Supplement to *Generalization Errors and Learning Curves for Regression with Multi-task Gaussian Processes*

Kian Ming A. Chai
School of Informatics, University of Edinburgh,
10 Crichton Street, Edinburgh EH8 9AB, UK
k.m.a.chai@ed.ac.uk

S.1 Introduction

We provide material supplementary to the the main paper. Much of this will be detailed proofs for the propositions and corollaries. To distinguish from the equations in the main paper, the equations here are prefixed with "S.".

S.2 Proof for Proposition 5

In this section, we give the proof for Proposition 5 in the main text.

S.2.1 Proof for Proposition 5a

Recall from (3) that $\sigma_T^2(\rho)$ is given by

$$\sigma_T^2(\rho) \stackrel{\text{def}}{=} k_{**} - \begin{pmatrix} \boldsymbol{k}_{T*}^{\text{x}} \\ \rho \boldsymbol{k}_{S*}^{\text{x}} \end{pmatrix}^{\text{T}} \begin{pmatrix} K_{TT}^{\text{x}} + \sigma_{\text{n}}^2 I & \rho K_{TS}^{\text{x}} \\ \rho K_{ST}^{\text{x}} & K_{SS}^{\text{x}} + \sigma_{\text{n}}^2 I \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{k}_{T*}^{\text{x}} \\ \rho \boldsymbol{k}_{S*}^{\text{x}} \end{pmatrix} \tag{S.1}$$

To perform the matrix inverse in the above equation, we use the following formula for inverting block matrices:

Theorem S.1. *Banachiewicz inversion formula (see e.g., [1]).*

$$\begin{aligned} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} &= \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1} A_{12} C^{-1} A_{21} A_{11}^{-1} & -A_{11}^{-1} A_{12} C^{-1} \\ -C^{-1} A_{21} A_{11}^{-1} & C^{-1} \end{pmatrix} \\ &= \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} -A_{11}^{-1} A_{12} \\ I \end{pmatrix} C^{-1} \begin{pmatrix} -A_{21} A_{11}^{-1} & I \end{pmatrix}, \end{aligned}$$

where $$C \stackrel{\text{def}}{=} A_{22} - A_{21} A_{11}^{-1} A_{12}.$$

The role of C in the above theorem is played by $A(\rho)$ defined as

$$A(\rho) \stackrel{\text{def}}{=} K_{SS}^{\text{x}} + \sigma_{\text{n}}^2 I - \rho^2 K_{ST}^{\text{x}} \left(K_{TT}^{\text{x}} + \sigma_{\text{n}}^2 I \right)^{-1} K_{TS}^{\text{x}}. \tag{S.2}$$

In addition, we let

$$\begin{aligned} \boldsymbol{v}(\rho) &\stackrel{\text{def}}{=} \begin{pmatrix} -\rho K_{ST}^{\text{x}} (K_{TT}^{\text{x}} + \sigma_{\text{n}}^2 I)^{-1} & I \end{pmatrix} \begin{pmatrix} \boldsymbol{k}_{T*}^{\text{x}} \\ \rho \boldsymbol{k}_{S*}^{\text{x}} \end{pmatrix} \\ &= -\rho K_{ST}^{\text{x}} (K_{TT}^{\text{x}} + \sigma_{\text{n}}^2 I)^{-1} \boldsymbol{k}_{T*}^{\text{x}} + \rho \boldsymbol{k}_{S*}^{\text{x}} \\ &= \rho \left(\boldsymbol{k}_{S*}^{\text{x}} - K_{ST}^{\text{x}} (K_{TT}^{\text{x}} + \sigma_{\text{n}}^2 I)^{-1} \boldsymbol{k}_{T*}^{\text{x}} \right). \end{aligned} \tag{S.3}$$

Then

$$\sigma_T^2(\rho) = k_{**} - \begin{pmatrix} \boldsymbol{k}_{T*}^{\mathrm{x}} \\ \rho \boldsymbol{k}_{S*}^{\mathrm{x}} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} (K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I)^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{k}_{T*}^{\mathrm{x}} \\ \rho \boldsymbol{k}_{S*}^{\mathrm{x}} \end{pmatrix} - \boldsymbol{v}(\rho)^{\mathrm{T}} [A(\rho)]^{-1} \boldsymbol{v}(\rho) \tag{S.4}$$

$$= k_{**} - (\boldsymbol{k}_{T*}^{\mathrm{x}})^{\mathrm{T}} \left(K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I\right)^{-1} \boldsymbol{k}_{T*}^{\mathrm{x}} - \boldsymbol{v}(\rho)^{\mathrm{T}} [A(\rho)]^{-1} \boldsymbol{v}(\rho) \tag{S.5}$$

We can identify $\boldsymbol{v}(\rho) = \rho \boldsymbol{v}(1)$. If we also write $\boldsymbol{v}_1$ for $\boldsymbol{v}(1)$, then

$$\sigma_T^2(\rho) = k_{**} - (\boldsymbol{k}_{T*}^{\mathrm{x}})^{\mathrm{T}} \left(K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I\right)^{-1} \boldsymbol{k}_{T*}^{\mathrm{x}} - \rho^2 \boldsymbol{v}_1^{\mathrm{T}} \left[A(\rho)\right]^{-1} \boldsymbol{v}_1 \tag{S.6}$$

By substituting 0 for ρ above, the first two terms on the right of the equation can be identified with $\sigma_T^2(0)$. Thus

$$\sigma_T^2(\rho) = \sigma_T^2(0) - \rho^2 \boldsymbol{v}_1^{\mathrm{T}} \left[A(\rho)\right]^{-1} \boldsymbol{v}_1. \tag{S.7}$$

