Validation of Machine Learning Prediction Models

Luc Pronzato, Maria Joao Rendas

The New England Journal of Statistics in Data Science Volume 0, 1–17 (2020)
DOI: https://doi.org/0000

Validation of Machine Learning Prediction Models

Luc PRONZATO* and Maria-João RENDAS

Abstract

We address the estimation of the Integrated Squared Error (ISE) of a predictor $\eta(x)$ of an unknown function f learned using data acquired on a given design $\mathbf{X}_n$. We consider ISE estimators that are weighted averages of the residuals of the predictor $\eta(x)$ on a set of selected points $\mathbf{Z}_m$. We show that, under a stochastic model for f, minimisation of the mean squared error of these ISE estimators is equivalent to minimisation of a Maximum Mean Discrepancy (MMD) for a non-stationary kernel that is adapted to the geometry of $\mathbf{X}_n$. Sequential Bayesian quadrature then yields sequences of nested validation designs that minimise, at each step of the construction, the relevant MMD. The optimal ISE estimate can be written in terms of the integral of a linear reconstruction, for the assumed model, of the square of the interpolator residuals over the domain of f. We present an extensive set of numerical experiments which demonstrate the good performance and robustness of the proposed solution. Moreover, we show that the validation designs obtained are space-filling continuations of $\mathbf{X}_n$, and that correct weighting of the observed interpolator residuals is more important than the precise configuration $\mathbf{Z}_m$ of the points at which they are observed.

KEYWORDS AND PHRASES: Model Validation, Bayesian Quadrature, Maximum Mean Discrepancy, Experimental Design.

1. INTRODUCTION AND MOTIVATION

This paper proposes a methodology to estimate the quality of an interpolator learned on a given experimental design. More precisely, we suppose that data gathered on the points of an experimental design $\mathbf{X}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ with n points in a compact set[1] $\mathscr{X}$ has been used to build a predictor of the value of the function $f : \mathscr{X} \to \mathbb{R}$ that produced the collected samples.

We denote by $\mathbf{y}_n = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^\top$ the vector collecting the n evaluations of f at the design points $\mathbf{x}_i$, by $\mathscr{F}_n = (\mathbf{X}_n, \mathbf{y}_n)$ the learning dataset, and by $\eta_{\mathscr{F}_n}(\mathbf{x})$ the resulting prediction of $f(\mathbf{x})$. The quality of $\eta_{\mathscr{F}_n}$ is assessed through a widely used measure of the precision of interpolators, the Integrated Squared Error (ISE):

$$\mathsf{ISE}(\eta_{\mathscr{F}_n}) = \int_{\mathscr{X}} \left[\eta_{\mathscr{F}_n}(\mathbf{x}) - f(\mathbf{x})\right]^2 \mu(\mathrm{d}\mathbf{x})\,. \tag{1.1}$$

In the definition above the (user-defined) measure μ enables penalization of the interpolation errors over regions of $\mathscr{X}$ which are considered to be of particular importance. We stress that we consider that the experimental design $\mathbf{X}_n$ – also referred to as the "learning design" – is given, making no assumptions on how it is has been chosen.

Estimation of the integral (1.1) must necessarily resort to the evaluation of the prediction error $\varepsilon(\mathbf{x}) = \eta_{\mathscr{F}_n}(\mathbf{x}) - f(\mathbf{x})$ over only a finite set of points $\mathbf{Z}_m = \{\mathbf{z}_1, \dots \mathbf{z}_m\} \subset \mathscr{X}$, which we designate by "validation design". The integral is then approximated by replacing μ by a point mass measure $\zeta = \zeta(\mathbf{w}, \mathbf{Z}_m) = \sum_i \mathbf{w}_i \delta_{\mathbf{z}_i}$ supported on $\mathbf{Z}_m$ only. We generically refer to ζ as the validation measure, using the notation ζ_m to make explicit the dependency on the size of the validation set. Although ζ is not necessarily the uniform distribution supported on $\mathbf{Z}_m$, and with a slight abuse of terminology, we refer to the corresponding ISE estimators

$$\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n}; \zeta) = \sum_{i=1}^{m} \mathbf{w}_i \left[\eta_{\mathscr{F}_n}(\mathbf{z}_i) - f(\mathbf{z}_i)\right]^2 \,, \tag{1.2}$$

as *empirical* ISE estimators.

We address the choice of the validation measure ζ – both of the *validation design* $\mathbf{Z}_m$ and of the *validation weights* $\mathbf{w}$ – and investigate the properties of the resulting estimates $\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n}; \zeta)$ given by (1.2). The algorithms presented are iterative, defining increasing sequences of nested validation designs $\mathbf{Z}_m \subset \mathbf{Z}_{m+1} \subset \mathbf{Z}_{m+2} \subset \cdots$ such that the performance of $\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n}; \zeta)$ improves as m increases. A preliminary version of this work has been presented in [5], in the context of a comprehensive comparison of validation methodologies.

The paper is organised as follows. Section 2.1 first relates the ISE estimators (1.2) to other ISE estimators. Then, assuming that the interpolated function f is a realisation of a Gaussian process with known moments, we present in Section 2.2 a computable criterion $\mathcal{R}(\zeta, \mathscr{F}_n)$ that evaluates the precision of empirical estimators of the form (1.2). In Section 3 we discuss optimisation of $\mathcal{R}(\zeta, \mathscr{F}_n)$, detailing appli-

*Corresponding author.
[1] We will often consider $\mathscr{X} = [0, 1]^d$.

cation of related existing algorithms to the specific conditions of the validation problem of interest here, and providing an instrumental interpretation of the corresponding "optimal" empirical ISE estimators. Since the "optimal" validation measure depends on the assumed GP model, the robustness and performance of the validation methodology presented are investigated numerically in Section 4, leading to two major conclusions. One concerns the validation weights $\mathbf{w}$, stating that the contributions of the individual errors $\varepsilon(\mathbf{z}_i)$ to $\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n};\zeta)$ must be down-weighted – with respect to taking ζ to be the uniform distribution over $\mathbf{Z}_m$ – to avoid overestimation of $\mathsf{ISE}(\eta_{\mathscr{F}_n})$. The second concerns the geometry of the validation design $\mathbf{Z}_m$, whose optimality is seen to be much less important that correct choice of the weights $\mathbf{w}$. Based on these numerical studies we propose a default choice for the covariance kernel of the GP model used, including its scale parameter. Finally, Section 5 summarises our findings and proposes some directions for future studies.

2. A CRITERION FOR VALIDATION MEASURES

Since f is unknown, we can at best expect to find an ISE estimator that will perform well for most functions f consistent with $\mathscr{F}_n$. To characterise this set of functions we adopt the Gaussian process framework – briefly recalled below – enabling us to subsequently derive a criterion to choose the validation measure ζ.

Before doing that, the next section puts our approach in perspective in relation to other (non-parametric) model validation methods.

2.1 Empirical ISE estimation

Non-parametric estimation of the ISE of a computational model learned on a dataset $\mathscr{F}_n$ is most commonly done using $\mathscr{F}_n$ itself. In cross-validation (CV), see e.g. [3, 2], the residuals $\varepsilon_i^{cv} = \mathbf{y}_i - \eta_{\mathscr{F}_n\setminus(\mathbf{x}_i,\mathbf{y}_i)}(\mathbf{x}_i)$ at each data point $(\mathbf{x}_i,\mathbf{y}_i)$ of a predictor fit to all other $n-1$ points of $\mathscr{F}_n$ are computed, and the ISE is estimated by their average:

$$\widehat{\mathsf{ISE}}_{cv} = \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i^{cv})^2 \ . \tag{2.1}$$

The setup considered in this paper is in some sense dual of CV. On the one hand, CV requires more information about η, assuming the ability to build the n new predictors $\eta_{\mathscr{F}_n\setminus(\mathbf{x}_i,\mathbf{y}_i)}$ (one for each point that is "left out") and assumes thus knowledge of how $\eta_{\mathscr{F}_n}$ is learned, while we consider $\eta_{\mathscr{F}_n}$ as a black-model delivered by a third party, using an undisclosed modelling approach. On the other hand, CV requires no any additional observations of f, while $\widehat{\mathsf{ISE}}(\zeta,\mathscr{F}_n)$ requires m new evaluations, one at each point of $\mathbf{Z}_m$.

Given the observations of f over a validation set $\mathbf{Z}_m$, a straightforward estimate of the ISE is the simple arithmetic mean of the squared values of the m residuals $\varepsilon_i = f(\mathbf{z}_i) - \eta_{\mathscr{F}_n}(\mathbf{z}_i)$ observed over the $\mathbf{z}_i \in \mathbf{Z}_m$:

$$\widehat{\mathsf{ISE}}_{un} = \frac{1}{m}\sum_{i=1}^{m}\varepsilon_i^2 \ , \tag{2.2}$$

a special case of (1.2), obtained by letting ζ be the uniform distribution over $\mathbf{Z}_m$: $\zeta = (1/m)\sum_i \delta_{\mathbf{z}_i}$, with $\delta_{\mathbf{a}}$ denoting the unit point-mass at $\mathbf{x}=\mathbf{a}$.

Let p_η denote the (unknown) probability density of the residuals $\varepsilon(\mathbf{x})$ when $\mathbf{x} \sim \mu$. For expression (2.2) to be a Monte Carlo estimate of the ISE integral, the observed $\varepsilon(\mathbf{Z}_m) = \{\varepsilon_i\}_{i=1}^m$ must be a plausible i.i.d.[2] sample from p_η, which is generally not true. Consider for instance that $\mathbf{Z}_m$ is a space filling continuation of $\mathbf{X}_n$, sampling the regions of $\mathscr{X}$ the most distant from $\mathbf{X}_n$. In this situation we can anticipate that $\varepsilon(\mathbf{Z}_m)$ will be biased towards the upper limit of the support of p_η, and thus that $\widehat{\mathsf{ISE}}_{un}$ will over-estimate ISE. The contribution of the observed residuals to $\widehat{\mathsf{ISE}}$ must thus be adjusted, counterbalancing this poor sampling of regions where the weakest residual values are expected. The validation measures ζ proposed in this paper automatically implement this variable residual scaling, relying on a prior stochastic model for f to infer how well the observed $\varepsilon(\mathbf{Z}_m)$ are expected to be representative of the errors over the entire $\mathscr{X}$. It follows from the above that there is no reason for imposing that the validation measure ζ be a probability distribution i.e., that $\sum_i \mathbf{w}_i = 1$. Our methodology drops this common constraint, defining an un-normalised measure ζ adjusted to the geometry of $\mathbf{Z}_m$ relative to $\mathbf{X}_n$. To corroborate this choice, note that when $\eta_{\mathscr{F}_n}$ is an interpolator, so that $\varepsilon(\mathbf{x}_i)=0$ for all $\mathbf{x}_i \in \mathbf{X}_n$, incorporation of these n zero residuals in (2.2), which should lead to a better estimator of $\mathsf{ISE}(\eta_{\mathscr{F}_n})$, yields

$$\widehat{\mathsf{ISE}}^{\star}_{un} = \frac{1}{m+n}\sum_{i=1}^{m}\varepsilon_i^2 < \widehat{\mathsf{ISE}}_{un} \ ,$$

for which $\sum_i \mathbf{w}_i = m/(n+m) < 1$.

