}_{\tilde{\mathbf{X}}_s\tilde{\mathbf{X}}_s}\mathbf{K}_{\tilde{\mathbf{X}}_s\mathbf{X}_s}$, which is the low-rank approximation of $\mathbf{K}_{\mathbf{X}_s\mathbf{X}_s}$ in equation (7).

Appendix C Proof of Theorems

C.1 Proof of Proposition 1

Proof. For any $i \in [m_s]$, let $\mathbf{u}_i$ denote the covariance vector between $\mathbf{z}$ and the first i elements of $\tilde{\mathbf{X}}_s$, and let $\mathbf{v}_i$ denote the covariance vector between the $(i+1)^{\text{th}}$ element of $\tilde{\mathbf{X}}_s$ and the first i elements of $\tilde{\mathbf{X}}_s$. That is, $\mathbf{u}_i = [\phi(\mathbf{z}, \tilde{\mathbf{x}_1}), \ldots, \phi(\mathbf{z}, \tilde{\mathbf{x}}_i)]^T$, and $\mathbf{v}_i = [\phi(\tilde{\mathbf{x}}_{i+1}, \tilde{\mathbf{x}}_1), \ldots, \phi(\tilde{\mathbf{x}}_{i+1}, \tilde{\mathbf{x}}_i)]^T$. Also let $\mathbf{K}_i$ denote the covariance matrix between the first i elements of $\tilde{\mathbf{X}}_s$ themselves. We now prove by induction on i. For the base case, i.e, $i = 1$, the claim clearly holds,

$$\mathbb{E}_{\Omega_s}(\mathbf{u}_1^T\mathbf{K}_1^{-1}\mathbf{u}_1) = \mathbb{E}_{\Omega_s}(\phi(\mathbf{z}, \tilde{\mathbf{x}}_1)\phi(\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_1)^{-1}\phi(\mathbf{z}, \tilde{\mathbf{x}}_1)) = \frac{1}{h}\mathbb{E}_{\Omega_s}(\phi^2(\mathbf{z}, \tilde{\mathbf{x}}_1)) = \frac{1}{h}\mathbb{E}_{\Omega_s}(\phi^2(\mathbf{x}, \mathbf{x}')). \quad (35)$$

Suppose the claim holds for $m_s - 1$, we show that it also holds for m_s. Expanding $\mathbf{u}_{m_s}^T\mathbf{K}_{m_s}^{-1}\mathbf{u}_{m_s}$ gives

$$\mathbf{u}_{m_s}^T\mathbf{K}_{m_s}^{-1}\mathbf{u}_{m_s} = \begin{bmatrix}\mathbf{u}_{m_s-1}^T & \phi(\mathbf{z}, \tilde{\mathbf{x}}_{m_s})\end{bmatrix}\begin{bmatrix}\mathbf{K}_{m_s-1} & \mathbf{v}_{m_s-1} \\ \mathbf{v}_{m_s-1}^T & h\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{u}_{m_s-1} \\ \phi(\mathbf{z}, \tilde{\mathbf{x}}_{m_s})\end{bmatrix} \quad (36a)$$

$$= \begin{bmatrix}\mathbf{u}_{m_s-1}^T & \phi(\mathbf{z}, \tilde{\mathbf{x}}_{m_s})\end{bmatrix}\begin{bmatrix}\mathbf{K}_{m_s-1}^{-1} + c\mathbf{K}_{m_s-1}^{-1}\mathbf{v}_{m_s-1}\mathbf{v}_{m_s-1}^T\mathbf{K}_{m_s-1}^{-1} & -c\mathbf{K}_{m_s-1}^{-1}\mathbf{v}_{m_s-1} \\ -c\mathbf{v}_{m_s-1}^T\mathbf{K}_{m_s-1}^{-1} & c\end{bmatrix}\begin{bmatrix}\mathbf{u}_{m_s-1}^T \\ \phi(\mathbf{z}, \tilde{\mathbf{x}}_{m_s})\end{bmatrix} \quad (36b)$$

$$= \mathbf{u}_{m_s-1}^T\mathbf{K}_{m_s-1}^{-1}\mathbf{u}_{m_s-1} + \frac{(\mathbf{v}_{m_s-1}^T\mathbf{K}_{m_s-1}^{-1}\mathbf{u}_{m_s-1})^2 + \phi^2(\mathbf{z}, \tilde{\mathbf{x}}_{m_s}) - 2\mathbf{v}_{m_s-1}^T\mathbf{K}_{m_s-1}^{-1}\mathbf{u}_{m_s-1}\phi(\mathbf{z}, \tilde{\mathbf{x}}_{m_s})}{c} \quad (36c)$$

$$= \mathbf{u}_{m_s-1}^T\mathbf{K}_{m_s-1}^{-1}\mathbf{u}_{m_s-1} + \frac{(\mathbf{v}_{m_s-1}^T\mathbf{K}_{m_s-1}^{-1}\mathbf{u}_{m_s-1} - \phi(\mathbf{z}, \tilde{\mathbf{x}}_{m_s}))^2}{c} \quad (36d)$$

$$\geq \mathbf{u}_{m_s-1}^T\mathbf{K}_{m_s-1}^{-1}\mathbf{u}_{m_s-1}. \quad (36e)$$

where equality (36b) follows from the block matrix inversion lemma (Hager, 1989), and $c = (h - \mathbf{v}_{m_s-1}^T \mathbf{K}_{m_s-1}^{-1} \mathbf{v}_{m_s-1})^{-1}$, which is always non-negative.

By (36) and the induction step,

$$\mathbb{E}_{\Omega_s}(\mathbf{u}_{m_s}^T \mathbf{K}_{m_s}^{-1} \mathbf{u}_{m_s}) \geq \mathbb{E}_{\Omega_s}(\mathbf{u}_{m_s-1}^T \mathbf{K}_{m_s-1}^{-1} \mathbf{u}_{m_s-1}) \geq \frac{1}{h} \mathbb{E}_{\Omega_s}(\phi^2(\mathbf{x}, \mathbf{x}')). \tag{37}$$

□

C.2 Proof of Theorem 1

First, we prove the following lemma

Lemma 3. *For the random variables $z_1 \sim \mathcal{U}(a, a+e)$ and $z_2 \sim \mathcal{U}(b, b+e)$, where $a \leq b$ and $a, b, e \geq 0$, define $v = (z_1 - z_2)^2$. Then $\mathbb{E}_v(\exp(-cv)) \leq \mathbb{E}_v(\exp(-cv) \mid a = b)$ for any $c > 0$.*