Now, observe that $K_{ST}^{\mathrm{x}}(K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I)^{-1} K_{TS}^{\mathrm{x}}$ is positive semi-definite, since we can factorize it into the form XX^{T} for some matrix X. Proceeding from this, we have

$$\begin{aligned}
& K_{ST}^{\mathrm{x}}(K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I)^{-1} K_{TS}^{\mathrm{x}} \succcurlyeq 0 \\
\iff \quad & (1-\rho^2) K_{ST}^{\mathrm{x}}(K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I)^{-1} K_{TS}^{\mathrm{x}} \succcurlyeq 0 \qquad \text{since } \rho^2 \in [0,1] \\
\iff \quad & -\rho^2 K_{ST}^{\mathrm{x}}(K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I)^{-1} K_{TS}^{\mathrm{x}} \succcurlyeq -K_{ST}^{\mathrm{x}}(K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I)^{-1} K_{TS}^{\mathrm{x}} \\
\iff \quad & K_{SS}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I - \rho^2 K_{ST}^{\mathrm{x}}(K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I)^{-1} K_{TS}^{\mathrm{x}} \succcurlyeq K_{SS}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I - K_{ST}^{\mathrm{x}}(K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I)^{-1} K_{TS}^{\mathrm{x}} \\
\text{i.e.,} \quad & A(\rho^2) \succcurlyeq A(1) \\
\iff \quad & [A(\rho^2)]^{-1} \preccurlyeq [A(1)]^{-1} \\
\iff \quad & \boldsymbol{v}_1^{\mathrm{T}} [A(\rho^2)]^{-1} \boldsymbol{v} \leqslant \boldsymbol{v}^{\mathrm{T}} [A(1)]^{-1} \boldsymbol{v} \\
\iff \quad & \sigma_T^2(0) - \rho^2 \boldsymbol{v}_1^{\mathrm{T}} [A(\rho^2)]^{-1} \boldsymbol{v} \geqslant \sigma_T^2(0) - \rho^2 \boldsymbol{v}^{\mathrm{T}} [A(1)]^{-1} \boldsymbol{v} \\
\text{i.e.,} \quad & \sigma_T^2(\rho^2) \geqslant \sigma_T^2(0) - \rho^2 \boldsymbol{v}^{\mathrm{T}} [A(1)]^{-1} \boldsymbol{v}
\end{aligned}$$

We write A_1 for $A(1)$. To complete the proof, we use the identity $\boldsymbol{v}_1^{\mathrm{T}} A_1^{-1} \boldsymbol{v}_1 = \sigma_T^2(0) - \sigma_T^2(1)$, which is obtained by by substituting 1 for ρ into (S.7). Further re-grouping of terms leads to the result. □

Remark The expression for $\sigma_T^2(\rho)$ given by (S.7) can also be obtained by repeated conditioning. Consider the following covariance matrix between the query f_*^T, the noisy observations $\boldsymbol{y}_S^S$ at X_S for task S, and the noisy observations $\boldsymbol{y}_T^T$ at X_T for task T,

$$\mathbb{C} \begin{pmatrix} f_*^T \\ \boldsymbol{y}_S^S \\ \boldsymbol{y}_T^T \end{pmatrix} = \begin{pmatrix} k_{**} & \rho(\boldsymbol{k}_{S*}^{\mathrm{x}})^{\mathrm{T}} & (\boldsymbol{k}_{T*}^{\mathrm{x}})^{\mathrm{T}} \\ \rho \boldsymbol{k}_{S*}^{\mathrm{x}} & K_{SS}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I & \rho K_{ST}^{\mathrm{x}} \\ \boldsymbol{k}_{T*}^{\mathrm{x}} & \rho K_{TS}^{\mathrm{x}} & K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I \end{pmatrix}. \tag{S.8}$$

By conditioning on $\boldsymbol{y}_T^T$, we obtain

$$\mathbb{C} \left(\begin{matrix} f_*^T \\ \boldsymbol{y}_S^S \end{matrix} \,\middle|\, \boldsymbol{y}_T^T \right) = \begin{pmatrix} \sigma_T(0) & (\boldsymbol{v}(\rho))^{\mathrm{T}} \\ \boldsymbol{v}(\rho) & A(\rho) \end{pmatrix}. \tag{S.9}$$

Conditioning subsequently on $\boldsymbol{y}_S^S$ gives

$$\sigma_T^2(\rho) \stackrel{\text{def}}{=} \mathbb{C}(f_*^T | \boldsymbol{y}_T^T, \boldsymbol{y}_S^S) = \sigma_T^2(0) - \boldsymbol{v}(\rho)^{\mathrm{T}} [A(\rho)]^{-1} \boldsymbol{v}(\rho). \tag{S.10}$$

To complete, we simply write $\rho \boldsymbol{v}_1$ for $\boldsymbol{v}(\rho)$.

S.2.2 Proof for Proposition 5b

Recall that the exact posterior variance is

$$\sigma_T^2(\rho) = \sigma_T^2(0) - \rho^2 \boldsymbol{v}_1^{\mathrm{T}} \left[A(\rho)\right]^{-1} \boldsymbol{v}_1, \tag{S.11}$$

i.e., (S.7). Denote the lower bound by $\underline{\sigma}_T^2(\rho)$. Then from the proof for Proposition 5a, we have

$$\underline{\sigma}_T^2(\rho) = \sigma_T^2(0) - \rho^2 \boldsymbol{v}_1^{\mathrm{T}} A_1^{-1} \boldsymbol{v}_1. \tag{S.12}$$

Define the gap between the exact posterior variance and its lower bound as

$$\begin{aligned} g(\rho^2) &\stackrel{\text{def}}{=} \sigma_T^2(\rho) - \underline{\sigma}_T^2(\rho) \\ &= -\rho^2 \boldsymbol{v}_1^{\mathrm{T}} [A(\rho)]^{-1} \boldsymbol{v}_1 + \rho^2 \boldsymbol{v}_1^{\mathrm{T}} A_1^{-1} \boldsymbol{v}_1. \end{aligned} \tag{S.13}$$