2.2 Choosing the validation measure: a GP-based criterion

The estimation error $|\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n};\zeta) - \mathsf{ISE}(\eta_{\mathscr{F}_n})|$ is not a computable criterion that we can optimise to choose ζ. A possible approach would be to consider that f belongs to some class of functions $\mathcal{S}$ and optimise the worst estimation performance over all $f \in \mathcal{S}$. Here we follow an alternative and simpler route, assuming that f is a realization of a Gaussian Process (GP), or Gaussian Random Field, and minimising a moment of the ISE estimation error under the assumed model.

[2] independent and identically distributed.

Assume thus that f is a sample function $\mathcal{F}_x$ from a GP indexed by $\mathscr{X}$, with known second-order characteristics $\mathsf{E}\{\mathcal{F}_x\mathcal{F}_{x'}\} = \sigma^2 K(\mathbf{x},\mathbf{x}')$: $f \sim \mathcal{GP}^f(m(\mathbf{x}), \sigma^2 K(\mathbf{x},\mathbf{x}'))$. Kernel K is supposed to be Strictly Positive Definite (SPD), and, for the sake of simplicity, we consider that $m(\mathbf{x}) = \mathsf{E}\{\mathcal{F}_x\} = 0$ for all $\mathbf{x} \in \mathscr{X}$. Extension of the material presented below to the case of a linearly parameterized mean, with $\mathsf{E}\{\mathcal{F}_x\} = \boldsymbol{\beta}^\top \mathbf{h}(\mathbf{x})$ for a vector $\boldsymbol{\beta}$ of unknown parameters and a vector $\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), \ldots, h_p(\mathbf{x}))^\top$ of p known functions of $\mathbf{x}$ is possible via some adaptation.

Under the assumption above $\mathsf{ISE}(\eta_{\mathscr{F}_n})$ given by (1.1) is a random variable. The statistical moments of $\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n};\zeta)$ under this stochastic model for f provide computable and pertinent criteria to chose ζ. We use the Mean Squared Error (MSE) of $\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n};\zeta)$ given $\mathscr{F}_n$,

$$\begin{aligned}\mathcal{R}(\zeta_m,\mathscr{F}_n) &= \mathsf{E}\left\{ \left[\mathsf{ISE}(\eta_{\mathscr{F}_n}) - \widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n};\zeta_m)\right]^2 \middle| \mathscr{F}_n \right\} \\ &= \mathsf{E}\left\{ \left[\int_{\mathscr{X}} [\mathcal{F}_x - \eta_{\mathscr{F}_n}(\mathbf{x})]^2 \,(\zeta_m - \mu)(\mathrm{d}\mathbf{x})\right]^2 \middle| \mathscr{F}_n \right\},\end{aligned}$$

as a criterion to choose the validation design: $\zeta_m^\star(\mathscr{F}_n) \in \arg\min_{\zeta_m} \mathcal{R}(\zeta_m,\mathscr{F}_n)$.

The GP assumption defines a prior distribution for f, which given $\mathscr{F}_n$ can be updated into the posterior distribution of its values over the unobserved points, with mean $\mathsf{E}\{\mathcal{F}_x|\mathscr{F}_n\} = \mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\mathbf{y}_n$ and covariance $\mathsf{E}\{\mathcal{F}_x\mathcal{F}_{x'}|\mathscr{F}_n\} = \sigma^2 K_{|n}(\mathbf{x},\mathbf{x}')$, with $K_{|n}$ defined by

$$K_{|n}(\mathbf{x},\mathbf{x}') = K(\mathbf{x},\mathbf{x}') - \mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x}') \geq 0\,, \qquad (2.3)$$

for any $\mathbf{x}$, $\mathbf{x}'$ in $\mathscr{X}$, where

$$\begin{aligned}\mathbf{k}_n(\mathbf{x}) &= (K(\mathbf{x},\mathbf{x}_1)\ \ldots, K(\mathbf{x},\mathbf{x}_n))^\top \\ \{\mathbf{K}_n\}_{i,j} &= K(\mathbf{x}_i,\mathbf{x}_j)\,,\ i,j = 1,\ldots,n\,.\end{aligned}$$

The $n \times n$ matrix $\mathbf{K}_n$ is SPD as K is SPD (we assume that the $\mathbf{x}_i$ in $\mathbf{X}_n$ are pairwise distinct). Note that $K_{|n}(\mathbf{x}_i,\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathscr{X}$ and all $\mathbf{x}_i \in \mathbf{X}_n$. The Integrated Mean Squared Error (IMSE) is thus

$$\begin{aligned}\mathsf{IMSE}(\mathscr{F}_n) &= \int_{\mathscr{X}} \mathsf{E}\left\{[\eta_{\mathscr{F}_n}(\mathbf{x}) - f(\mathbf{x})]^2\,|\mathscr{F}_n\right\} \mu(\mathrm{d}\mathbf{x}) \\ &= \int_{\mathscr{X}} \mathsf{E}\left\{\left[\eta_{\mathscr{F}_n}(\mathbf{x}) - \mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\mathbf{y}_n\right]^2|\mathscr{F}_n\right\} \mu(\mathrm{d}\mathbf{x}) \\ &\qquad + \sigma^2 \int_{\mathscr{X}} K_{|n}(\mathbf{x},\mathbf{x})\,\mu(\mathrm{d}\mathbf{x})\,.\end{aligned}$$

$\mathsf{IMSE}(\mathscr{F}_n)$ is minimum when $\eta_{\mathscr{F}_n}(\mathbf{x})$ is the posterior mean $\mathbf{k}_n(\mathbf{x})^\top\mathbf{K}_n^{-1}\mathbf{y}_n$. This minimum value depends only on the learning design $\mathbf{X}_n$ and is given by

$$\mathsf{IMSE}^\star(\mathbf{X}_n) = \sigma^2 \int_{\mathscr{X}} K_{|n}(\mathbf{x},\mathbf{x})\,\mu(\mathrm{d}\mathbf{x}) \leq \mathsf{IMSE}(\mathscr{F}_n)\,. \qquad (2.4)$$

In this situation, direct calculation yields $\mathcal{R}(\zeta_m,\mathscr{F}_n) = \mathcal{R}(\zeta_m,\mathbf{X}_n)$, with

$$\begin{aligned}\mathcal{R}(\zeta_m,\mathbf{X}_n) &= \sigma^4 \int_{\mathscr{X}^2} \overline{K}_{|n}(\mathbf{x},\mathbf{x}')(\zeta_m - \mu)(\mathrm{d}\mathbf{x})(\zeta_m - \mu)(\mathrm{d}\mathbf{x}') \\ &= \sigma^4 \mathscr{E}_{\overline{K}_{|n}}(\zeta_m - \mu)\,, \qquad (2.5)\end{aligned}$$

proportional to the energy of the signed measure $\zeta_m - \mu$ for the kernel $\overline{K}_{|n}(\mathbf{x},\mathbf{x}') = (1/\sigma^4)\mathrm{E}\left[\varepsilon^2(\mathbf{x})\varepsilon^2(\mathbf{x})'\middle|\mathscr{F}_n\right]$, a scaled version of the second order moment of the squared residuals. Under $\mathcal{GP}^f$,

$$\overline{K}_{|n}(\mathbf{x},\mathbf{x}') = 2\,K_{|n}^2(\mathbf{x},\mathbf{x}') + K_{|n}(\mathbf{x},\mathbf{x})K_{|n}(\mathbf{x}',\mathbf{x}')\,. \qquad (2.6)$$

This still means that $\mathcal{R}(\zeta_m,\mathbf{X}_n)$ is proportional to the squared Maximum Mean Discrepancy (MMD) between the measures ζ_m and μ for the kernel $\overline{K}_{|n}$ [16, Def. 10]. Under the GP modelling framework assumed we are thus lead to

$$\zeta_m^\star(\mathscr{F}_n) \in \arg\min_{\zeta_m} \mathscr{E}_{\overline{K}_{|n}}(\zeta_m - \mu)\,,$$

with $\overline{K}_{|n}(\mathbf{x},\mathbf{x}')$ given by (2.6).

When $\eta_{\mathscr{F}_n}$ does not interpolate $\mathbf{y}_n$, and under the same GP for f, similar developments still give $\mathcal{R}(\zeta_m,\mathscr{F}_n) = \sigma^4\mathscr{E}_{\overline{K}_{|n}}(\zeta_m - \mu)$, with now

$$\begin{aligned}\overline{K}_{|n}(\mathbf{x},\mathbf{x}') &= 2\left[K_{|n}(\mathbf{x},\mathbf{x}') + 2\,\widehat{\delta}_n(\mathbf{x})\widehat{\delta}_n(\mathbf{x}')\right]K_{|n}(\mathbf{x},\mathbf{x}') \\ &\quad + \left[\widehat{\delta}_n^2(\mathbf{x}) + K_{|n}(\mathbf{x},\mathbf{x})\right]\left[\widehat{\delta}_n^2(\mathbf{x}') + K_{|n}(\mathbf{x}',\mathbf{x}')\right],\end{aligned}$$

where $\widehat{\delta}_n(\mathbf{x}) = \mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\mathbf{y}_n - \eta_{\mathscr{F}_n}(\mathbf{x})$. In the following, we always consider that $\eta_{\mathscr{F}_n}$ is the optimal interpolator $\mathbf{k}_n(\mathbf{x})^\top\mathbf{K}_n^{-1}\mathbf{y}_n$, and thus that (2.6) holds.

Kernels $\overline{K}_{|n}$ present a number of features which are not shared by the most commonly used GP kernels. The assumption that $\eta_{\mathscr{F}_n}$ is an interpolator, i.e. $\varepsilon(\mathbf{x}_i) = 0$, implies that $\overline{K}_{|n}(\mathbf{x}_i,\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathscr{X}$ and all $\mathbf{x}_i \in \mathbf{X}_n$. The squared error process is thus non-stationary, with a spatial coherency structure that is strongly dictated by the geometry of $\mathbf{X}_n$. Adapting the validation weights $\mathbf{w}_i$ to this correlation structure dictates the performance of $\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n};\zeta_m)$ in a critical manner. Yet, as the numerical studies of Section 4 show, exploiting the particular shape of $\overline{K}_n$ when choosing the validation points $\mathbf{Z}_m$ is less critical (as long as they do not fall in the vicinity of $\mathbf{X}_n$).