Proof. Let $z = z_1 - z_2$, then by convolution of probability distributions, we have:

$$f_z(t) = \int_{-\infty}^{+\infty} f_{z_1}(t + z_2) f_{z_2}(z_2) dz_2 = \frac{1}{e} \int_b^{b+e} f_{z_1}(t + z_2) dz_2, \tag{38}$$

where the last equation follows from the fact that $f_{z_2} = \frac{1}{e}$ if $b \leq z_2 \leq b+e$. Note that the integrand $f_{z_1}(z + z_2)$ is zero unless $a \leq t + z_2 \leq a + e$, which implies $a - t \leq z_2 \leq a + e - t$. Figure 6 shows the region defined by $a - t \leq z_2 \leq a + e - t$ and $b \leq z_2 \leq b + e$, for the case that $a + b < e$ and $a + e > b$. In the both cases, integration (38) can be calculated as follows:

$$f_z(t) = \begin{cases} \frac{1}{e^2} \int_{a-e-d}^{t} dz_2 & a - b - e \leq t < a - b \\ \frac{1}{e^2} \int_{t}^{a-e} dz_2 & a - b \leq t \leq a - b + e \end{cases} = \begin{cases} \frac{1}{e^2}(t + b - a + e) & a - b - e \leq t < a - b \\ \frac{-1}{e^2}(t + b - a - e) & a - b \leq t \leq a - b + e. \end{cases} \tag{39}$$

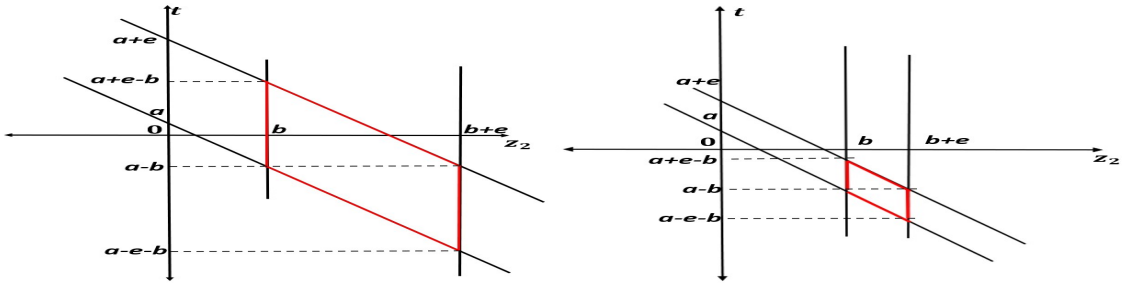


Figure 6: The region defined by $a - t \leq z_2 \leq a + e - t$ and $b \leq z_2 \leq b + e$. Left panel corresponds to the case when $a + b > e$ and right panel corresponds to the case when $a + e < b$.

Hence, $F_v(t) = p(v \leq t) = p(z^2 \leq t) = p(\sqrt{t} \leq z \leq \sqrt{t})$ can be written as

$$F_v(t) = \begin{cases} \frac{2\sqrt{t}}{e^2}(a - b + e) & 0 \leq \sqrt{t} < b - a, \\ \frac{1}{e^2}(2\sqrt{t}e - t - (a - b)^2) & b - a \leq \sqrt{t} < a - b + e, \\ 1 - \frac{1}{2e^2}(\sqrt{t} + a - b - e)^2 & a - b + e \leq \sqrt{t} \leq b - a + e. \end{cases} \tag{40}$$

Moreover, $G_v(t) = p(v \leq t \mid a = b) = p(z^2 \leq t \mid a = b) = p(\sqrt{t} \leq z \leq \sqrt{t} \mid a = b)$ can be derived by setting $a = b$ in CDF (40)

$$G_v(t) = \frac{1}{e^2}(2\sqrt{t}e - t) \;\; 0 \leq \sqrt{t} \leq e. \tag{41}$$

Comparing $G_v(t)$ and $F_v(t)$ for all possible values of t gives

- $\sqrt{t} < 0$: $G_v(t) = F_v(t) = 0$.
- $0 \leq \sqrt{t} < b - a$: then $F_v(t) - G_v(t) = \frac{1}{e^2}(2\sqrt{t}(a - b) + t)$. Since $\sqrt{t} < b - a \Rightarrow t < \sqrt{t}(b - a) \Rightarrow t + \sqrt{t}(a - b) < 0 \Rightarrow t + 2\sqrt{t}(a - b) < 0 \Rightarrow F_v(t) - G_v(t) < 0 \Rightarrow F_v(t) < G_v(t)$.
- $b - a \leq \sqrt{t} < a - b + e$: then $F_v(t) - G_v(t) = -\frac{(a-b)^2}{e^2} < 0 \Rightarrow F_v(t) - G_v(t) < 0 \Rightarrow F_v(t) < G_v(t)$.
- $a - b + e \leq \sqrt{t} < e$: then $F_v(t) - G_v(t) = 1 - \frac{1}{2e^2}(\sqrt{t} + a - b - e)^2 + \frac{1}{e^2}(t - 2\sqrt{t}e)$.

  Note that $\frac{e(F_v(t) - G_v(t))}{et} = \frac{1}{2e^2}(1 - \frac{a-b+e}{\sqrt{t}}) > 0$, and therefore, $F_v(t) - G_v(t)$ is a monotonically increasing function. Due to the monotonicity of $F_v(t) - G_v(t)$, the maximum occurs at e, so

$\max_t F_v(t) - G_v(t) = F_v(e) - G_v(e) = -\frac{(a-b)^2}{e^2} < 0$. Therefore, $F_v(t) - G_v(t) \leq F_v(e) - G_v(e) < 0 \Rightarrow F_v(t) \leq G_v(t)$.

- $e \leq \sqrt{t} < b - a + e$: in this case $G_v(t)$ is always 1, hence, $F_v(t) \leq G_v(t)$.
- $b - a + e \leq \sqrt{t}$: in this case $G_v(t) = F_v(t) = 1$

Therefore, we can conclude that

$$p(v \leq t) \leq p(v \leq t \mid a = b)\ \forall t \in \mathbb{R} \Rightarrow p(-cv \geq t') \leq p(-cv \geq t' \mid a = b)\ \forall t' \in \mathbb{R} \text{ and } c > 0,$$

which implies that random variable $(-cv)$ is stochastically less than random variable $(-cv \mid a = b)$, i.e., $-cv \preceq_{st} -cv \mid a = b$. Consequently, the expectation of any non-decreasing function of these two variables are ordered, i.e., $\mathbb{E}_v(\exp(-cv)) \leq \mathbb{E}_v(\exp(-cv) \mid a = b)$ for any $c > 0$. □

To proceed to the proof of Theorem 1, we use the following characterization for the cutting hyperplanes and subdomains. Assuming that the cutting hyperplanes are equidistant with distant $W = L/S$ from each other, we can characterize the $\ell^{\text{th}} \in [S-1]$ cutting hyperplane on Ω with respect to k^{th} primary axis of $\mathbb{R}^p$ using the vector of angles $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_p\} \backslash \{\theta_k\}$,

$$H_{\boldsymbol{\theta},k,W,\ell} = \{\mathbf{x} \in \Omega \mid x_k - \sum_{j \in [p] \backslash \{k\}} \tan(\theta_j) x_j - \ell W = 0\} \quad \forall \ell \in [S-1]. \tag{42}$$

Note that this cutting hyperplane is orthogonal to the axis k only if $\boldsymbol{\theta} = \mathbf{0}$, that is $\theta_j = 0$ for $j \in [p] \backslash \{k\}$.