Ignoring the first term, which is negative, gives

$$\begin{aligned} g(\rho^2) &\leqslant \rho^2 \boldsymbol{v}_1^{\mathrm{T}} A_1^{-1} \boldsymbol{v}_1 \\ &= \rho^2 \left[\sigma_T^2(0) - \sigma_T^2(1)\right]. \end{aligned}$$ □

S.2.3 Proof for Proposition 5c

We rewrite the gap (S.13) between the exact posterior variance and its lower bound as

$$g(\rho^2) = \rho^2 \boldsymbol{v}_1^{\mathrm{T}} \left[A_1^{-1} - A(\rho))^{-1}\right] \boldsymbol{v}_1. \tag{S.14}$$

Next, express $A(\rho)$ as

$$A(\rho) = A_1 + (1-\rho^2) K_{ST}^{\mathrm{x}} (K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I)^{-1} K_{TS}^{\mathrm{x}}, \tag{S.15}$$

so that we can use the Woodbury identity to expand its inverse in S.14, and write

$$g(\rho^2) = \boldsymbol{v}_1^{\mathrm{T}} A_1^{-1} K_{ST}^{\mathrm{x}} [B(\rho^2)]^{-1} K_{TS}^{\mathrm{x}} A_1^{-1} \boldsymbol{v}_1, \tag{S.16}$$

where

$$B(\rho^2) \stackrel{\text{def}}{=} D(\rho^2) + \frac{1}{\rho^2} K_{TS}^{\mathrm{x}} A_1^{-1} K_{ST}^{\mathrm{x}} \tag{S.17}$$

$$D(\rho^2) \stackrel{\text{def}}{=} \frac{1}{\rho^2(1-\rho^2)} (K_{TT}^{\mathrm{x}} + \sigma_{\mathrm{n}}^2 I). \tag{S.18}$$

Notice that the dependence of $g(\rho^2)$ on ρ^2 is only through $B(\rho^2)$. We differentiate $g(\rho^2)$ with respect to ρ^2:

$$\frac{\mathrm{d}g}{\mathrm{d}\rho^2} = \boldsymbol{v}_1^{\mathrm{T}} A_1^{-1} K_{ST}^{\mathrm{x}} [B(\rho^2)]^{-1} C(\rho^2) [B(\rho^2)]^{-1} K_{TS}^{\mathrm{x}} A_1^{-1} \boldsymbol{v}_1, \tag{S.19}$$

where

$$C(\rho^2) \stackrel{\text{def}}{=} -\frac{\mathrm{d}B}{\mathrm{d}\rho^2} = \frac{1-2\rho^2}{\rho^2(1-\rho^2)} D + \frac{1}{\rho^4} K_{TS}^{\mathrm{x}} A_1^{-1} K_{ST}^{\mathrm{x}} = \frac{1}{\rho^2} B - \frac{1}{1-\rho^2} D. \tag{S.20}$$

Substituting the last expression for $C(\rho^2)$ back into (S.19) gives

$$\frac{\mathrm{d}g}{\mathrm{d}\rho^2} = \frac{1}{\rho^2} g - \frac{1}{1-\rho^2} h, \tag{S.21}$$

where

$$h(\rho^2) \stackrel{\text{def}}{=} \boldsymbol{v}_1^{\mathrm{T}} A_1^{-1} K_{ST}^{\mathrm{x}} [B(\rho^2)]^{-1} D(\rho^2) [B(\rho^2)]^{-1} K_{TS}^{\mathrm{x}} A_1^{-1} \boldsymbol{v}_1. \tag{S.22}$$

Put inequality $D(\rho^2) \preccurlyeq B(\rho^2)$ into $h(\rho^2)$ to give

$$\begin{aligned} h(\rho^2) &\leqslant \boldsymbol{v}_1^{\mathrm{T}} A_1^{-1} K_{ST}^{\mathrm{x}} [B(\rho^2)]^{-1} B(\rho^2) [B(\rho^2)]^{-1} K_{TS}^{\mathrm{x}} A_1^{-1} \boldsymbol{v}_1 \\ &= \boldsymbol{v}_1^{\mathrm{T}} A_1^{-1} K_{ST}^{\mathrm{x}} [B(\rho^2)]^{-1} K_{TS}^{\mathrm{x}} A_1^{-1} \boldsymbol{v}_1 \\ &= g(\rho^2). \end{aligned} \tag{S.23}$$

Putting the above inequality into (S.21) leads to

$$\frac{\mathrm{d}g}{\mathrm{d}\rho^2} \geqslant f(\rho^2)\, g(\rho^2), \qquad \text{where} \quad f(\rho^2) \stackrel{\text{def}}{=} \frac{1}{\rho^2} - \frac{1}{1-\rho^2}. \tag{S.24}$$

Since $\underline{\sigma}_T^2(\rho)$ is a lower bound, $g(\rho^2) \geqslant 0$ (also see the quadratic form in (S.16)). In addition, the multiplicative factor $f(\rho^2)$ is positive for $\rho^2 \in [0, 1/2[$, zero at $\rho^2 = 1/2$, and negative for $\rho^2 \in]1/2, 1]$. Thus $\mathrm{d}g/\mathrm{d}\rho^2 > 0$ for $\rho^2 \in [0, 1/2[$. Therefore g is monotonically increasing within $\rho^2 \in [0, 1/2[$, and its maximum value must be at $\hat{\rho}^2 \geqslant 1/2$. □

S.3 Proof for Proposition 7

Recall that we seek an upper bound $\bar{\sigma}_{\mathrm{n}}^2$ for $\tilde{\sigma}_{\mathrm{n}}^2$ such that $\Delta(\rho, \sigma_{\mathrm{n}}^2, \bar{\sigma}_{\mathrm{n}}^2) \leqslant 0$ for all test locations, where

$$\Delta(\rho, \sigma_{\mathrm{n}}^2, s^2) \stackrel{\mathrm{def}}{=} (\boldsymbol{k}_*^{\mathrm{x}})^{\mathrm{T}} \left[(\Sigma(1, \sigma_{\mathrm{n}}^2, s^2))^{-1} - (\Sigma(\rho, \sigma_{\mathrm{n}}^2, \sigma_{\mathrm{n}}^2))^{-1} \right] \boldsymbol{k}_*^{\mathrm{x}}. \tag{S.25}$$