Finally, notice that $\overline{K}_{|n}$ is PD. Indeed, the Hadamard product $\mathbf{C}_n^{\circ 2}$ with elements $\{\mathbf{C}_n^{\circ 2}\}_{i,j} = C^2(\mathbf{x}_i,\mathbf{x}_j)$, $i,j = 1,\ldots,n$, is PD when the matrix $\mathbf{C}_n$ with elements $\{\mathbf{C}_n\}_{i,j} = C(\mathbf{x}_i,\mathbf{x}_j)$ is PD. Hence, the positive definiteness of $K_{|n}$ implies that $K_{|n}^2$ is PD, which in turn implies that $\overline{K}_{|n}$ is also PD.

3. MINIMISATION OF $\mathscr{E}_{\overline{K}_{|n}}$

We address now the minimisation of $\mathscr{E}_{\overline{K}_{|n}}(\zeta_m - \mu)$ with respect to ζ_m.

We drop the two constraints usually imposed on weights: besides the sum-of weights-equals-one constraint (see Section 2.1), we also do not impose $\mathbf{w}_i \geq 0$. Imposing positivity would be natural if the observations were noisy independent random samples of the interpolation error, but here the ε_i are noise-free and, more importantly, strongly linked by a coherency structure dictated by both the regularity characteristics of f and the quality of $\eta_{\mathscr{F}_n}$ as an interpolator. Nonetheless, our numerical experiments show that the $\mathbf{w}_i$ are almost always positive; see for example Figures 12 and 13.

Since for a given f and $\eta_{\mathscr{F}_n}$ the validation residuals are deterministic, repeating validation points or choosing $\mathbf{z}_i \in \mathbf{X}_n$ brings no additional information. We thus restrict $\mathbf{Z}_m$ to configurations of m distinct points in $\mathscr{X} \setminus \mathbf{X}_n$. The minimisation of $\mathscr{E}_{\overline{K}_{|n}}(\zeta_m - \mu)$ with respect to the parameters of ζ_m is a non-linear optimisation problem over a large dimensional space ($m(d+1)$ scalar parameters when $\mathscr{X} \subset \mathbb{R}^d$). As briefly evoked in the introduction, rather than fixing upfront the size m of the validation design, we are interested in finding nested sequences of validation designs, generated by a sequence of identical steps, each one increasing the design size by one:

$$\mathbf{Z}_{m+1} = \mathbf{Z}_m \cup \{\mathbf{z}_{m+1}\}\,, \tag{3.1}$$

where $\mathbf{z}_{m+1}$ is restricted to $\mathscr{X}_m = \mathscr{X} \setminus \{\mathbf{Z}_m \cup \mathbf{X}_n\}$.

Before we present in Section 3.2 the sequential Bayesian quadrature algorithm that performs this iterative construction, greedily decreasing $\mathscr{E}_{\overline{K}_{|n}}(\zeta_m - \mu)$ at each step, we present background on relevant literature on iterative energy (or, equivalently, MMD) minimisation.

3.1 Background

Kernel herding (KH) [19] can be seen to correspond to the Frank-Wolfe conditional gradient algorithm [1] applied to MMD minimisation, that is, to the vertex-direction method with predefined step-length, commonly used in optimal experimental design since the pioneering work of H.P. Wynn [20] and V.V. Fedorov [4]. It is an accretive method[3], generating a sequence $\mathbf{z}_1, \mathbf{z}_2, \ldots$ which can be incrementally grown to any target size m.

In Bayesian quadrature (BQ) [11, 15] the goal is to choose samples that best approximate an integral by exploiting the assumption that the integrated function is the realisation of a GP. Sequential BQ (SBQ) sequentially expands the set of sampled points by adding a new sample at the point that decreases the variance of the integral estimate the most. This variance is shown to be the MMD between the target integral measure and the discrete measure that implements the quadrature rule for the kernel of the assumed GP model.

KH and SBQ are closely related, see e.g. [8], both attempting to minimise the same MMD. The two techniques embed the problem in consideration in the RKHS of a positive definite kernel that is chosen to reflect the characteristics of the underlying data distribution (in the original formulation of KH) or of the integrated functions (in SBQ). As stressed in [8], a major distinction between the two techniques concerns the weights assigned to each sample, which are uniform for standard KH, while they are optimally selected in SBQ. The two methods differ both in complexity and performance: SBQ is superior to standard (uniform weight) KH, this improvement coming at the cost of an increased complexity, $O(n)$ for KH and $O(n^2)$ for SBQ when constructing an n-point design; see [12].

Experiments combining the two methodologies, by using the optimal BQ weights for a design found by standard KH, show that correct weighting is more critical than sample placement [8], affecting in particular the algorithm's convergence rate: KH has performance similar to SBQ for small design sizes, but displays worse performance as design size grows.

The validation setup of this paper coincides with the framework assumed by BQ, our final goal being to estimate an integral from a small number of samples, and we also resort to a GP assumption. As in BQ, the weights of our empirical estimator do not need to sum to 1 and are not necessarily positive, and the optimal solution minimises an MMD. Placing the GP assumption not directly on the function we wish to integrate – in our case $\varepsilon^2(\mathbf{x})$ – but on the interpolated f, leads to the identification of the pertinent MMD kernel under our validation framework as the non-stationary kernel $\overline{K}_{|n}$, whose structure encodes the geometry of the learning design $\mathbf{X}_n$.

Both KH and BQ assume that the RKHS kernel is characteristic, meaning that the corresponding MMD between two probability measures is zero if and only if these two measures coincide. Kernel $\overline{K}_{|n}$ is not characteristic, and in particular it cannot differentiate between measures that differ only over the finite set $\mathbf{X}_n$, where $\overline{K}_{|n}$ is zero. However, as we stressed before, since we know that $\varepsilon(\mathbf{x}) = 0$ for $\mathbf{x} \in \mathbf{X}_n$, the set of target measures over which we optimise $\mathscr{E}_{\overline{K}_{|n}}(\zeta_m - \mu)$ all put zero mass on $\mathbf{X}_n$, and thus it still makes sense to minimise it.

3.2 Greedy optimisation of $\mathscr{E}_{\overline{K}_{|n}}(\zeta_m - \mu)$

In this section we briefly present the SBQ method, reinterpreting it in the validation setup of interest to us.

By noting that $\mathscr{E}_{\overline{K}_{|n}}(\zeta_m - \mu)$ is quadratic in the $\{\mathbf{w}_i\}_{i=1}^m$, the weights $\tilde{\mathbf{w}}(\mathbf{Z}_m)$ that minimise it for a given $\mathbf{Z}_m$ can be

[3] However, it does not provide the optimal design for a fixed m: the construction of one-shot m-point designs minimizing a MMD criterion is considered for instance in [10, 14]; we do not develop this aspect here.

seen to be given by

$$\tilde{\mathbf{w}}(\mathbf{Z}_m) = \overline{K}_{|n}(\mathbf{Z}_m,\mathbf{Z}_m)^{-1} P_{\overline{K}_{|n}}(\mathbf{Z}_m)\,, \tag{3.2}$$

where the $m \times m$ matrix $\overline{K}_{|n}(\mathbf{Z}_m,\mathbf{Z}_m)$ has generic element $\overline{K}_{|n}(\mathbf{z}_i,\mathbf{z}_j)$ and the i-th entry of the m-dimensional column vector $P_{\overline{K}_{|n}}(\mathbf{Z}_m)$ is the potential of μ associated with kernel $\overline{K}_{|n}$ at validation point $\mathbf{z}_i$:

$$\left[P_{\overline{K}_{|n}}(\mathbf{Z}_m)\right]_i = P_{\overline{K}_{|n}}(\mathbf{z}_i) = \int_{\mathscr{X}} \overline{K}_{|n}(\mathbf{z}_i,\mathbf{x})\mu(d\mathbf{x})\,. \tag{3.3}$$

Remembering that $\sigma^4\overline{K}_{|n}(\mathbf{x},\mathbf{x}') = E\left[\varepsilon^2(\mathbf{x})\varepsilon^2(\mathbf{x}')|\mathscr{F}_n\right]$, $P_{\overline{K}_{|n}}(\mathbf{z})$ can be recognised as

$$P_{\overline{K}_{|n}}(\mathbf{z}) = \frac{1}{\sigma^4} E\left[\varepsilon^2(\mathbf{z})\,\mathsf{ISE}(\eta_{\mathscr{F}_n})\middle|\,\mathscr{F}_n\right]\,. \tag{3.4}$$

Define

$$\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m) = \overline{K}_{|n}(\mathbf{x},\mathbf{Z}_m)\overline{K}_{|n}(\mathbf{Z}_m,\mathbf{Z}_m)^{-1}\varepsilon^2(\mathbf{Z}_m)\,. \tag{3.5}$$

Under the posterior model, i.e., given $\mathscr{F}_n$, $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)$ is the Minimum MSE (MMSE) linear estimate of $\varepsilon^2(\mathbf{x})$ given the residuals observed over $\mathbf{Z}_m$. When the weights $\mathbf{w}_i$ of the validation measure are given by (3.2), $\widehat{\mathsf{ISE}}$ has thus the following simple and enlightening expression:

$$\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n},\mathbf{Z}_m) = \sum_i \tilde{\mathbf{w}}_i(\mathbf{Z}_m)\varepsilon^2(\mathbf{z}_i) = \int_{\mathscr{X}} \widehat{\varepsilon^2}_{\mathscr{F}_n}(x|\mathbf{Z}_m)\,\mu(dx)\,. \tag{3.6}$$

Note that the weights $\tilde{\mathbf{w}}_i(\mathbf{Z}_m)$, and thus the estimator $\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n},\mathbf{Z}_m)$ itself, are independent of σ^2. The estimators $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)$ and $\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n},\mathbf{Z}_m)$ rely on the assumed GP model $\mathcal{GP}^f$ for f, but as explained in [17, Sect. 3.2], model misspecification has a much smaller effect on linear predictions than on evaluation of the MSE. One important strength of the approach is thus that our estimator of $\mathsf{ISE}(\eta_{\mathscr{F}_n})$ does not require estimation of the MSE associated with a prediction. As shown in Appendix A, this is no longer the case when one attempts at removing the bias of $\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n},\mathbf{Z}_m)$, which leads to estimators that are much less robust to model misspecification.