Denoting, respectively, the hyperplanes containing the "bottom" and the "top" faces of Ω as

$$H_{\boldsymbol{\theta},k,W,0} = \{\mathbf{x} \in \Omega \mid x_k = 0\} \quad \text{and} \quad H_{\boldsymbol{\theta},k,W,S} = \{\mathbf{x} \in \Omega \mid x_k - L = 0\},$$

we define the s^{th} subdomain as the intersection of area between two consecutive hyperplanes and Ω, specifically,

$$\Omega_{\boldsymbol{\theta},k,W,s} = \{\mathbf{x} \in \Omega \mid \min_{\mathbf{x}' \in H_{\boldsymbol{\theta},k,W,s-1}} ||\mathbf{x} - \mathbf{x}'||_2 \leq W \quad \text{and} \quad \min_{\mathbf{x}' \in H_{\boldsymbol{\theta},k,W,s}} ||\mathbf{x} - \mathbf{x}'||_2 \leq W\}, \tag{43}$$

where $\|\cdot\|_2$ denotes the Euclidean norm.

Proof of Theorem 1. Let $\mathbf{x}_{\{k\}} = \{x_1, \ldots, x_p\} \backslash \{x_k\}$ for any $\mathbf{x} \in \Omega$. Then, based on how each $\Omega_{\boldsymbol{\theta},k,W,s}$ in (43) is constructed and considering the distribution of the data points in Ω according to (16), all variables $x_j \in \mathbf{x}_{\{i\}}$ are independent and have the uniform distribution $\mathcal{U}(0, L)$. Moreover, by the definition of the hyperplanes in (42), and given $\mathbf{x}_{\{k\}}$, the corresponding values of the variable x_k on the hyperplanes $H_{\boldsymbol{\theta},k,W,s-1}$ and $H_{\boldsymbol{\theta},k,W,s}$ are

$$\sum_{j \in [p] \backslash \{k\}} \tan(\theta_j) x_j + (s-1)w \quad \& \sum_{j \in [p] \backslash \{k\}} \tan(\theta_j) x_j + sw. \tag{44}$$

Therefore, the conditional distribution $x_k | \mathbf{x}_{\{k\}}$ in the parallelogram subdomain $\Omega_{\boldsymbol{\theta},k,W,s}$ has a uniform distribution whose support is bounded by the values calculated in (44). Consequently, given a parallelogram subdomain $\Omega_{\boldsymbol{\theta},k,W,s}$, for any $\mathbf{x} \in \Omega_{\boldsymbol{\theta},k,W,s-1}$,

$$x_j \sim \mathcal{U}(0, L) \quad \forall j \in [p] \backslash \{k\}, \tag{45a}$$

$$x_i | \mathbf{x}^i \sim \mathcal{U}\Big(\sum_{j \in [p] \backslash \{k\}} \tan(\theta_j) x_j + (s-1)w, \sum_{j \in [p] \backslash \{k\}} \tan(\theta_j) x_j + sw \Big). \tag{45b}$$

Now that we have the distribution (45), we expand $\mathbb{E}_{\Omega_{\boldsymbol{\theta},k,W,s}}\big(\phi(\mathbf{x}, \mathbf{x}')\big)$ by conditioning, that is

$$\mathbb{E}_{\Omega_{\boldsymbol{\theta},k,W,s}}\big(\phi(\mathbf{x}, \mathbf{x}')\big) = \mathbb{E}_{\mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}}} \Big(\mathbb{E}_{x_k, x'_k}\big(\phi(\mathbf{x}, \mathbf{x}') \mid \mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}}\big) \Big) \tag{46a}$$

$$= \mathbb{E}_{\mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}}} \Big(\exp\Big(- \sum_{j \in [p] \backslash \{k\}} \gamma_j (x_j - x'_j)^2 \Big) \mathbb{E}_{x_k, x'_k} \Big(\exp\big(- \gamma_k (x_k - x'_k)^2 \big) \mid \mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}} \Big) \Big) \tag{46b}$$

$$= \mathbb{E}_{\mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}}} \big(g(\mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}}) h(\mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}}) \big). \tag{46c}$$

Note that the function $g(\mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}})$ is always positive and independent of $\boldsymbol{\theta}$, and function $h(\mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}})$ is positive that attains its maximum for any given $\mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}}$ at $\boldsymbol{\theta} = \mathbf{0}$ by Lemma (3). Therefore, $\boldsymbol{\theta} = \mathbf{0}$,

$$g(\mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}}) h(\mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}}) \leq g(\mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}}) h(\mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}} \mid \boldsymbol{\theta} = \mathbf{0}) \quad \forall \mathbf{x}_{\{k\}}, \mathbf{x}'_{\{k\}},$$

which results in

$$\mathbb{E}_{\Omega_{\boldsymbol{\theta},k,W,s}}\big(\phi(\mathbf{x}, \mathbf{x}')\big) \leq \mathbb{E}_{\Omega_{\boldsymbol{\theta},k,W,s}}\big(\phi(\mathbf{x}, \mathbf{x}') \mid \boldsymbol{\theta} = \mathbf{0}\big)$$

$$\Rightarrow \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\Omega_{\boldsymbol{\theta},k,W,s}}\big(\phi(\mathbf{x},\mathbf{x}')\big) = \mathbf{0}.$$

□

C.3 Proof of Theorem 2

First, we prove the following lemma

Lemma 4. $\mathbb{E}_{z_1,z_2}\Big(\exp\big(-c(z_1-z_2)^2\big)\Big) = \int_0^{b^2}\exp(-ct)(\frac{1}{b\sqrt{t}}-\frac{1}{b^2})dt$, *where* $z_1, z_2 \overset{i.i.d}{\sim} \mathcal{U}(a, a+b)$.