For the derived bound to be applicable in general, it is necessary that the condition "for all test locations" be equivalent to the condition "for all $\boldsymbol{k}_*^{\mathrm{x}} \in \mathbb{R}^n$", which is easier to handle; we shall remark on this later. Thus, we start from the requirement that $\Delta(\rho, \sigma_{\mathrm{n}}^2, \bar{\sigma}_{\mathrm{n}}^2) \leqslant 0$ for all $\boldsymbol{k}_*^{\mathrm{x}} \in \mathbb{R}^n$:

$$\Delta(\rho, \sigma_{\mathrm{n}}^2, \bar{\sigma}_{\mathrm{n}}^2) \leqslant 0 \quad \forall \boldsymbol{k}_*^{\mathrm{x}} \in \mathbb{R}^n \tag{S.26a}$$

$$\iff \quad (\Sigma(1, \sigma_{\mathrm{n}}^2, \bar{\sigma}_{\mathrm{n}}^2))^{-1} - (\Sigma(\rho, \sigma_{\mathrm{n}}^2, \sigma_{\mathrm{n}}^2))^{-1} \preccurlyeq 0 \tag{S.26b}$$

$$\iff \quad (\Sigma(1, \sigma_{\mathrm{n}}^2, \bar{\sigma}_{\mathrm{n}}^2))^{-1} \preccurlyeq (\Sigma(\rho, \sigma_{\mathrm{n}}^2, \sigma_{\mathrm{n}}^2))^{-1} \tag{S.26c}$$

$$\iff \quad \Sigma(1, \sigma_{\mathrm{n}}^2, \bar{\sigma}_{\mathrm{n}}^2) \succcurlyeq \Sigma(\rho, \sigma_{\mathrm{n}}^2, \sigma_{\mathrm{n}}^2) \tag{S.26d}$$

$$\iff \quad \begin{pmatrix} K_{TT}^{\mathrm{x}} & K_{TS}^{\mathrm{x}} \\ K_{ST}^{\mathrm{x}} & K_{SS}^{\mathrm{x}} \end{pmatrix} + \begin{pmatrix} \sigma_{\mathrm{n}}^2 I & 0 \\ 0 & \bar{\sigma}_{\mathrm{n}}^2 I \end{pmatrix} \succcurlyeq \begin{pmatrix} K_{TT}^{\mathrm{x}} & K_{TS}^{\mathrm{x}} \\ K_{ST}^{\mathrm{x}} & \rho^{-2} K_{SS}^{\mathrm{x}} \end{pmatrix} + \begin{pmatrix} \sigma_{\mathrm{n}}^2 I & 0 \\ 0 & \rho^{-2}\sigma_{\mathrm{n}}^2 I \end{pmatrix} \tag{S.26e}$$

$$\iff \quad \begin{pmatrix} 0 & 0 \\ 0 & \beta K_{SS}^{\mathrm{x}} \end{pmatrix} \preccurlyeq \begin{pmatrix} 0 & 0 \\ 0 & (\bar{\sigma}_{\mathrm{n}}^2 - \rho^{-2}\sigma_{\mathrm{n}}^2) I \end{pmatrix} \tag{S.26f}$$

$$\iff \quad \beta K_{SS}^{\mathrm{x}} \preccurlyeq (\bar{\sigma}_{\mathrm{n}}^2 - \rho^{-2}\sigma_{\mathrm{n}}^2) I \tag{S.26g}$$

$$\iff \quad K_{SS}^{\mathrm{x}} \preccurlyeq \frac{\bar{\sigma}_{\mathrm{n}}^2 - \rho^{-2}\sigma_{\mathrm{n}}^2}{\beta} I \tag{S.26h}$$

$$\iff \quad \bar{\lambda} \leqslant \frac{\bar{\sigma}_{\mathrm{n}}^2 - \rho^{-2}\sigma_{\mathrm{n}}^2}{\beta} \tag{S.26i}$$

$$\iff \quad \bar{\sigma}_{\mathrm{n}}^2 \geqslant \beta\bar{\lambda} + \rho^{-2}\sigma_{\mathrm{n}}^2 = \beta(\bar{\lambda} + \sigma_{\mathrm{n}}^2) + \sigma_{\mathrm{n}}^2 \tag{S.26j}$$

Therefore we have the minimum of the upper bound is

$$\bar{\bar{\sigma}}_{\mathrm{n}}^2 \stackrel{\mathrm{def}}{=} \beta(\bar{\lambda} + \sigma_{\mathrm{n}}^2). \tag{S.27}$$

The tightness of the bound is evident from the construction of $\bar{\bar{\sigma}}_{\mathrm{n}}^2$. □

Remark For the bound to hold in general, we have claimed in the above proof that the condition "for all test locations" must be equivalent to the condition "for all $\boldsymbol{k}_*^{\mathrm{x}} \in \mathbb{R}^n$". To show this, we shall give a particular example that demands this equivalence. Consider input domain $[-1, 1]$ and $k^{\mathrm{x}}(x, x') = xx'$. We fix x_*. If the observed locations are densely located on $[-1, 1] \setminus \{x_*\}$, then the entries in $\boldsymbol{k}_*^{\mathrm{x}}$ are densely located on $[-x_*, x_*] \setminus \{x_*^2\}$. Since scaling $\boldsymbol{k}_*^{\mathrm{x}}$ does not effect the inequality $\Delta(\rho, \sigma_{\mathrm{n}}^2, \bar{\sigma}_{\mathrm{n}}^2) \leqslant 0$, we have, equivalently, $\boldsymbol{k}_*^{\mathrm{x}} \in \mathbb{R}^n$.