For a given ζ_m define $\mathcal{E}_m(\mathbf{x}) = \mathscr{E}_{\overline{K}_{|n}}(\zeta^\star_{m+1}-\mu)$, the energy for measure $\zeta^\star_{m+1}$ having support $\mathbf{Z}_{m+1}(\mathbf{x}) = \mathbf{Z}_m \cup \{\mathbf{x}\}$ and optimal weights $\tilde{\mathbf{w}}(\mathbf{Z}_{m+1}(\mathbf{x}))$ given by (3.2). If $\zeta_m = \zeta(\tilde{\mathbf{w}}(\mathbf{Z}_m),\mathbf{Z}_m)$, and for $\mathbf{x} \in \mathscr{X}_m$, we have

$$\mathcal{E}_m(\mathbf{x}) = \mathscr{E}_{\overline{K}_{|n}}(\zeta_m-\mu) - \frac{\left(P_{\overline{K}_{|n}}(\mathbf{x}) - \overline{K}_{|n}(\mathbf{x},\mathbf{Z}_m)\overline{K}_{|n}(\mathbf{Z}_m,\mathbf{Z}_m)^{-1}P_{\overline{K}_{|n}}(\mathbf{Z}_m)\right)^2}{s^2(\mathbf{x})}\,,$$

where,

$$s^2(\mathbf{x}) = \overline{K}_{|n}(\mathbf{x},\mathbf{x}) - \overline{K}_{|n}(\mathbf{x},\mathbf{Z}_m)\overline{K}_{|n}(\mathbf{Z}_m,\mathbf{Z}_m)^{-1}\overline{K}_{|n}(\mathbf{Z}_m,\mathbf{x}) = \frac{1}{\sigma^4}E\left[\left(\varepsilon^2(\mathbf{x}) - \widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)\right)^2\middle|\,\mathscr{F}_n\right]\,.$$

The next validation point is thus a maximiser of the second term in $\mathcal{E}_m(\mathbf{x})$, which can equivalently be written as

$$\mathbf{z}_{m+1} \in \arg\max_{\mathbf{x}\in\mathscr{X}_m} \frac{E\left[\mathsf{ISE}(\eta_{\mathscr{F}_n})\left(\varepsilon^2(\mathbf{x}) - \widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)\right)\middle|\,\mathscr{F}_n\right]}{E\left[\left(\varepsilon^2(\mathbf{x}) - \widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)\right)^2\middle|\,\mathscr{F}_n\right]}\,. \tag{3.7}$$

The numerator measures how much $\mathsf{ISE}(\eta_{\mathscr{F}_n})$ and the error of $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{z}|\mathbf{Z}_m)$ as an estimate of $\varepsilon^2(\mathbf{x})$ are statistically associated. Points where this term is large are good candidates to extend the current design. The denominator penalises points $\mathbf{x}$ where $\varepsilon^2(\mathbf{x})$ is estimated with a large MSE, tending to keep $\mathbf{z}_{m+1}$ away from the boundaries of $\mathscr{X}$ (where the uncertainty is in general large), as the numerical studies presented later will show.

The recursive extension of the validation measure is initiated with $\mathbf{Z}_1 = \{\mathbf{z}_1\}$ solution of

$$\mathbf{z}_1 = \max_{\mathbf{x}\in(\mathscr{X}\setminus\mathbf{X}_n)} \frac{P_{\overline{K}_{|n}}(\mathbf{x})^2}{\overline{K}_{|n}(\mathbf{x},\mathbf{x})}\,. \tag{3.8}$$

In practice, a finite set $\mathscr{X}_L \subset \mathscr{X}$, for instance the L first elements of a low-discrepancy sequence in $\mathscr{X}$, of a regular grid in $\mathscr{X}$ if d is not too large, is substituted for $\mathscr{X}$ in (3.7) and (3.8). The determination of $\mathbf{z}_{m+1}$, $m \geq 0$, then requires the evaluation of $P_{\overline{K}_{|n}}(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{X}_L \setminus \mathbf{X}_n$. This calculation is done once for all, at the initialisation of the algorithm. In the numerical examples of Section 4, $P_{\overline{K}_{|n}} = P_{\overline{K}_{|n},\mu}$ is replaced by $P_{\overline{K}_{|n},\mu_L}$, with μ_L the uniform (discrete) measure uniform on $\mathscr{X}_L$, see Appendix C for details. When K is a tensor-product kernel and μ is uniform on $\mathscr{X} = [0,1]^d$, $P_{\overline{K}_{|n}}(\mathbf{x})$ can often be calculated explicitly; see Appendix B.

With the aid of a one-dimensional example we formulate now a number of comments about the expected behaviour and properties of the estimators $\widehat{\mathsf{ISE}}$ obtained by repeated application of (3.7) – to extend $\mathbf{Z}_m$ to $\mathbf{Z}_{m+1}$ – and (3.2) – fixing the weights of ζ_{m+1}, and thus $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_{m+1})$ for the subsequent design extension. The red bold curve in the top panel of Figure 1 plots the squared residuals $\varepsilon^2(\mathbf{x})$ of the interpolator $\eta_{\mathscr{F}_n}$ for the function f plotted in the bottom panel (where $\eta_{\mathscr{F}_n}$ and f are in red and green, respectively), trained on the learning design of size 10 indicated by the red stars. The blue and green curves on the top panel are the squared residuals $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)$ predicted by two distinct ζ_m ($m = 10$), both generated using (3.7) and (3.2), but assuming distinct kernels $K(\mathbf{x},\mathbf{x}')$: Cauchy (in blue) and Matérn 3/2 (in green), with range parameters θ as indicated in the legend[4] The (nearly coincident) validation designs

[4]The exact definition of these kernels is given in Appendix C.

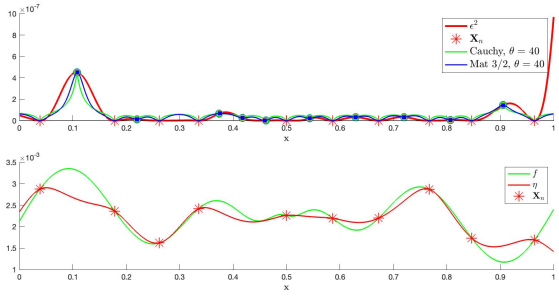

Figure 1: Top: $\varepsilon(\mathbf{x})$ (red) and $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)$ (blue and green) for two distinct GP models. Down: f, η and $\mathbf{X}_n$.

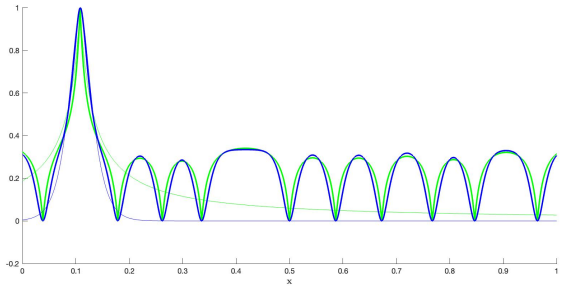

Figure 2: $\overline{K}_{|n}(\mathbf{z}_1,\mathbf{x})/\overline{K}_{|n}(\mathbf{z}_1,\mathbf{z}_1)$ (bold lines) and $K(\mathbf{z}_1,\mathbf{x})/K(\mathbf{z}_1,\mathbf{z}_1)$ (thin lines) for the Cauchy and Matérn kernels used in Figure 1 (same colour code) and $z_1 \simeq 0.1$.

$\mathbf{Z}_m$ are indicated by the squares and circles filled with the corresponding colours.

Remark first that as anticipated both designs $\mathbf{Z}_m$ have no points in the boundaries of $\mathscr{X}$, even if the uncertainty affecting $\varepsilon^2(\mathbf{x})$ is large in those regions. Those familiar with optimal interpolation using monotonically decreasing stationary covariance kernels may be surprised by the fact that in intervals between learning points containing no validation points (e.g. around $\mathbf{x} \simeq 0.3$) the interpolated squared residual is non-zero, i.e., $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m) > 0$. This is a consequence of the particular shape of kernel $\overline{K}_{|n}$, strongly dictated by the geometry of $\mathbf{X}_n$, which has larger values between residual values at pairs of points at large distance than the original K, as shown in Figure 2. For $\mathbf{z}_1 \simeq 0.1 \in \mathbf{Z}_m$, the figure plots normalised versions of both the assumed (stationary) signal correlation $K(\mathbf{z}_1 - \mathbf{x})$ (in thin coloured lines) as well as kernel $\overline{K}_{|n}(\mathbf{z}_1,\mathbf{x})$ (bold lines), with the same colour code as in Figure 2. The similarity of the two $\overline{K}_{|n}$ allows us to expect that the estimator will have some robustness with respect to the assumed GP model. The numerical studies presented in Section 4 confirm this expectation.

Above, we recognised $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)$ as the MMSE linear estimator of $\varepsilon^2(\mathbf{x})$ given $\varepsilon^2(\mathbf{Z}_m)$. Being agnostic with respect to the expected values of the involved random variables, estimators $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)$, and thus $\widehat{\mathsf{ISE}}$, are biased. We investigate in Appendix A the possibility of exploiting knowledge of the first moments, namely $E\left[\varepsilon^2(\mathbf{x})\middle|\mathscr{F}_n\right] = \sigma^2 K_{|n}(\mathbf{x},\mathbf{x})$ and $E\left[\mathsf{ISE}(\eta_{\mathscr{F}_n})\middle|\mathscr{F}_n\right] = \mathsf{IMSE}(\mathscr{F}_n)$, to replace $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)$ in (3.6) with an unbiased estimator. Unfortunately, bias correction comes at the price of loosing robustness with respect to the assumed GP model for f, as we might expect given the explicit dependency on σ^2 of both expected values. Thus, the unbiased estimators in Appendix A cannot be considered as instrumental alternatives to $\widehat{\mathsf{ISE}}$, and we will not consider them in the numerical study of Section 4.

4. NUMERICAL EXPERIMENTS

Section 4.1 presents numerical studies that demonstrate the robustness of $\widehat{\mathsf{ISE}}$ with respect to the assumed GP model, with ζ_m found by SBQ. Section 4.2 confirms the importance of using $\overline{K}_{|n}$ to define the energy minimised by SBQ. We then study, in Section 4.3, the impact of using KH, which has slightly smaller computational complexity, rather than SBQ, to find the validation support of ζ_m, concluding that it leads to worse performance and is subject to numerical instability. Finally, Section 4.4 illustrates via some examples the properties of the validation measures, in particular their space-filling properties and the fact that they down-weight the observed squared residuals. In all examples $\mathscr{X} = [0,1]^d$, with $d = 1$, 2 or 3. Use of larger values of d lead to similar conclusions, see [13].

Our analysis resorts to simulations from several (zero mean) GP models, and the MSE of the ISE estimates is approximated by averaging the squared errors of $\widehat{\mathsf{ISE}}^{(i)}$ on $M = 500$ realisations $\{f^{(i)}\}_{i=1}^M$ of the assumed GP model. We reserve the notation $Q(\cdot,\cdot;\theta_0)$ for the kernel of the GP model from which is f sampled, θ_0 being thus "the true" scale parameter. The scale parameter is adapted to the size of the learning design, $\theta_0 = n^{1/d}$, such that good interpolation performance over $\mathscr{X}$ can be attained with n points. Designs $\mathbf{X}_n$ are always space-filling, and $\eta_{\mathscr{F}_n}$ is the optimal Bayesian interpolator for the simulated GP model. See Appendix C for details.