Proof. Let $v = (z_1 - z_2)^2$, then Philip (2007) shows that v has the following PDF:

$$f_v(t) = \frac{1}{\sqrt{t}b} - \frac{1}{b^2} \quad \forall\, 0 \le t \le b^2;$$

therefore,

$$\mathbb{E}_{z_1,z_2}\Big(\exp\big(-c(z_1-z_2)^2\big)\Big) = \mathbb{E}_v\Big(\exp(-cv)\Big) =$$

$$\int_0^{b^2}\exp(-ct)f_s(t)dt = \int_0^{b^2}\exp(-ct)(\frac{1}{b\sqrt{t}}-\frac{1}{b^2})dt.$$

□

Proof of Theorem 2. By the assumptions of uniform distribution of points in Ω (16), and independence of the dimensions due to geometry of $\Omega_{\mathbf{0},k,W,s}$, for any $\mathbf{x} \in \Omega_{\mathbf{0},i,W,s}$,

$$x_k \sim \mathcal{U}\big((s-1)W, sW\big) \quad \& \quad x_j \sim \mathcal{U}\big(0, L\big) \quad \forall j \in [p]\backslash\{k\}. \tag{47}$$

Letting $G_k = \mathbb{E}_{\Omega_{\mathbf{0},k,W,s}}(\phi(\mathbf{x},\mathbf{x}'))$, and using distribution (47),

$$G_k = \mathbb{E}_{x_k}\Big(\exp\big(-\gamma_k(x_k-x_k')^2\big)\Big)\prod_{j\in[p]\backslash\{k\}}\mathbb{E}_{x_j}\Big(\exp\big(-\gamma_j(x_j-x_j')^2\big)\Big) \tag{48a}$$

$$= \mathbb{E}_{v_k}\Big(\exp(-\gamma_k v_k)\Big)\prod_{j\in[p]\backslash\{k\}}\mathbb{E}_{v_j}\Big(\exp(-\gamma_j v_j)\Big) \tag{48b}$$

$$= \Big(\int_0^{W^2}\exp(-\gamma_k t)(\frac{1}{W\sqrt{t}}-\frac{1}{W^2})dt\Big)\Big(\prod_{j\in[p]\backslash\{k\}}\Big(\int_0^{L^2}\exp(-\gamma_j t)(\frac{1}{L\sqrt{t}}-\frac{1}{L^2})dt\Big)\Big) \tag{48c}$$

$$= \Big(\int_0^{W^2} g_k^W(t)dt\Big)\Big(\prod_{j\in[p]\backslash\{k\}}\Big(\int_0^{L^2} g_j^L(t)dt\Big)\Big), \tag{48d}$$

where equality (48a) follows from the independence of dimensions in each $\Omega_{\mathbf{0},i,W,s}$, equalities (48b) and (48c) follow from Lemma (4) with $f_{v_k}(t) = \frac{1}{\sqrt{t}W} - \frac{1}{W^2}$ $0 \le t \le W^2$ and $f_{v_j}(t) = \frac{1}{\sqrt{t}L} - \frac{1}{L^2}$ $0 \le t \le L^2$, and $g_\ell^m(t) = \exp(-\gamma_\ell t)(\frac{1}{m\sqrt{t}} - \frac{1}{m^2})$ in (48d).

To show that $G_p - G_k \ge 0$ for any $k \in [p]$, We first expand $G_p - G_k$,

$$G_p - G_k = \left(\int_0^{W^2} g_p^W(t)dt\right)\left(\prod_{j\in[p]\setminus\{p\}}\left(\int_0^{L^2} g_j^L(t)dt\right)\right) - \left(\int_0^{W^2} g_k^W(t)dt\right)\left(\prod_{j\in[p]\setminus\{k\}}\left(\int_0^{L^2} g_j^L(t)dt\right)\right)$$

$$= \left(\prod_{j\in[p]\setminus\{k,p\}}\left(\int_0^{L^2} g_j^L(t)dt\right)\right)\left(\int_0^{W^2} g_p^W(t)dt\int_0^{L^2} g_k^L(t)dt - \int_0^{W^2} g_k^W(t)dt\int_0^{L^2} g_p^L(t)dt\right) = A * B.$$

Note that A is always positive, since each $\int_0^{L^2} g_j^L(t)dt$ is the expectation of the random variable $\exp(-\gamma_j v_j)$ which is positive. Hence, it is enough to show that B is positive. Expanding B further,

$$B = \left(\int_0^{W^2} g_p^W(t)dt\right)\left(\int_0^{W^2} g_k^L(t)dt + \int_{W^2}^{L^2} g_k^L(t)dt\right) - \left(\int_0^{W^2} g_k^W(t)dt\right)\left(\int_0^{W^2} g_p^L(t)dt + \int_{W^2}^{L^2} g_p^L(t)dt\right) \tag{49a}$$

$$= \int_{t_k:0}^{W^2}\int_{t_p:0}^{W^2} g_p^W(t_k)g_k^L(t_p)dt_k dt_p + \int_{t_k:0}^{W^2}\int_{t_p:W^2}^{L^2} g_p^W(t_k)g_k^L(t_p)dt_k dt_p$$

$$- \int_{t_k:0}^{w^2}\int_{t_p:0}^{w^2} g_k^w(t_k)g_p^L(t_p)dt_k dt_p - \int_{t_k:0}^{w^2}\int_{t_p:w^2}^{L^2} g_k^w(t_k)g_p^L(t_p)dt_k dt_p \tag{49b}$$

$$= \int_{t_k:0}^{W^2}\int_{t_p:0}^{W^2} \big(\exp(-\gamma_p t_k - \gamma_k t_p) - \exp(-\gamma_k t_k - \gamma_p t_p)\big)\big(\frac{1}{W\sqrt{t_k}} - \frac{1}{W^2}\big)\big(\frac{1}{L\sqrt{t_p}} - \frac{1}{L^2}\big)dt_k dt_p$$

$$+ \int_{t_k:0}^{W^2}\int_{t_p:W^2}^{L^2} \big(\exp(-\gamma_p t_k - \gamma_k t_p) - \exp(-\gamma_k t_k - \gamma_p t_p)\big)\big(\frac{1}{W\sqrt{t_k}} - \frac{1}{W^2}\big)\big(\frac{1}{L\sqrt{t_p}} - \frac{1}{L^2}\big)dt_k dt_p \tag{49c}$$

$$= \int_{t_k:0}^{W^2}\int_{t_p:0}^{W^2} c(t_k,t_p)dt_k dt_p + \int_{t_k:0}^{W^2}\int_{t_p:W^2}^{L^2} c(t_k,t_p)dt_k dt_p. \tag{49d}$$

Note that for any member of set

$$\{(W, L, t_p, t_k, \gamma_p, \gamma_k) \mid 0 < W < L,\ 0 < \gamma_k < \gamma_p,\ 0 \le t_k \le W^2,\ W^2 \le t_p \le L^2\}, \tag{50}$$

we have

$$\big(\frac{1}{w\sqrt{t_k}} - \frac{1}{w^2}\big)\big(\frac{1}{L\sqrt{t_p}} - \frac{1}{L^2}\big) > 0, \tag{51}$$

and also

$$(-\gamma_p t_k - \gamma_k t_p) - (-\gamma_k t_k - \gamma_p t_p) = (\gamma_p - \gamma_k)(t_p - t_k) > 0, \tag{52}$$

where the latter results in

$$\exp(-\gamma_p t_k - \gamma_k t_p) - \exp(-\gamma_k t_k - \gamma_p t_p) > 0. \tag{53}$$