S.4 Proof for Theorem 12

Theorem 12 upon which Proposition 13 is built depends on the Lemma 4 in Ferrari Trecate et al. [2]. Before we examine a variant of that lemma, some additional notations are needed. Let $\boldsymbol{y} \stackrel{\mathrm{def}}{=} (y_1 \dots y_n)^{\mathrm{T}}$ be the n values observed at the set of data locations $X \stackrel{\mathrm{def}}{=} \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$. In addition to the matrix Φ defined in Theorem 12, we introduce an (infinite) vector function $\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}) \stackrel{\mathrm{def}}{=} (\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \dots)^{\mathrm{T}}$, and an (infinite) diagonal matrix Λ_κ with κ_i on the diagonals.[1] Given the data, consider the space of functions

$$\mathcal{H}_0 \stackrel{\mathrm{def}}{=} \{\boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}) L \boldsymbol{y} \mid L \in \mathbb{R}^{\infty \times n}\}.$$

The aim is to use a function g from $\mathcal{H}_0$ to estimate the true function $f^\star$ drawn from the GP with covariance function $k^{\mathrm{x}}(\cdot, \cdot)$. The quality of this estimation may be evaluated using the following variant of Lemma 4 from [2]:

[1] In [2], where the focus is on finite-dimensional approximation of GP, only m major eigenvalues and eigenfunctions are used. However, our focus is to obtain an upper bound on the learning curve, and so we follow [3] and let $m \to \infty$.

Lemma S.2. *(cf. [2, Lemma 4]) The generalization error of a function $g \in \mathcal{H}_0$ is*

$$\epsilon(g \in \mathcal{H}_0, X) \stackrel{\text{def}}{=} \left\langle (f^\star(\boldsymbol{x}) - g(\boldsymbol{x}))^2 \right\rangle_{f^\star, \boldsymbol{y}, \boldsymbol{x}}$$
$$= \sum_{i=1}^{\infty} \kappa_i + \operatorname{tr}\left(L \left\langle \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} \right\rangle_{\boldsymbol{y}} L^{\mathrm{T}} \right) - 2 \operatorname{tr}\left(L \left\langle \langle \boldsymbol{y} f^\star(\boldsymbol{x}) \rangle_{f^\star, \boldsymbol{y}} \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}) \right\rangle_{\boldsymbol{x}} \right)$$

Proof.

$$\left\langle (f^\star(\boldsymbol{x}) - g(\boldsymbol{x}))^2 \right\rangle_{f^\star, \boldsymbol{y}, \boldsymbol{x}}$$
$$= \left\langle \left(f^\star(\boldsymbol{x}) - \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}) L \boldsymbol{y} \right)^2 \right\rangle_{f^\star, \boldsymbol{y}, \boldsymbol{x}}$$
$$= \langle f^\star(\boldsymbol{x}) f^\star(\boldsymbol{x}) \rangle_{f^\star, \boldsymbol{x}} + \left\langle \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}) L \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} L^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}) \right\rangle_{\boldsymbol{y}, \boldsymbol{x}} - 2 \left\langle \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}) L \langle \boldsymbol{y} f^\star(\boldsymbol{x}) \rangle_{f^\star, \boldsymbol{y}} \right\rangle_{\boldsymbol{x}}$$
$$= \langle k^{\mathrm{x}}(\boldsymbol{x}, \boldsymbol{x}) \rangle_{\boldsymbol{x}} + \operatorname{tr}\left(L \left\langle \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} \right\rangle_{\boldsymbol{y}} L^{\mathrm{T}} \left\langle \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}) \right\rangle_{\boldsymbol{x}} \right) - 2 \operatorname{tr}\left(L \left\langle \langle \boldsymbol{y} f^\star(\boldsymbol{x}) \rangle_{f^\star, \boldsymbol{y}} \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}) \right\rangle_{\boldsymbol{x}} \right)$$
$$= \sum_{i=1}^{\infty} \kappa_i + \operatorname{tr}\left(L \left\langle \boldsymbol{y}\boldsymbol{y}^{\mathrm{T}} \right\rangle_{\boldsymbol{y}} L^{\mathrm{T}} \right) - 2 \operatorname{tr}\left(L \left\langle \langle \boldsymbol{y} f^\star(\boldsymbol{x}) \rangle_{f^\star, \boldsymbol{y}} \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}) \right\rangle_{\boldsymbol{x}} \right)$$

We have used $\left\langle \boldsymbol{\phi}(\boldsymbol{x}) \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}) \right\rangle_{\boldsymbol{x}} = I$ for the last expression. □

To proceed further, it is necessary to specify how $\boldsymbol{y}$ is obtained. For the single-task GP with isotropic noise, and under correct prior specification, each entry in $\boldsymbol{y}$ is generated via

$$y(\boldsymbol{x}) \sim \mathcal{N}(f^\star(\boldsymbol{x}), \sigma_{\mathrm{n}}^2). \tag{S.28}$$

This is the setting considered in [2], where it is also shown that minimizing the generalization error with respect to L leads to the GP mean predictor [2, Theorem 5]; see also [4], and [5, Proposition V.1]. The single-task FWO bound on the learning curve of the GP is obtained by minimizing $\langle \epsilon(g \in \mathcal{H}_0, X) \rangle_X$ with respect to g within only the sub-space of functions

$$\mathcal{H}_1 \stackrel{\text{def}}{=} \{ \boldsymbol{\phi}^{\mathrm{T}}(\boldsymbol{x}) D \Phi^{\mathrm{T}} \boldsymbol{y} \mid D \text{ is a diagonal matrix} \}.$$

This results in an upper bound on the learning curve. This is because $\mathcal{H}_1 \subseteq \mathcal{H}_0$, so that minimizing $\langle \epsilon(g \in \mathcal{H}_0, X) \rangle_X$ naturally gives predictors that cannot outperform the GP mean predictor. The form of functions in $\mathcal{H}_1$ is motivated by Projected Bayes Regression [2, Definition 1], wherein the L in $\mathcal{H}_0$ is constrained to be $M\Phi^{\mathrm{T}}$ for a square matrix M.