$K(\cdot,\cdot;\theta)$ denotes the kernel of the GP model assumed by the design algorithm that produces ζ_m, with θ its scale parameter. In all numerical examples we will always consider $\sigma^2 = 1$. The influence of θ is studied for $\theta \in \left[n^{1/d}/4, \max(n^{1/d}, 2\,(n+m)^{1/d})\right]$, an interval that always contains

$$\theta_c(n,m,d) = (n+m)^{1/d}$$

(as well as $\theta_0 = n^{1/d}$). All plots consider the normalisation θ/θ_c, such that $\theta_c \leftrightarrow 1$ in the plots shown. In all plots of this section the special symbols in the plotted curves indicate that the design algorithm uses $\theta = \theta_0$, the scale parameter of the simulated GP model.

4.1 Robustness with respect to assumed GP model

We address robustness by studying how much the MSE of $\widehat{\mathsf{ISE}}$ is affected by model mismatch. Figure 3 plots empirical estimates of $\mathcal{R}(\zeta_m, \mathscr{F}_n)$. Kernel Q is the Matérn 3/2 kernel, $\mathbf{X}_n$ has $n = 10$ points and $d = 1$, $\theta_0 = n$. The panels correspond to different values of the regularity parameter – $\nu = 1/2, 3/2, 5$, from left to right – of the Matérn kernel K.

The three curves in each plot correspond to different sizes of $\mathbf{Z}_m$: $m \in \{5, 10, 20\}$ (in blue, red and brown, respectively), plotting $\mathcal{R}$ as a function of θ. The black stars indicate $\theta = \theta_0$. Comparison of the three panels confirms the anticipated robustness of the estimator. When K has higher regularity than Q, as in the rightmost panel ($\nu = 5/2$), the curves are almost identical to the central panel, where the correct model is used. However, the assumption of a less regular model, as in the leftmost panel, may significantly degrade performance. The estimators are reasonably robust with respect to precise choice of the scale parameter if values $\theta \simeq \theta_c$ are used.

Figure 4 reproduces the same study for simulations from a process with a Cauchy kernel and larger $\mathbf{Z}_m$: $m \in \{10, 20, 30\}$ (left to right). As in previous figure, K is the Matérn kernel and the three panels correspond to different smoothness parameters $\nu \in \{1/2, 3/3, 5/2\}$. Here the simulated model has a weaker regularity than the models assumed, and a noticeable performance degradation is now observed for the smaller designs and the more regular Matérn kernel with $\nu = 5/2$. Similar results were obtained when simulating from other models and for higher values of d.

Finally, Figure 5 shows, for the same validation designs $\mathbf{Z}_m$ as in Figure 3, the MSE of $\widehat{\mathsf{ISE}}_{un}$ given by equation (2.2), estimated over 500 realisations of a GP with the same Matérn 3/2 model. We can see that proper residual weighting leads to a significant decrease of the estimation error, which is nearly one order magnitude larger in Figure 5 than for the optimal BQ designs of Figure 3.

The experiments in this section suggest a rule-of-thumb to chose the kernel used by the design algorithm: K should model functions with a reasonably large degree of smoothness (Matérn 3/2 was found to be a good compromise), with a scale parameter θ dependent on the sizes of the learning and validation sets. For the Matérn family used in our experiments a good choice is $\theta \simeq (n + m)^{1/d}$, automatically adjusting to the actual total number of residual samples.

4.2 Impact of $\overline{K}_{|n}$

Our main novel contribution is the identification of $\overline{K}_{|n}$ as the kernel that appears in the MMD that the validation measure ζ_m, both its weights and its support, must minimise. One may question the importance of using the non-stationary conditional kernel $\overline{K}_{|n}$ to find $\mathbf{Z}_m$, instead of directly using kernel K. We now compare the performance of the empirical estimator $\widehat{\mathsf{ISE}}$ with $\mathbf{Z}_m$ determined by SBQ for kernel $\overline{K}_{|n}$, as in Section 3.2, which from now on we denote by $\zeta^{\mathsf{BQ}\star}$, with use of a validation measure ζ_K^{BQ} whose support $\mathbf{Z}_m$ is incrementally found by SBQ for kernel K, the continuation of $\mathbf{X}_n$ that is optimal to integrate the function f. Independently of how $\mathbf{Z}_m$ was found, the validation measures ζ_m used by the estimators $\widehat{\mathsf{ISE}}$ always have optimal weights given by (3.2).

Figures 6 ($d = 1$) and 7 ($d = 3$) show the empirical MSE of $\widehat{\mathsf{ISE}}$ for $\zeta^{\mathsf{BQ}\star}$ (black lines) and ζ_K^{BQ} (red lines) observed when Q is the Matérn 3/2 kernel (top) and the Cauchy kernel (bottom), for a learning design of size $n = 10\,d$. From left to right, K is a Matérn kernel with $\nu = 1/2$, 3/2 and 5/2. The size of the validation designs, $m \in \{10\,d, 20\,d, 30\,d\}$, is indicated by the line symbols ($+$, $\star$ and $\circ$, respectively). We can see that the two estimators display similar performance and robustness with respect to mis-modelling. When m is small $\zeta^{\mathsf{BQ}\star}$ often yields smaller MSE, see top curves, but the red and black curves are almost coincident for the larger values of m. These results, which are representative of those obtained for other choices of Q and d, indicate that correct residual weighting is more important than the detailed placement of the validation points $\mathbf{Z}_m$.

Note that, in the configurations tested, the default rule-of-thumb for the choice of K and θ presented in Section 4.1 leads indeed to good and stable performance.

4.3 Comparison with Kernel Herding

Considering only validation measures ζ with uniform weights ($1/m$), standard KH also minimises an MMD, incrementally extending $\mathbf{Z}_m$ with the point that minimises the numerator of the second term of the BQ criterion, see equation (3.7). Since KH has smaller complexity than SBQ, and the results of the previous section suggest that optimal choice of $\mathbf{Z}_m$ is less important than correct determination of the weights $\mathbf{w}_i$, we compare now $\zeta^{\mathsf{BQ}\star}$ to two other validation measures, whose designs $\mathbf{Z}_m$ are found by extending $\mathbf{X}_n$ by KH: ζ_K^{KH}, that performs KH for kernel K, and $\zeta^{\mathsf{KH}\star}$ that considers kernel $\overline{K}_{|n}$. As we will see, the SBQ design is a superior alternative, both in terms of performance and numerical stability, to the KH designs.

Since ζ_K^{KH} considers only, at each step, measures with uniform weights, and $\zeta^{\mathsf{KH}\star}$ does not take into account the optimal weights that will be applied when $\mathbf{Z}_m$ is extended to $\mathbf{Z}_{m+1}$, we can expect the following ranking of these estimators:

$$\mathcal{R}(\zeta_K^{\mathsf{KH}}; \mathscr{F}_n) \geq \mathcal{R}(\zeta^{\mathsf{KH}\star}; \mathscr{F}_n) \geq \mathcal{R}(\zeta^{\mathsf{BQ}\star}; \mathscr{F}_n)\,, \tag{4.1}$$

which has already been remarked in [8].

Figures 8 and 9 plot, for $d = 1$ and $d = 2$, respectively, the MSE of estimators $\widehat{\mathsf{ISE}}$ that use $\zeta^{\mathsf{BQ}\star}$, $\zeta^{\mathsf{KH}\star}$ and ζ_K^{KH}. Kernels (Q and K) and designs sizes m are as in the previous examples, see the figures' captions. We can see that $\zeta^{\mathsf{BQ}\star}$ has virtually always smaller MSE than the validation

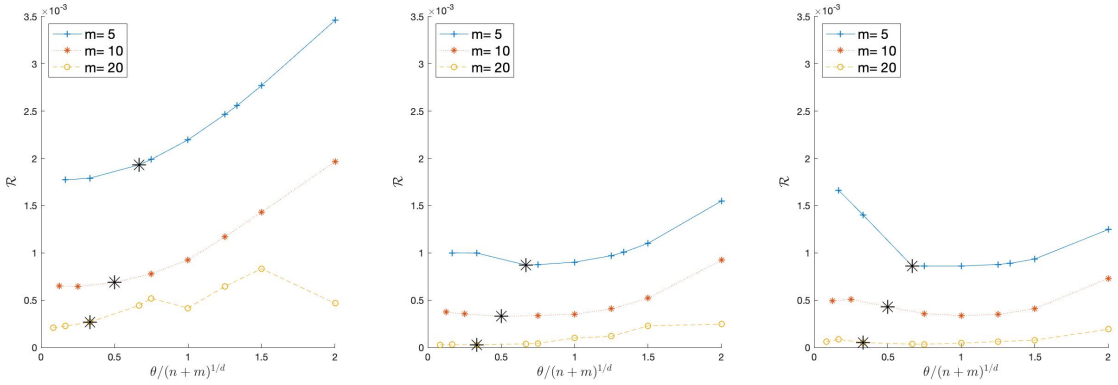

Figure 3: MSE of $\widehat{\mathsf{ISE}}$. Statistics over 500 realisations. Q is a Matérn 3/2 kernel; K is a Matérn kernel with $\nu = 1/2$ (left), $\nu = 3/2$ (middle) and $\nu = 5/2$ (right); $d = 1$.

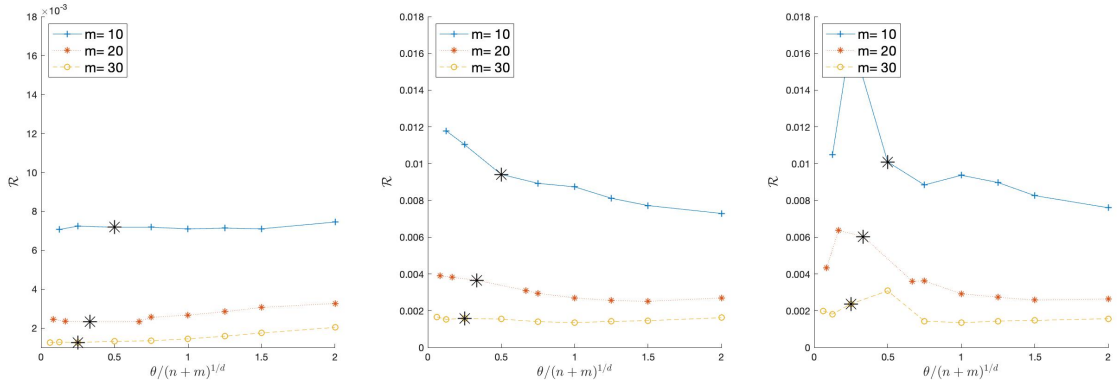

Figure 4: MSE of $\widehat{\mathsf{ISE}}$. Statistics over 500 realisations. Q is a Cauchy kernel, K is as in Figure 3; $d = 1$.