Therefore, by (51) and (53), the integrand $c(t_k, t_p)$ in (49d) is positive for any member of set (50), so is integral $\int_{t_k:0}^{W^2} \int_{t_p:W^2}^{L^2} c(t_k, t_p) dt_k dt_p$. Hence, to complete the proof we need to show integral $\int_{t_k:0}^{w^2} \int_{t_p:0}^{w^2} c(t_k, t_p) dt_k dt_p$ in (49d) is also positive. To show this, we expand the integral,

$$\int_{t_k:0}^{W^2} \int_{t_p:0}^{W^2} c(t_k, t_p) dt_k dt_p = \int_{t_k:0}^{W^2} \int_{t_p:t_k}^{W^2} c(t_k, t_p) dt_k dt_p + \int_{t_p:0}^{W^2} \int_{t_k:t_p}^{W^2} c(t_k, t_p) dt_p dt_k \tag{54a}$$

$$= \int_{t_k:0}^{W^2} \int_{t_p:t_k}^{W^2} c(t_k, t_p) dt_k dt_p + \int_{t_k:0}^{W^2} \int_{t_p:t_k}^{W^2} c(t_p, t_k) dt_k dt_p = \int_{t_k:0}^{W^2} \int_{t_p:t_k}^{W^2} \big(c(t_k, t_p) + c(t_p, t_k)\big) dt_k dt_p \tag{54b}$$

$$= \frac{1}{wL} \int_{t_k:0}^{W^2} \int_{t_p:t_k}^{W^2} \big(\exp(-\gamma_p t_k - \gamma_k t_p) - \exp(-\gamma_k t_k - \gamma_p t_p)\big)(\frac{1}{\sqrt{t_k}} - \frac{1}{\sqrt{t_p}})(\frac{1}{W} - \frac{1}{L}) dt_k dt_p. \tag{54c}$$

Similar to (50)-(53), for any member of set

$$\{(W, L, t_p, t_k, \gamma_p, \gamma_k) \mid 0 < W < L,\ 0 < \gamma_k < \gamma_p,\ 0 \le t_k \le W^2,\ t_k \le t_p \le W^2\}, \tag{55}$$

we have $(\frac{1}{\sqrt{t_k}} - \frac{1}{\sqrt{t_p}}) > 0$, $(\frac{1}{W} - \frac{1}{L}) > 0$, and $\big(\exp(-\gamma_p t_k - \gamma_k t_p) - \exp(-\gamma_k t_k - \gamma_p t_p)\big) > 0$. Hence the integrand in (54c) is positive for any member of set (55), so is integral (54c), and the proof is complete.

□

Appendix D A simulation study on the relation between expected error (15) and $\mathbb{E}_{\Omega_s}(\phi^2(\mathbf{x}, \mathbf{x}'))$

Consider the squared exponential Gaussian kernel $\phi(x, x') = exp(-\gamma(x - x')^2)$ with $\gamma > 0$ defined on

$$\Omega_s = \{x \in \mathbb{R} | a \leq x \leq a + b\} \tag{56}$$

with uniform sampling distribution

$$x \sim \mathcal{U}(a, a + b) \quad \forall x \in \Omega_s. \tag{57}$$

To have a general simulation study, we need the following lemma.

Lemma 5. $\mathbb{E}_{z_1,z_2}\Big(\exp\big(-c(z_1 - z_2)^2\big)\Big)$, *where* $z_1, z_2 \overset{i.i.d}{\sim} \mathcal{U}(a, a+b)$, *is a monotonically decreasing function of* c *and* b.

Proof. We need to show that $\nabla g(b, c) = [\frac{\partial g(b,c)}{\partial b}, \frac{\partial g(b,c)}{\partial c}]^T < 0$ for all $[b, c]^T > 0$, where

$$g(b, c) = \mathbb{E}_{z_1,z_2}\Big(\exp\big(-c(z_1 - z_2)^2\big)\Big) = \int_0^{b^2} \exp(-ct)(\frac{1}{b\sqrt{t}} - \frac{1}{b^2})dt$$

by Lemma 4.

We can write $\frac{\partial g(b,c)}{\partial b}$ as

$$\frac{\partial g(b, c)}{\partial b} = \frac{1}{b^2}\int_0^{b^2} \exp(-ct)(\frac{2}{b} - \frac{1}{\sqrt{t}})dt \tag{58a}$$

$$= \frac{1}{b^2}\left(\left[\exp(-ct)(\frac{2t}{b} - 2\sqrt{t})\right]_0^{b^2} - \int_0^{b^2} -c\exp(-ct)(\frac{2t}{b} - 2\sqrt{t})\right) \tag{58b}$$

$$= \frac{2c}{b^2}\int_0^{b^2} \exp(-ct)(\frac{t}{b} - \sqrt{t}), \tag{58c}$$

where equalities (58a) and (58b) follow from the Leibniz integral differentiation and the integration by part rules, respectively. It is easy to check that integrand $\exp(-ct)(\frac{t}{b} - \sqrt{t})$ is always negative for any member of set $\{(b, c, t) \mid 0 < b, 0 < c, 0 \leq t \leq b^2\}$; therefore, we always have $\frac{\partial g(b,c)}{\partial b} < 0$.

Moreover, for $\frac{\partial g(b,c)}{\partial c}$,

$$\frac{\partial g(b,c)}{\partial c} = \int_0^{b^2} -t\exp(-ct)(\frac{1}{b\sqrt{t}} - \frac{1}{b^2})dt = \frac{-1}{b}\int_0^{b^2} t\exp(-ct)(\frac{1}{\sqrt{t}} - \frac{1}{b})dt.$$

It is again easy to check that the integrand $t\exp(-ct)(\frac{1}{\sqrt{t}} - \frac{1}{b})$ is positive for any member of set $\{(b,c,t) \mid 0 < b, 0 < c, 0 \leq t \leq b^2\}$. Therefore, $\frac{\partial g(b,c)}{\partial c}$ is always negative. □

By Lemma 5, expectation function

$$\mathbb{E}_{\Omega_s}(\phi^2(\mathbf{x},\mathbf{x}')) = \mathbb{E}_{x,x'}(\exp(-2\gamma(x-x')^2) \tag{59}$$

is a monotonically decreasing function of γ and b. This means that there are only two ways to increase expectation $\mathbb{E}_{\Omega_s}(\phi^2(\mathbf{x},\mathbf{x}'))$, which are either decreasing γ or decreasing b. The approximation of expected error function (15) on domain (56) and sampling distribution (57) for varying values of γ and b and a fixed value of m_s using a heat map plot is shown in Figure 7. We observe that as the values of γ or b decrease, or equivalently, $\mathbb{E}_{\Omega_s}(\phi^2(\mathbf{x},\mathbf{x}'))$ increases, the approximation of the expected error function decreases.