Theorem 12, our variant of Ferrari Trecate et al.'s Theorem 6 for correlated noise γ^2, for example as defined by (14), uses

$$y(\boldsymbol{x}) \sim \mathcal{GP}(f^\star(\boldsymbol{x}), \gamma^2(\boldsymbol{x}, \boldsymbol{x}')) \tag{S.29}$$

instead of (S.28). Generating $\boldsymbol{y}$ in this manner and further restricting g to be from $\mathcal{H}_1$ leads to the following generalization error using Lemma S.2:

$$\epsilon(g \in \mathcal{H}_1, X) = \sum_{i=1}^{\infty} \kappa_i + \operatorname{tr}\left(D \Phi^{\mathrm{T}} H \Phi D \right) - 2 \operatorname{tr}\left(D \Phi^{\mathrm{T}} \Phi \Lambda_\kappa \right), \tag{S.30}$$

where

$$H_{ij} \stackrel{\text{def}}{=} k^{\mathrm{x}}(\boldsymbol{x}_i, \boldsymbol{x}_j) + \gamma^2(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{S.31}$$

The right of (S.30) is quadratic in the diagonal entries of D. Minimizing $\langle \epsilon(g \in \mathcal{H}_1, X) \rangle_X$ with respect to D results in Theorem 12.

S.5 Proof for Proposition 13

We prove the expression for c_i given by (15), which forms the crux of Proposition 13. In order to present the proof, some notations are required. Recall that the cardinality of these sets are $|X| = n$,

$|X_S| = n_S$ and $|X_T| = n_T$, and that $\pi_S \stackrel{\text{def}}{=} n_S/n$. We partition the index set $\mathcal{I} \stackrel{\text{def}}{=} \{1 \dots n\}$ into $\mathcal{I}_T \stackrel{\text{def}}{=} \{1 \dots n_T\}$, and $\mathcal{I}_S \stackrel{\text{def}}{=} \{(n_T+1) \dots n\}$. We enumerate and order the elements of X so that $X = \{\boldsymbol{x}_i\}_{i=1}^n$, $X_T = \{\boldsymbol{x}_i \mid i \in \mathcal{I}_T\}$, and $X_S = \{\boldsymbol{x}_i \mid i \in \mathcal{I}_S\}$. We shall represent by $\langle \cdots \rangle_X$ the expectation over the set X, and write $p(X)\mathrm{d}X \stackrel{\text{def}}{=} \prod_{i=1}^n p(\boldsymbol{x}_i)\mathrm{d}\boldsymbol{x}_i$, where the distributions over the $\boldsymbol{x}_i$s are identical; expressions $\langle \cdots \rangle_{X_S}$ and $\langle \cdots \rangle_{X_T}$, and $p(X_S)\mathrm{d}X_S$ and $p(X_T)\mathrm{d}X_T$ have similar meanings.

Recall the definition of eigenvalues and eigenfunctions using the integral equation and the orthogonality of eigenfunctions:

$$\int k^{\mathrm{x}}(\boldsymbol{x}, \boldsymbol{x}')\phi_i(\boldsymbol{x}')p(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}' = \kappa_i\phi_i(\boldsymbol{x}) \qquad \int \phi_i(\boldsymbol{x})\phi_i(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1. \tag{S.32}$$

Using the these two equalities, we can show that

$$\begin{aligned}
\int k^{\mathrm{x}}(\boldsymbol{x}, \boldsymbol{x}')\phi_i(\boldsymbol{x})\phi_i(\boldsymbol{x}')p(\boldsymbol{x})p(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' &= \int \left(\int k^{\mathrm{x}}(\boldsymbol{x}, \boldsymbol{x}')\phi_i(\boldsymbol{x}')p(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}' \right) \phi_i(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int \left(\kappa_i\phi_i(\boldsymbol{x})\right) \phi_i(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \kappa_i \int \phi_i(\boldsymbol{x})\phi_i(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \kappa_i
\end{aligned} \tag{S.33}$$

The next two equalities will be used in the main part of the proof:

$$\begin{aligned}
\left\langle \sum_{p,q \in \mathcal{I}} \delta_{pq}\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q) \right\rangle_X &= \int \sum_{p=1}^n \phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_p)p(X)\mathrm{d}X \\
&= \sum_{p=1}^n \int \phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_p)p(\boldsymbol{x}_p)\mathrm{d}\boldsymbol{x}_p \\
&= n \int \left[\phi_i(\boldsymbol{x})\right]^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= n
\end{aligned} \tag{S.34}$$

$$\begin{aligned}
\left\langle \sum_{p,q \in \mathcal{I}} k^{\mathrm{x}}(\boldsymbol{x}_p, \boldsymbol{x}_q)\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q) \right\rangle_X &= \int \sum_{p,q \in \mathcal{I}} k^{\mathrm{x}}(\boldsymbol{x}_p, \boldsymbol{x}_q)\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q)p(X)\mathrm{d}X \\
&= \sum_{\substack{p,q \in \mathcal{I} \\ p \neq q}} \int k^{\mathrm{x}}(\boldsymbol{x}_p, \boldsymbol{x}_q)\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q)p(\boldsymbol{x}_p)p(\boldsymbol{x}_q)\mathrm{d}\boldsymbol{x}_p\mathrm{d}\boldsymbol{x}_q \\
&\qquad + \sum_{p=1}^n \int k^{\mathrm{x}}(\boldsymbol{x}_p, \boldsymbol{x}_p)\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_p)p(\boldsymbol{x}_p)\mathrm{d}\boldsymbol{x}_p \\
&= \sum_{\substack{p,q \in \mathcal{I} \\ p \neq q}} \kappa_i + \sum_{p=1}^n \int k^{\mathrm{x}}(\boldsymbol{x}, \boldsymbol{x}) \left[\phi_i(\boldsymbol{x})\right]^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \qquad (*) \\
&= n(n-1)\kappa_i + n \int k^{\mathrm{x}}(\boldsymbol{x}, \boldsymbol{x}) \left[\phi_i(\boldsymbol{x})\right]^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x},
\end{aligned} \tag{S.35}$$

where (S.33) is used in getting to $(*)$. By similar arguments we can show equivalent results when the summations and expectations are taken only over data points in X_S and X_T.