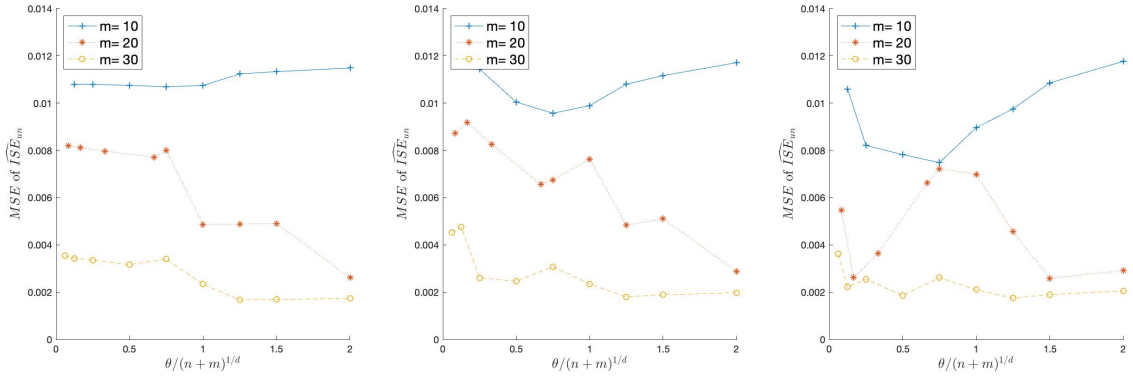

Figure 5: MSE of $\widehat{\mathsf{ISE}}_{un}$. Statistics over 500 realisations for the example in Figure 3

measures using validation designs $\mathbf{Z}_m$ found by KH, in particular for small design sizes m and the more regular models, and appears to be more robust with respect to the choice of the GP kernel. We remark that the design found by KH for kernel $\overline{K}_{|n}$, i.e., the validation measure $\zeta^{\mathsf{KH}\star}$ (in blue), often leads to the poorest performance. That use of $\overline{K}_{|n}$ may lead to worse performance than simply using K has already been noticed in [5], where only validation sets generated with KH were considered.

Our experiments reveal that the designs $\zeta^{\mathsf{KH}\star}$ can some-

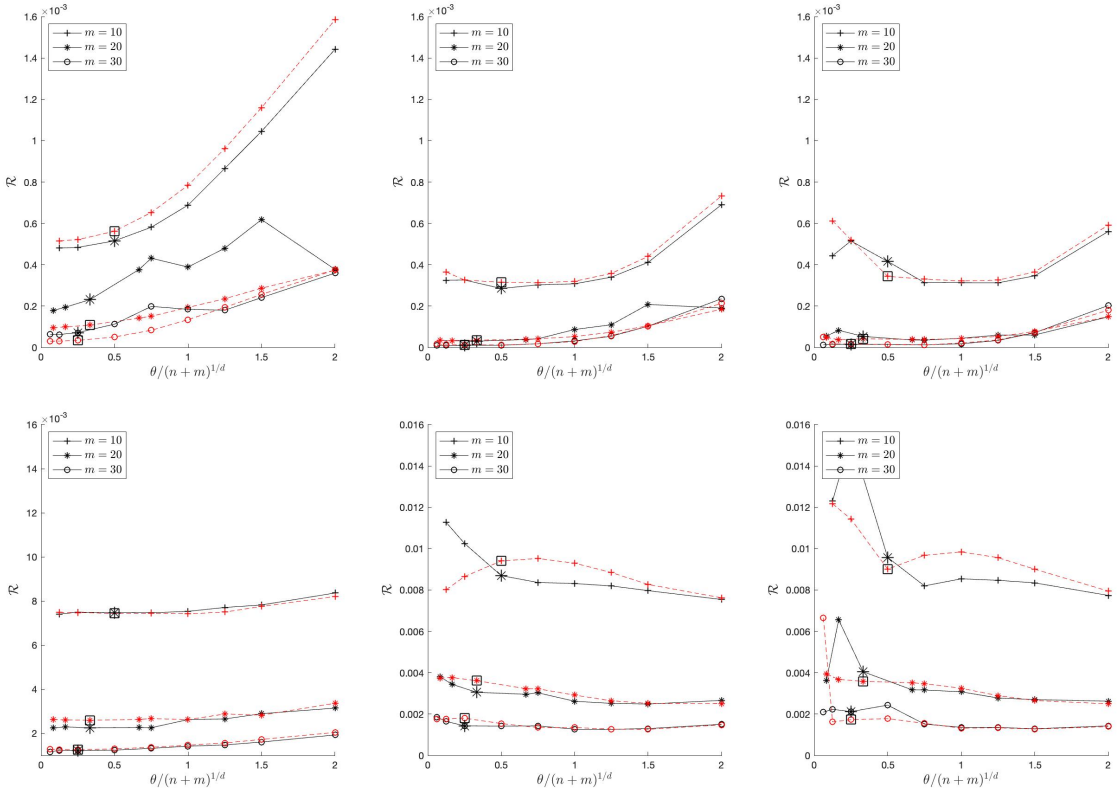

Figure 6: MSE of $\widehat{\mathsf{ISE}}$ for $\zeta^{\mathsf{BQ}\star}$ (black) and ζ_K^{BQ} (red) for $m \in \{10, 20, 30\}$. From left to right $\nu = 1/2, 3/2, 5/2$. Statistics over 500 realisations. Top: Q is the Matérn 3/2 kernel, bottom: Q is the Cauchy kernel; $d = 1$.

times lead to very large errors. This happens when KH places design points close $\mathbf{X}_n$, which are subsequently given very large weights. In fact, the implementation of standard KH for kernel $\overline{K}_{|n}$ needs careful handling of possible repetition of design points, as already noted in [13] where an algorithm is proposed to accommodate this eventuality. As shown there, this corresponds to situations where, when adding a new point to the current design $\mathbf{Z}_m$, the total mass of the optimal uniform measure over $\mathbf{Z}_{m+1}$ must be decreased. Since our implementation simply imposes $\mathbf{z}_{m+1} \notin (\mathbf{X}_n \cup \mathbf{Z}_m)$, a grid point very close to $\mathbf{X}_n \cup \mathbf{Z}_m$ is chosen in these situations, as shown below.

Figure 10 shows the designs $\mathbf{Z}_m$, $m = 10$, for Matérn kernels with $\theta = \theta_0$, and regularity parameter (top to bottom panels) $\nu = 1/2, 3/2$ and $5/2$. The vertical red lines indicate $\mathbf{X}_n$ and the black stars, blue circles and red squares the position of points of $\zeta^{\mathsf{BQ}\star}, \zeta^{\mathsf{KH}\star}$ and ζ^{KH}, respectively. A vertical offset is used to facilitate the visualisation of each design (from top to bottom, $\zeta_K^{\mathsf{KH}}, \zeta^{\mathsf{KH}\star}$ and $\zeta^{\mathsf{BQ}\star}$). Remark first that the SBQ designs are always space-filling continuations of $\mathbf{X}_n$, presenting a good stability with respect to ν, mainly moving points closer to the boundaries of $\mathscr{X}$ when ν increases. The other two designs place a few points in the vicinity of $\mathbf{X}_n$.

4.4 Properties of the design measures $\zeta^{\mathsf{BQ}\star}$

For the same set of kernels K and design sizes considered in Figure 3 (with $d = 1$ thus), we plot in Figure 11 the sum of the design weights, $S(\theta) = \sum_i \mathbf{w}_i(\theta)$, as a function of the (normalised) scale parameter of K. K is always a Matérn kernel, with regularity parameter $\nu = 1/2$, $3/2, 5/2$ (top to bottom), as indicated in the legends. The learning design $\mathbf{X}_n$ ($n = 10$) is the same for all cases.

Three values of m are considered, $m = 10$, 20 and 30 (blue, red and cyan curves, respectively). In each curve the black squares indicate the value $\theta_0 = n^{1/d}$. We can see that $S(\theta)$ increases with m. For θ larger than a certain value S becomes nearly constant and smaller than (note that the value of the scale parameter θ_c prescribed by our rule of thumb, which corresponds to the normalised value of θ equal to one) is always inside this interval) while for $\theta = n^{1/d}$, under the more regular model with a Matérn 5/2 kernel, S may be larger than 1.

Figures 12, 13 and 14 present the designs for three values of θ: $\theta = n^{1/d}$ (the value used in the simulations of Figure 3, and indicated by the squares in Figure 11), for the value prescribed by our rule of thumb, $\theta = (n+m)^{1/d}$, and for $\theta = 2\,(n+m)^{1/d}$, the upper limit considered in Figure 3. In the Figures the weights of ζ_m are shown multiplied m, to en-

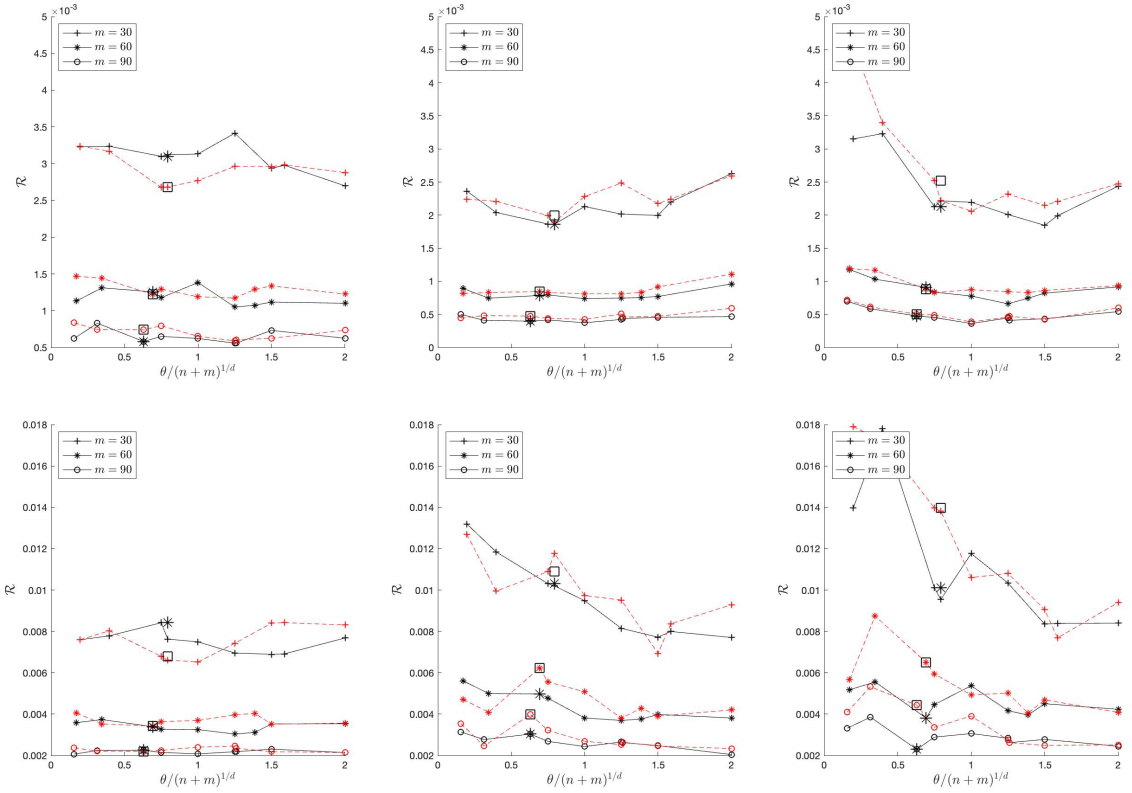

Figure 7: MSE of $\widehat{\mathsf{ISE}}$ for $\zeta^{\mathsf{BQ}\star}$ (black) and ζ_K^{BQ} (red) for $m \in \{20, 40, 60\}$. Top: Q is a Matérn 3/2 kernel; bottom: Q is the Cauchy kernel. K is always a Matérn kernel, from left to right $\nu = 1/2, 3/2, 5/2$. $d = 3$