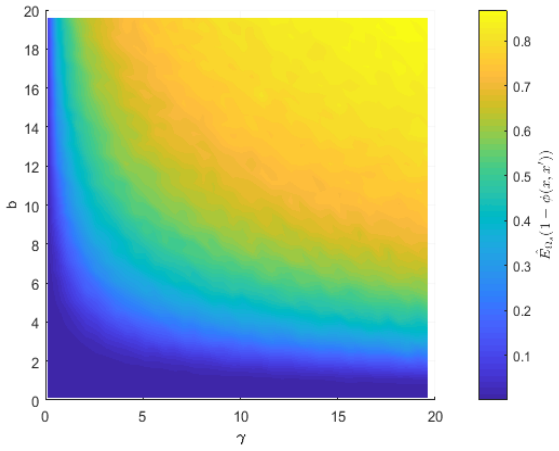


Figure 7: Heat map of the approximation of expected error function (15) on domain (56) and sampling distribution (57) for varying values of γ and b and a fixed value of m_s

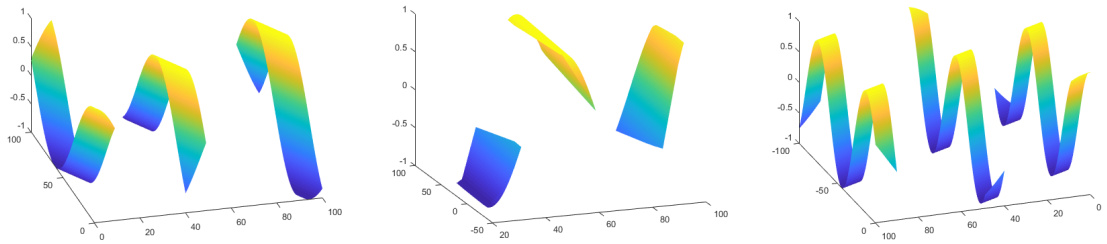
Our simulation study can be used to infer a more general case. Consider the covariance function as $\phi(\mathbf{x},\mathbf{x}') = \exp(-\sum_{i=1}^{p}\gamma_k(x_k - x'_k))$ defined on Ω_s as a p-dimensional hyper-rectangle with side lengths $b_1,\ldots,b_p$ with a uniform sampling distribution, i.e., $x_k \sim \mathcal{U}(a_k, a_k + b_k) \quad \forall \mathbf{x} \in \Omega_s$. With

this setup, we can write

$$\mathbb{E}_{\Omega_s}(\phi^2(\mathbf{x},\mathbf{x}')) = \prod_{i=1}^{p} \mathbb{E}_{x_k,x_k'}(\exp(-2\gamma_k(x_k - x_k')), \tag{60}$$

which is a monotonic function in each b_i and γ_i by lemma 5. Therefore, our simulation results are valid for this generalized case as well.

Finally, we present some intuition behind the theoretical results in Section 3. The reason why the direction $\mathbf{a}$, found by solving optimization problem (20), results in a better covariance approximation in each subdomain can be visually perceived for a two-dimensional domain. Suppose we can partition the domain of two-dimensional function $f(\mathbf{x}) = \cos(0.05x_1 + 0.1x_2)$ by cutting orthogonal to either of three directions $[1,0]$, $[0.43,0.9]$, or $[0,1]$, where direction $[0.43,0.9]$ is the direction of the fastest covariance decay obtained by optimizing (20). Figure 8 shows the 3-D presentations of three local functions created by cutting orthogonal to each direction. We observe that the local functions created by cutting orthogonal to the desired direction have a less fluctuating behaviour compared to those of directions $[1,0]$ and $[0,1]$. That the function has less fluctuation allows a random point on the local functions of Figure 8b to have (on average) higher correlation to its neighboring data points. Therefore, we can obtain a better approximation of local covariance structures by using the same number of pseudo data points located in each subdomain.



(a) $\mathbf{a} = [1,0]^t$ (b) $\mathbf{a} = [0.43,0.9]^t$ (c) $\mathbf{a} = [0,1]^t$

Figure 8: Local functions created by cutting orthogonal to directions $[1,0]$, $[0.43,0.9]$ (solution of (19)), and $[0,1]$ on a synthetic dataset

Appendix E Solving optimization problem (20)

Let first write the partial derivatives of objective function in (20),

$$\frac{\partial \mathcal{L}(\bar{\mathbf{a}})}{\partial a_k} = -\mathbf{y}_n^T(\mathbf{K}_n^{\bar{\mathbf{a}}} + \sigma^2\mathbf{I_n})^{-1}\frac{\partial \mathbf{K}_n^{\bar{\mathbf{a}}}}{\partial a_k}(\mathbf{K}_n^{\bar{\mathbf{a}}} + \sigma^2\mathbf{I_n})^{-1}\mathbf{y}_n + \mathrm{tr}((\mathbf{K}_n^{\bar{\mathbf{a}}} + \sigma^2\mathbf{I_n})^{-1}\frac{\partial \mathbf{K}_n^{\bar{\mathbf{a}}}}{\partial a_k}), \tag{61}$$

where $\frac{\partial \mathbf{K}_n^{\bar{\mathbf{a}}}}{\partial a_k}$ is the matrix of element-wise derivatives with respect to the k^{th} element of $\bar{\mathbf{a}}$. Note that each element of $\frac{\partial \mathbf{K}_n^{\bar{\mathbf{a}}}}{\partial a_k}$ involves the term $\frac{1}{\sqrt{1-\bar{\mathbf{a}}^T\bar{\mathbf{a}}}}$. Therefore, the gradient of the objective function in (20) does not exist on the boundary of the feasible region, i.e., $\nabla\mathcal{L}(\bar{\mathbf{a}}) \to \infty$ as $\bar{\mathbf{a}}^T\bar{\mathbf{a}} \to 1$. Therefore, to avoid an undefined gradient on the boundary, we modify the optimization by making the feasible region slightly tighter, i.e.,

$$\begin{aligned} \min_{\bar{\mathbf{a}}} \quad & \mathcal{L}(\bar{\mathbf{a}}) = \mathbf{y}_n^T(\mathbf{K}_n^{\bar{\mathbf{a}}} + \sigma^2\mathbf{I_n})^{-1}\mathbf{y}_n + \log|\mathbf{K}_n^{\bar{\mathbf{a}}} + \sigma^2\mathbf{I}_n| \\ \text{subject to} \quad & \bar{\mathbf{a}}^T\bar{\mathbf{a}} \leq 1-\epsilon, \end{aligned} \tag{62}$$

where ϵ is a very small number. In our experiments, we set $\epsilon = 0.001$.