We now turn to the main part of the proof. The observation noise (co)variance for our upper bound can be expressed as

$$\gamma^2(\boldsymbol{x}_p, \boldsymbol{x}_q) \stackrel{\text{def}}{=} \begin{cases} \delta_{pq}\sigma_{\mathrm{n}}^2 & \text{if } p \in \mathcal{I}_T \text{ and } q \in \mathcal{I}_T \\ \beta k^{\mathrm{x}}(\boldsymbol{x}_p, \boldsymbol{x}_q) + \rho^{-2}\delta_{pq}\sigma_{\mathrm{n}}^2 & \text{if } p \in \mathcal{I}_S \text{ and } q \in \mathcal{I}_S \\ 0 & \text{otherwise,} \end{cases} \tag{S.36}$$

where δ_{pq} is the Kronecker delta function. In the definition of γ^2, the first case is when both input locations are for task T, and the second case is when both input locations are for task S. It follows that

$$\begin{aligned}
&\left\langle \sum_{p,q\in\mathcal{I}} \gamma^2(\boldsymbol{x}_p,\boldsymbol{x}_q)\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q) \right\rangle_X \\
&= \left\langle \sum_{p,q\in\mathcal{I}_T} \delta_{pq}\sigma_{\mathrm{n}}^2\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q) \right\rangle_{X_T} + \left\langle \sum_{p,q\in\mathcal{I}_S} \left(\beta k^{\mathrm{x}}(\boldsymbol{x}_p,\boldsymbol{x}_q) + \rho^{-2}\delta_{pq}\sigma_{\mathrm{n}}^2\right)\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q) \right\rangle_{X_S} \\
&= \sigma_{\mathrm{n}}^2\left\langle \sum_{p,q\in\mathcal{I}_T} \delta_{pq}\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q) \right\rangle_{X_T} + \beta\left\langle \sum_{p,q\in\mathcal{I}_S} k^{\mathrm{x}}(\boldsymbol{x}_p,\boldsymbol{x}_q)\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q) \right\rangle_{X_S} \\
&\qquad\qquad + \rho^{-2}\sigma_{\mathrm{n}}^2\left\langle \sum_{p,q\in\mathcal{I}_S} \delta_{pq}\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q) \right\rangle_{X_S} \\
&= n_T\sigma_{\mathrm{n}}^2 + \beta\left[n_S(n_S-1)\kappa_i + \beta n_S\int k^{\mathrm{x}}(\boldsymbol{x},\boldsymbol{x})\left[\phi_i(\boldsymbol{x})\right]^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right] + \rho^{-2}n_S\sigma_{\mathrm{n}}^2 \qquad (*) \\
&= \beta n_S(n_S-1)\kappa_i + \beta n_S\int k^{\mathrm{x}}(\boldsymbol{x},\boldsymbol{x})\left[\phi_i(\boldsymbol{x})\right]^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + n\sigma_{\mathrm{n}}^2 + \beta n_S\sigma_{\mathrm{n}}^2, \qquad \text{(S.37)}
\end{aligned}$$

where the X_S and X_T equivalents of (S.34) and (S.35) are applied to get $(*)$. We are now ready to apply Theorem 12 to obtain c_i:

$$\begin{aligned}
c_i &= \frac{1}{n}\left\langle \left(\Phi^{\mathrm{T}}H\Phi\right)_{ii}\right\rangle_X \\
&= \frac{1}{n}\left\langle \sum_{p,q\in\mathcal{I}} \left(k^{\mathrm{x}}(\boldsymbol{x}_p,\boldsymbol{x}_q) + \gamma^2(\boldsymbol{x}_p,\boldsymbol{x}_q)\right)\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q) \right\rangle_X \\
&= \frac{1}{n}\left\langle \sum_{p,q\in\mathcal{I}} k^{\mathrm{x}}(\boldsymbol{x}_p,\boldsymbol{x}_q)\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q) \right\rangle_X + \frac{1}{n}\left\langle \sum_{p,q\in\mathcal{I}} \gamma^2(\boldsymbol{x}_p,\boldsymbol{x}_q)\phi_i(\boldsymbol{x}_p)\phi_i(\boldsymbol{x}_q) \right\rangle_X \\
&= (n-1)\kappa_i + \int k^{\mathrm{x}}(\boldsymbol{x},\boldsymbol{x})\left[\phi_i(\boldsymbol{x})\right]^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \qquad (*) \\
&\qquad\qquad + \beta\pi_S(n_S-1)\kappa_i + \beta\pi_S\int k^{\mathrm{x}}(\boldsymbol{x},\boldsymbol{x})\left[\phi_i(\boldsymbol{x})\right]^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \sigma_{\mathrm{n}}^2 + \beta\pi_S\sigma_{\mathrm{n}}^2 \\
&= \left[(n-1) + \beta n_S(n_S-1)\right]\kappa_i + (1+\beta\pi_S)\int k^{\mathrm{x}}(\boldsymbol{x},\boldsymbol{x})\left[\phi_i(\boldsymbol{x})\right]^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + (1+\beta\pi_S)\sigma_{\mathrm{n}}^2 \\
&= \left[(1+\beta\pi_S^2)n - (1+\beta\pi_S)\right]\kappa_i + (1+\beta\pi_S)\int k^{\mathrm{x}}(\boldsymbol{x},\boldsymbol{x})\left[\phi_i(\boldsymbol{x})\right]^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + (1+\beta\pi_S)\sigma_{\mathrm{n}}^2
\end{aligned}$$

where (S.35) and (S.37) are used to obtain $(*)$. □

Remark. Although we do not have the proof by Ferrari Trecate et al. [2] for their upper bound on the learning curve for single-task GP with isotropic noise, it is conceivable that some variation of the above proof has been used by them.