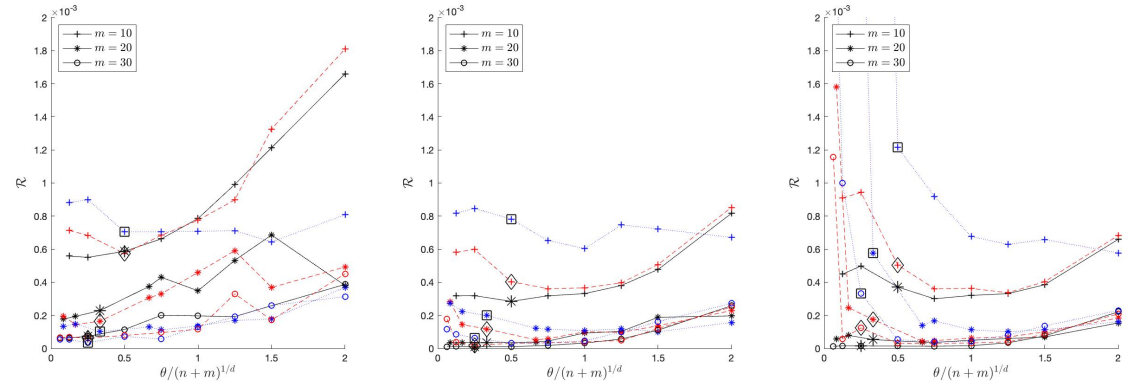

Figure 8: MSE of $\widehat{\mathsf{ISE}}$ for $\zeta^{\mathsf{BQ}\star}$ (black solid lines), ζ_K^{KH} (red dashed lines) and $\zeta^{\mathsf{KH}\star}$ (dotted blue lines), for $m = 10$ (+), $m = 20$ ($\star$) and $m = 30$ ($\circ$). From left to right $\nu = 1/2, 3/2, 5/2$. Q is a Matérn 3/2 kernel; $d = 1$.

able comparison. The distinct kernels K correspond to the three panels, as indicated in the Figure (regularity increasing from top to bottom). The dotted black vertical lines (the same in the three panels) indicate the learning design $\mathbf{X}_n$. The colours code the validation design size: $m = 10$ in blue, $m = 20$ in red and $m = 30$ in cyan. Remark the striking similarity of the validation measures obtained for the different kernels in Figure 13 and 14, supporting our observations concerning the robustness of the estimator. The Figures also show that the validation designs are, as expected, space filling continuations of $\mathbf{X}_n$, and that as m grows (remember $\mathbf{Z}_{10} \subset \mathbf{Z}_{20} \subset \mathbf{Z}_{30}$) the holes of $\mathbf{X}_n \cup \mathbf{Z}_m$ are refined. Note, however, the slow rate of population of the immediate neighborhood of $\mathbf{X}_n$, $\mathbf{Z}_m$ tending first, as m grows, to refine the

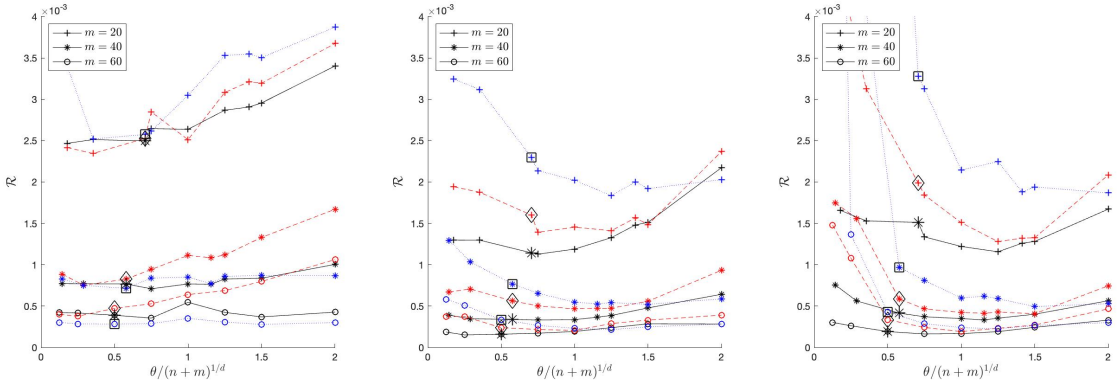

Figure 9: MSE of $\widehat{\mathsf{ISE}}$ for $\zeta^{\mathsf{BQ}\star}$ (black solid lines), ζ_K^{KH} (red dashed lines) and $\zeta^{\mathsf{KH}\star}$ (dotted blue lines), for $m = 10$ ($+$), $m = 20$ ($\star$) and $m = 30$ ($\circ$). From left to right $\nu = 1/2, 3/2, 5/2$. Q is a Matérn 3/2 kernel; $d = 2$.

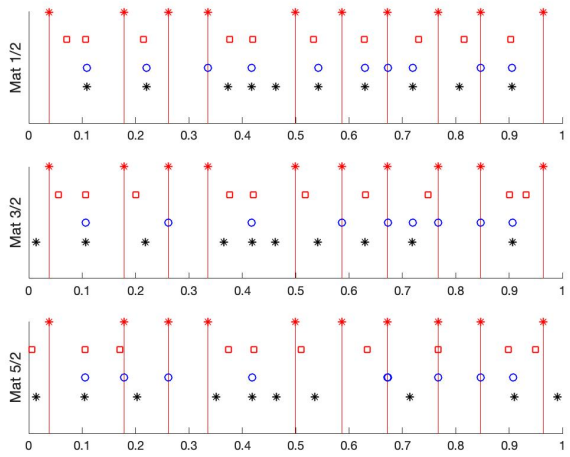

Figure 10: Designs for $\theta = \theta_0$ in Figure 8. From top to bottom: $\nu = 1/2, 3/2, 5/2$. $\mathbf{X}_n$: red $*$; $\mathbf{Z}_m^{\mathsf{BQ}\star}$: black $*$; $\mathbf{Z}_m^{\mathsf{KH}\star}$: blue $\circ$ and $\mathbf{Z}_m^{\mathsf{KH}}$: red $\square$.

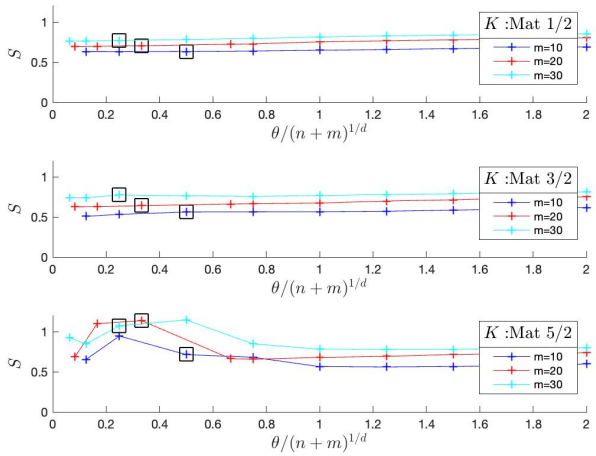

Figure 11: $S = \sum_i \mathbf{w}_i$ for measure $\zeta^{\mathsf{BQ}\star}$, $n = 10$.

interior of the wider holes of $\mathbf{X}_n$. For the Matérn 5/2 kernel and $\theta = n^{1/d}$ a few weights, corresponding to validation points close to $\mathbf{X}_n$, become very large, see Figure 12, explaining that $S(\theta)$ may be larger than one on the bottom panel of Figure 11. Analysis of the validation measures obtained assuming the larger value of θ in Figure 14 shows that as the assumed correlation length increases $\zeta^{\mathsf{BQ}\star}$ tends to a uniform measure, all weights having now a similar value. We note that even in this situation, use of the BQ measure, which down-weights the squared residuals, leads to a smaller error than use of the simple uniform measure over $\mathbf{Z}_m$, as the comparison of Figures 3 and 5 in Section 4.1 has shown.

5. CONCLUSIONS

The paper presents an estimator for the ISE of an interpolator based on knowledge of the design on which it has been learned, defined as the ISE for a finitely supported

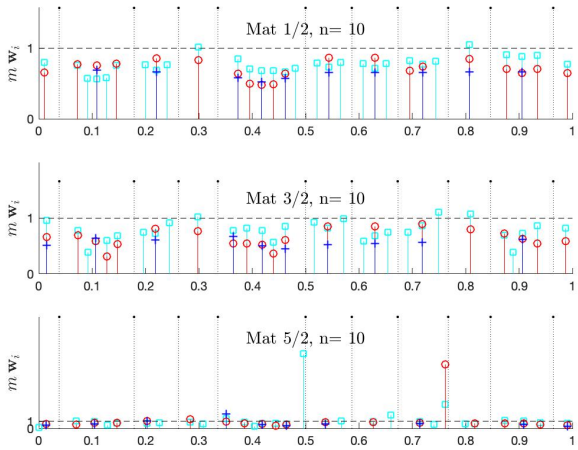

Figure 12: $\zeta^{\mathsf{BQ}\star}$ in Figure 11, $\theta = n^{1/d}$.