Due to the simple convex structure of constraint $\bar{\mathbf{a}}^T\bar{\mathbf{a}} \leq 1-\epsilon$, i.e., a $d-1$-dimensional hypersphere, optimization (62) can be solved by the Projected Gradient Descent algorithm (Nesterov and Nemirovskii, 1994). In this projection algorithm, the $(j+1)^{\text{th}}$ decent step is defined by

$$\bar{\mathbf{a}}^{j+1} = \mathcal{P}\big(\bar{\mathbf{a}}^j - \frac{\alpha}{||\nabla\mathcal{L}(\bar{\mathbf{a}}^j)||}\nabla\mathcal{L}(\bar{\mathbf{a}}^j)\big), \tag{63}$$

where $\frac{\alpha}{||\nabla\mathcal{L}(\bar{\mathbf{a}}^j)||}$ is a normalized length step, and

$$\begin{aligned} \mathcal{P}(\mathbf{z}) = \quad & \mathrm{argmin}_{\mathbf{w}}||\mathbf{w}-\mathbf{z}|| \\ \text{subject to} \quad & \mathbf{w}^T\mathbf{w} \leq 1-\epsilon. \end{aligned} \tag{64}$$

$\mathcal{P}(\mathbf{z}) = \mathbf{z}$ when $\mathbf{z}^T\mathbf{z} \leq 1-\epsilon$, otherwise the solution to $\mathcal{P}(\mathbf{z})$ occurs at the point that the line defined by $\mathbf{z}$ and the center of the hypersphere, $(\mathbf{0})$, crosses the boundary of the hypersphere, i.e, intersection of $\frac{w_1}{z_1} = \frac{w_2}{z_2} = \ldots = \frac{w_{p-1}}{z_{p-1}}$ and $\mathbf{w}^T\mathbf{w} = 1-\epsilon$. Therefore, the solution to $\mathcal{P}(\mathbf{z})$ has the

closed form,

$$\mathcal{P}(\mathbf{z}) = \begin{cases} \mathbf{z} & \mathbf{z}^T\mathbf{z} \leq 1-\epsilon \\ [\frac{z_1}{\sqrt{\mathbf{z}^T\mathbf{z}}}, \ldots, \frac{z_{p-1}}{\sqrt{\mathbf{z}^T\mathbf{z}}}]^T & \mathbf{z}^T\mathbf{z} > 1-\epsilon. \end{cases} \tag{65}$$

Appendix F Practical Considerations

F.1 Creating boundaries, control points, and boundary functions

The focus of this section is on the practical implementation of SPLK, and therefore, the characterization of cutting hyperplanes differs from the discussion in Section 3.2. Here, instead of using a vector of angles corresponding to primary axes of input space, we use a given direction, which can be the solution to optimization (20) or any other arbitrary direction, to define the cutting hyperplanes.

Recall that in our partitioning policy all the cutting hyperplanes are parallel to each other, and therefore, orthogonal to a unique direction, which is characterized by a vector $\mathbf{a} = [a_1, \ldots, a_p]^T$. Let $\mathbf{Z} = \{\mathbf{x}_i^T\mathbf{a} \mid \mathbf{x}_i \in \mathbf{X}\}$ denote the projection of all the input vectors onto $\mathbf{a}$. Next, consider the ordered set $\{z_1, \ldots, z_{S-1}\}$, where $\min \mathbf{Z} < z_1$ and $z_{S-1} < \max \mathbf{Z}$, and $z_\ell < z_{\ell+1}$, for $\ell \in [S-1]$.

Given the set $\{z_1, \ldots, z_{S-1}\}$ and direction $\mathbf{a}$, which is in fact the normal vector of all of the cutting hyperplanes, we define the ℓ^{th} cutting hyperplane orthogonal to $\mathbf{a}$ as $H_{\ell,\mathbf{a}} = \{\mathbf{x} \in \Omega \mid a_1x_1 + \ldots + a_px_p = z_\ell\}$ for $\ell \in [S-1]$. We use the data points close to $H_{\ell,\mathbf{a}}$ to locate the control points. To this end, we first define $\mathbf{\Delta}_\ell = \{\mathbf{x}_i \in \mathbf{X} | \ |\mathbf{x}_i^T\mathbf{a} - z_\ell| < \delta\}$ as the set of training data points whose Euclidean distance to $\mathcal{H}_{\ell,\mathbf{a}}$ is less than a predefined constant δ. Then, calculate the maximum and minimum of the k^{th} dimension of the data points in $\mathbf{\Delta}_\ell$, respectively,

$$\tau_{1,k,\ell} = \max_{\mathbf{x}_i \in \mathbf{\Delta}_\ell} \mathbf{x}_i^T\mathbf{e}_k \qquad \text{and} \qquad \tau_{0,k,\ell} = \min_{\mathbf{x}_i \in \mathbf{\Delta}_\ell} \mathbf{x}_i^T\mathbf{e}_k, \tag{66}$$

where $\mathbf{e}_k$ is the unit vector along the k^{th} primary axis of the space for $k \in [p]$. As such, the set $\mathbf{V}_\ell = \left\{[\tau_{b,1,\ell}, \ldots, \tau_{b,p,\ell}]^T \,|b = 0, 1\right\}$ characterizes the vertices of the hyper-rectangle inscribing $\mathbf{\Delta}_\ell$. Next, we uniformly sample $Q > 0$ points from $\mathbf{V}_\ell$ and denote the set of all these points as $\mathbf{U}_\ell$. We

obtain the set of control points on $H_{\ell,\mathbf{a}}$ denoted as $\mathbf{C}_\ell$ by projecting the points in $\mathbf{U}_\ell$ on $\mathcal{H}_{\ell,\mathbf{a}}$,

$$\mathbf{C}_\ell = \{(z_\ell - \mathbf{u}^T\mathbf{a})\mathbf{a} + \mathbf{u} \mid \forall \mathbf{u} \in \mathbf{U}_\ell\}. \tag{67}$$

There are several ways to choose the width of each subdomain, i.e., $z_{\ell+1} - z_\ell$ for $\ell \in [S-1]$. One way is to choose a fixed width for the subdomains; however, this approach results in subdomains with different numbers of local data points depending on their distribution on the domain. Also *adaptive mesh generation* techniques (Becker and Rannacher, 2001) can be used to vary the widths to balance the error among the subdomains. In Section 4, we use varying widths for the subdomains to balance the numbers of local data points across the subdomains. This approach helps us to control the computation time of the algorithm, because it is evenly distributed among the subdomains.

Furthermore, to impose connectivity on the optimization procedure discussed in Section 3.1, we need to specify the boundary values for each control point $\mathbf{c} \in \mathbf{C}_\ell$. To this end, we fit a boundary GPR over the hyper-rectangle defined by $\mathbf{V}_\ell$ using the data points in $\mathbf{\Delta}_\ell$. We then use the predictive mean function of this GPR to determine the boundary values. Letting $\mathcal{R}_\ell(.)$ denote as the predictive mean function of the GPR constructed by $\mathbf{\Delta}_\ell$, the boundary value for each $\mathbf{c} \in \mathbf{C}_\ell$ is

$$\mathcal{R}_\ell(\mathbf{c}) = \mathbf{k}_{\mathbf{c}\mathbf{\Delta}_\ell}(\mathbf{K}_{\mathbf{\Delta}_\ell\mathbf{\Delta}_\ell} + \sigma_\ell^2\mathbf{I}_\ell)^{-1}\mathbf{y}_{\mathbf{\Delta}_\ell}, \tag{68}$$

where $\mathbf{k}_{\mathbf{c}\mathbf{\Delta}_\ell}$ is the covariance vector between the control point $\mathbf{c} \in \mathbf{C}_\ell$ and the neighboring data points in $\mathbf{\Delta}_\ell$, and $\mathbf{K}_{\mathbf{\Delta}_\ell\mathbf{\Delta}_\ell}$ is the covariance matrix between the neighboring data points in $\mathbf{\Delta}_\ell$ themselves. In Section 3.1, with a slight abuse of notation, we denote $\mathcal{R}(.)$ as a function that takes a control point as an input and returns $\mathcal{R}_\ell(.)$, depending on the location of the control point. Note that since the set of neighboring data points $\mathbf{\Delta}_\ell$ is a small set, we use a full GPR to obtain functions 68.