S.6 Simulations of learning curve, details

In this section, we give additional details for our simulations of the learning curve.

S.6.1 Continuation of $\epsilon_T^{\mathrm{avg}}$ in $\pi_S n$

Recall that $\epsilon_T^{\mathrm{avg}}(\rho,\sigma_{\mathrm{n}}^2,\pi_S,n)$ is only defined for values of π_S and n such that $\pi_S n = n_S \in \mathbb{N}_0$. In our simulations, however, we extend the domain to allow $\pi_S n \in \mathbb{R}$, so that smooth curves can be plotted. For the theoretical bounds given by Propositions 11 and 13, this is done by simply using

the respective expressions verbatim. For the experimental bounds, which require sampling over $X \stackrel{\text{def}}{=} X_T \cup X_S$, this is achieved in the manner described next.

For a given π_S, we sample the sizes of the training sets X_T and X_S to satisfy

$$\lfloor \pi_S n \rfloor \leqslant n_S \leqslant \lceil \pi_S n \rceil \qquad \text{and} \qquad \langle n_S \rangle = \pi_S n, \tag{S.38}$$

where the expectation is taken over simulation runs. The first condition ensures that, within each simulation run, the size n_S of X_S is $\pi_S n$ whenever the latter is an integer. The second condition ensures that the ratio n_S/n is consistent with π_S when averaged over multiple simulation runs. For each simulation run, the training set is constructed sequentially by randomly drawing additional training locations. For each new location, we determine its task by using Algorithm 1.

Algorithm 1 Decide the task for a new input

Require: Ratio π_S, required cardinality n of X, and and previous cardinality n_S^{old} of X_S.
1: **if** $n_S^{\text{old}} < \lfloor \pi_S n \rfloor$ **then**
2: new input is for task S
3: **else if** $n_S^{\text{old}} = \lceil \pi_S n \rceil$ **then**
4: new input is for task T
5: **else**
6: new input is for task S with probability $(\pi_S n - \lfloor \pi_S n \rfloor)$
{or, equivalently, for task T with probability $1 - (\pi_S n - \lfloor \pi_S n \rfloor)$}
7: **end if**

S.6.2 Analytical averaging over test locations

The simulation study in section 5.3 uses the squared exponential (SE) covariance function with normally distributed inputs. As claimed in section 3.4, the expectation over test locations can be done analytically to obtain the generalization error ϵ_T exactly. In order to do this, we need to be able to compute

$$M_{pq} \stackrel{\text{def}}{=} \int k^{\text{x}}(\boldsymbol{x}_p, \boldsymbol{x}_*)\, k^{\text{x}}(\boldsymbol{x}_q, \boldsymbol{x}_*)\, p(\boldsymbol{x}_*) \mathrm{d}\boldsymbol{x}_* = \sum_{i=1}^{\infty} \kappa_i^2 \phi_i(\boldsymbol{x}_p) \phi_i(\boldsymbol{x}_q) \tag{S.39}$$

for a fixed pair of input locations $(\boldsymbol{x}_p, \boldsymbol{x}_q)$. In the one-dimensional case with the SE covariance function and normally distributed inputs, i.e.,

$$k^{\text{x}}(x, x') \stackrel{\text{def}}{=} \exp -\frac{(x - x')^2}{2l^2} \qquad \text{and} \qquad p(x) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}\sigma_x} \exp -\frac{x^2}{2\sigma_x^2}, \tag{S.40}$$

we can use the integral expression for M_{pq} to obtain

$$M_{pq} = \frac{l}{\sqrt{2\sigma_x^2 + l^2}} \exp -\frac{\sigma_x^2 (x_p - x_q)^2 + l^2 (x_p^2 + x_q^2)}{2l^2 (2\sigma_x^2 + l^2)}. \tag{S.41}$$

This can be easily generalized to input spaces of higher-dimensions. Note that the infinite sum expression for M_{pq} is not useful in this case, even though analytic expressions for the eigenfunctions are available [6]. This is because the eigenfunctions corresponding to the smaller eigenvalues exhibit larger oscillations around zero in regions where $p(x)$ is low, so that it is hard to determine when to truncate the infinite sum for certain pairs of (x_p, x_q).

References

[1] Simo Puntanen and George P. H. Styan. Historical introduction: Issai Schur and the early development of the Schur complement. In Fuzhen Zhang, editor, *The Schur complement and its applications*, Numerical Methods and Algorithms, pages 1–16. Springer, 2005.

[2] Giancarlo Ferrari Trecate, Christopher K. I. Williams, and Manfred Opper. Finite-dimensional approximation of Gaussian processes. In Michael J. Kearns, Sara A. Solla, and David A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 218–224. The MIT Press, 1999.

[3] Peter Sollich and Anason Halees. Learning curves for Gaussian process regression: Approximations and bounds. *Neural Computation*, 14(6):1393–1428, 2002.

[4] George S. Kimeldorf and Grace Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2): 495–502, 1970.

[5] Klaus Ritter. *Average-Case Analysis of Numerical Problems*, volume 1733 of *Lecture Notes in Mathematics*. Springer, 2000.

[6] Huaiyu Zhu, Christopher K. I. Williams, Richard Rohwer, and Michal Morciniec. Gaussian regression and optimal finite dimensional linear models. In Christopher M. Bishop, editor, *Neural Networks and Machine Learning*, volume 168 of *NATO ASI Series F: Computer and Systems Sciences*, pages 167–184. Springer-Verlag, Berlin, 1998.