Dataset	q	Time	MSE	NLPD
TCO	3	145.50	12.18	2.61
	4	145.61	12.15	2.61
	5	146.06	11.98	2.60
Levitus	3	134.48	25.50	2.60
	4	134.47	25.44	2.59
	5	135.36	25.25	2.59
Dasilva	3	157.62	0.42	4.01
	4	159.98	0.38	3.30
	5	167.79	0.38	3.05
Protein	2.2	147.53	17.41	2.67
	2.5	202.09	17.39	2.66
	3	3651.33	17.38	2.65

Table 1: Effect of q on efficiency of SPLK. $S = 30$ and $\kappa = 4$ across all the datasets

F.2 Control points density

As discussed in Section 3.4, we use a density parameter and the dimension of the boundary space, i.e., q and $p-1$, to determine the number of control points to be uniformly located on each boundary. Notably, our experiments show that setting q to small values usually results in satisfactory performance, while increasing it does not significantly affect the prediction accuracy, but increases the computation burden, particularly in higher dimensional domains. The results of testing SPLK on our four datasets with varying values of q and all other parameters fixed are reported in Table 1. An increase in the value of q slightly improves the prediction accuracy in terms of NLPD and MSE. Moreover, as the dimension of the domain of data increases, an increase in the value of q results in much longer computation time.

F.3 Hyperparameter learning

Maximizing the marginal likelihood of the training data, $p(\mathbf{y})$, is a popular method for learning the hyperparameters of a model (Rasmussen and Williams, 2006). In SPLK, instead of one global marginal likelihood function, there are S local functions $p(\mathbf{y}_s)$, each of which can be trained independently. Recall that our local predictors are in fact SPGP predictors that consider pseudo-inputs as parameters of the model. Therefore, we have two types of parameters: one is the location of local pseudo-inputs and the other is the hyperparameters of the underlying covariance function. Maximizing the logarithm of the local SPGP marginal likelihood functions using gradient descent with respect to local pseudo-inputs and hyperparameters provides local optimal locations. Specifically,

the logarithm of the marginal likelihood of SPLK's s^{th} local model is

$$\log(p(\mathbf{y}_s)) = -\frac{1}{2}\log|\mathbf{G}_s| - \frac{1}{2}\mathbf{y}_s^T\mathbf{G}_s^{-1}\mathbf{y}_s - \frac{n_s}{2}\log 2\pi, \tag{69}$$

where $\mathbf{G}_s$ is the same as that of Section 3.1.

Moreover, we use anti-isotropic squared exponential function as the choice of our local covariance functions,

$$\phi(\mathbf{x}, \mathbf{x}') = C\exp\big(-(\mathbf{x}-\mathbf{x}')^T\mathbf{\Gamma}(\mathbf{x}-\mathbf{x}')\big), \tag{70}$$

where $\mathbf{\Gamma}$ is a diagonal matrix with length-scale parameters $\gamma_1, \ldots, \gamma_p$ on the diagonal. This covarinace function automatically determines the significance of predictors after training its parameters by minimizing local likelihood function (69).

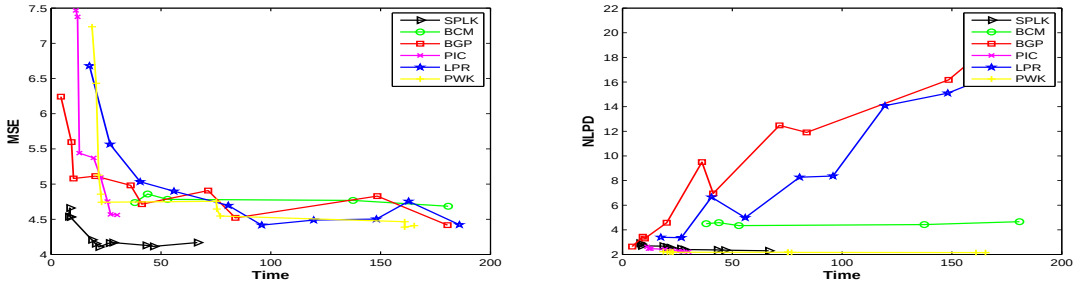
Appendix G A simulation study on the performance of SPLK

In this section, we conduct a simulation study to further investigate the performance of SPLK comparing to the other competing algorithms in terms of MSE. As mentioned in Section 4.2, when the rates of covariance decay highly vary in different directions (similar to the Dataset Dasilva), SPLK can perform better than the competing algorithms considered in this study. This is because SPLK partitions the domain of data orthogonal to the direction of the fastest rate of covariance decay, which potentially reduces the degree of mismatch on the boundaries compared with the other directions.

To test this claim we generate 10,000 samples from a Gaussian process with covariance function (17) and highly different length scale parameters $\gamma_1 = 50$, $\gamma_2 = 10$, and $\gamma_3 = 0.001$. To this end, we first generate 10,000 vectors, $\mathbf{x}_i$, uniformly from the cube $[0,5]\times[0,5]\times[0,5]$ and form the covariance matrix $\mathbf{K_{XX}}$. Then we draw 10,000 responses, y_i, using $\mathbf{K_{XX}}$ and add a noise to each response from distribution $\mathcal{N}(0,4)$. Finally, we use 9,000 of these samples for training and 1,000 fo r testing.

For this simulated dataset, SPLK partitions the domain of data from the first direction which has the largest associated length scale parameter. Figures 9a and 9b show the performance of all the

competing algorithms in terms of MSE and NLPD versus computation time. As expected, due to the designed covariance structure, i.e., highly varying rates of covariance decay, SPLK outperforms the other competing algorithms in terms of MSE, while performs as well as PWK and PIC in terms of NLPD.



(a) MSE vs. computation time

(b) NLPD vs. computation time

Figure 9: MSE and NLPD versus computation time. For SPLK, $q = 3$ and $k \in \{2, 4, 6, 8\}$. The value of parameter S is selected from the set $\{8, 16, 32, 64, 128, 256\}$.

References

Becker, R. and R. Rannacher (2001). An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica 10*, 1–102.

Philip, J. (2007). The probability distribution of the distance between two random points in a box. *TRITA MAT 10(7)*.