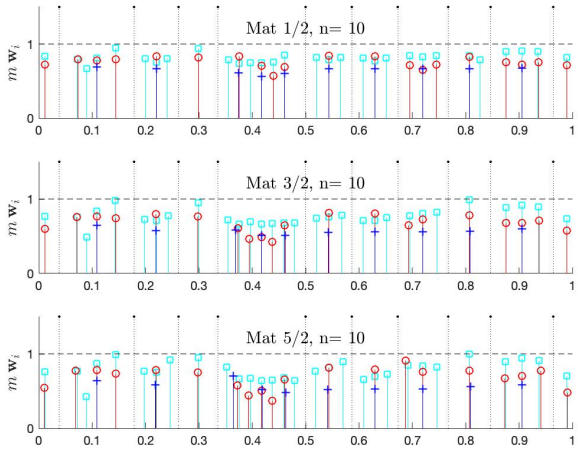

Figure 13: $\zeta^{\mathsf{BQ}\star}$ in Figure 11, $\theta = (n+m)^{1/d}$.

validation measure. The estimator proposed is the optimal MSE linear estimator under the assumption that the interpolated function is a realisation from a Gaussian process with known statistical moments. The support and weights of the validation measure are found by minimising an MMD for a non-stationary kernel that is adapted to the learning design, and a nested sequence of validation designs is greedily determined by SBQ. A default rule is proposed to select the covariance kernel of the assumed model.

The interpretation of the ISE estimator in terms of an interpolation of the squared residuals explains the utmost importance of accounting for the correct shape of their second order moment. Moreover, it unriddles the observed robustness of the estimator with respect to the covariance of the assumed GP model.

The work presented suggests several directions for future

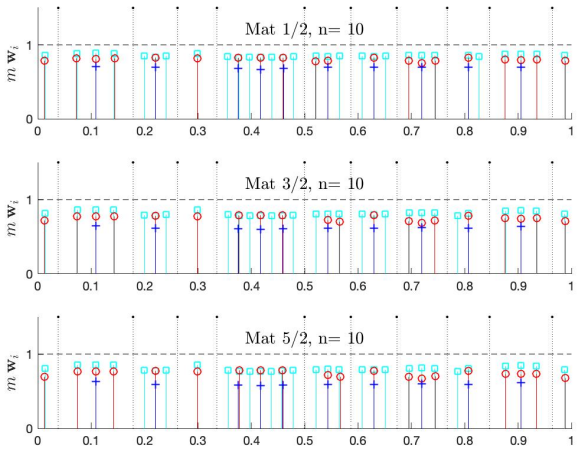

Figure 14: $\zeta^{\mathsf{BQ}\star}$ in Figure 11, $\theta = 2\,(n+m)^{1/d}$.

developments. One concerns the determination of indicators of the quality of the ISE estimate itself, ideally given by the risk function that is optimised. These could both be used to define stopping rules, indicating that incorporation of further residual observations should not yield a significant improvement on the confidence of the current ISE estimate, or to flag poor performance of the current interpolator, and trigger its update including some of the residuals observed over $\mathbf{Z}_m$ in the learning dataset $\mathscr{F}_n$. A major difficulty is related to the dependency of the MSE of the interpolator on the assumed process covariance, which is known to be difficult to estimate. A possible source of suboptimality of the estimator presented concerns the restriction to a linear estimator. The extension to more general estimators while preserving at the same time the robustness property of the method forms a challenging objective. Finally, we believe that the analysis presented here suggests possible approaches to define (down-)weighted CV estimators with better performance than standard ones.

ACKNOWLEDGEMENTS

The authors acknowledge the fruitful collaboration with the other partners of the ANR project INDEX, in particular Bertrand Iooss, Elias Fekhari and Joseph Muré from EDF R&D Chatou, France.

FUNDING

This work was partially funded by project ANR INDEX (ANR-18-CE91-0007)[5].

[5]https://sdb3.i3s.unice.fr/anrindex/

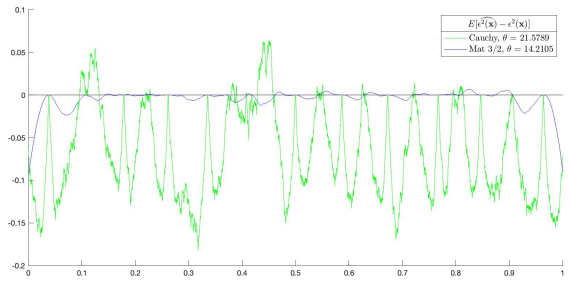

Figure 15: Bias of $\widehat{\varepsilon^2}$, $\mathbf{x} \in \mathscr{X}$. (simulations from Cauchy and Matérn 3/2 kernels).

APPENDIX A. BIAS CORRECTION

Under the assumed GP model for $f_{|\mathscr{F}_n}$ estimator $\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n}, \mathbf{Z}_m)$ has a non-zero bias:

$$\begin{aligned} B(\mathbf{Z}_m) &= E\left[\widehat{\mathsf{ISE}}(\eta_{\mathscr{F}_n}, \mathbf{Z}_m) - \mathsf{ISE}(\eta_{\mathscr{F}_n}) \middle| \mathscr{F}_n\right] \\ &= \sigma^2 \tilde{\mathbf{w}}(\mathbf{Z}_m)^T \mathbf{k}_{|n}(\mathbf{Z}_m) - \mathsf{IMSE}^\star(\mathbf{X}_n)\,. \end{aligned}$$

with $\mathbf{k}_{|n}(\mathbf{Z}_m)$ the m-dimensional column vector with components $[k_{|n}]_i = K_{|n}(\mathbf{z}_i, \mathbf{z}_i)$ (see equation (2.3)), and $\mathsf{IMSE}^\star(\mathbf{X}_n)$ given by (2.4).

By noting that $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)$ is also the optimal MMSE estimator under the zero mean model $\mathcal{GP}^0 = \mathcal{GP}(0, \overline{K}_{|n})$ (necessarily linear, since the model is Gaussian), equation (3.6) suggests that $B(\mathbf{Z}_m)$ may be negative: $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)$ being optimal for $\mathcal{GP}^0 = \mathcal{GP}(0, \overline{K}_{|n})$, it should tend to have smaller values than estimators that consider the correct first posterior moment, i.e., $\mathcal{GP}(\sigma^2 K_{|n}(\mathbf{x}, \mathbf{x}), \sigma^4 \overline{K}_{|n}(\mathbf{x}, \mathbf{x}'))$. Figure 15 displays the bias $(1/M)\sum_i \left((\widehat{\varepsilon^2}_{\mathscr{F}_n})^{(i)}(\mathbf{x}|\mathbf{Z}_m) - (\varepsilon^2)^{(i)}(\mathbf{x})\right)$ observed over $M = 500$ realisations from several GP models, supporting this conjecture (simulations are from the models considered in Figure 1).

Simply subtracting $B(\mathbf{Z}_m)$ from the biased linear estimator yields the following unbiased affine estimator

$$\begin{aligned} \widehat{\mathsf{ISE}}_{affine}(\mathbf{X}_n, \mathbf{Z}_m) &= \sum_{i=1}^{m} \tilde{\mathbf{w}}_i \varepsilon^2(\mathbf{z}_i) - B(\mathbf{Z}_m) \\ &= \mathsf{IMSE}^\star(\mathbf{X}_n) + \tilde{\mathbf{w}}^T \Delta_m\,. \end{aligned} \tag{A.1}$$

where Δ_m collects the mean corrected squared residuals at the validation points: $\Delta_m(\mathbf{z}_i) = \varepsilon^2(\mathbf{z}_i) - \sigma^2 K_{|n}(\mathbf{z}_i, \mathbf{z}_i), i = 1, \ldots, m$.

Alternatively, a linear (instead of affine) unbiased solution can be found by using weights $\mathbf{w}(\mathbf{Z}_m)$ that minimise the same quadratic cost function, but under the zero bias constraint. This leads to the following additive correction of the optimal weights of the biased linear estimator $\widehat{\mathsf{ISE}}$:

$$\mathbf{w}_{linear}(\mathbf{Z}_m) = \tilde{\mathbf{w}}(\mathbf{Z}_m) - \frac{\sigma^2 \tilde{\mathbf{w}}(\mathbf{Z}_m)\mathbf{k}_{|n}(\mathbf{Z}_m) - \mathsf{IMSE}^\star(\mathbf{X}_n)}{\sigma^2 \mathbf{k}_{|n}(\mathbf{Z}_m)^T \mathbf{t}} \mathbf{t}\,, \tag{A.2}$$

where $\mathbf{t} = \overline{K}_{|n}(\mathbf{Z}_m, \mathbf{Z}_m)^{-1} \mathbf{k}_{|n}(\mathbf{Z}_m)$.

Denote by $\widehat{\mathsf{ISE}}_{biased}$ the empirical ISE estimator that uses the validation measure presented in section 3.2, and let $\widehat{\mathsf{ISE}}_{linear}$ denote the linear unbiased estimator with weights given by (A.2).

As (A.2) is linear and $\widehat{\varepsilon^2}_{\mathscr{F}_n}(\mathbf{x}|\mathbf{Z}_m)$ in (3.5) is the MMSE linear estimate of $\varepsilon^2(\mathbf{x})$ given $\mathscr{F}_n$, $\widehat{\mathsf{ISE}}_{linear}$ will necessarily perform worse than $\widehat{\mathsf{ISE}}_{biased}$ when using the correct model for f. Similarly, $\widehat{\mathsf{ISE}}_{affine}$ performs better than $\widehat{\mathsf{ISE}}_{biased}$ for the right model for f. In fact, the numerical experiments presented below show that $\widehat{\mathsf{ISE}}_{linear}$ and $\widehat{\mathsf{ISE}}_{affine}$ have both bad performance and poor robustness: as both estimators explicitly incorporate the uncertainty predicted by the posterior distribution, they inherit, as we will see, the well known sensitivity of modelled prediction uncertainty with respect to the assumed model.

We performed $M = 500$ simulations from a GP with kernel $Q(\cdot, \cdot; \theta_0)$, the Matérn kernel with regularity parameter $\nu = 3/2$ and $\theta_0 = n$, over domain $\mathscr{X} = [0, 1]^d$ $(d = 1)$. The corresponding optimal Bayesian interpolators $\eta^{(i)}_{\mathscr{F}_n}$ all use the same learning design $\mathbf{X}_n$ of size $n = 10$ (see details in Appendix C).

Let $\widehat{\mathsf{ISE}}^{(i)}_c(K_\theta)$, $c \in \{biased, affine, linear\}$ denote the estimate $\widehat{\mathsf{ISE}}_c(\eta^{(i)}_{\mathscr{F}_n})$ when the validation measure assumes kernel $K(\cdot, \cdot; \theta)$. Figures 16 and 17 plot the average of these estimates. In Figure 16 $K = Q$, while in Figure 17 measures ζ_m are based on Matérn kernels with $\nu \in \{1/2, 5/2\}$. In both figures the horizontal dashed black lines indicate $\overline{\mathsf{ISE}}$, the empirical average of $\mathsf{ISE}^{(i)}(\eta_{\mathscr{F}_n})$ over the M realisations, and the solid black curve is $\mathsf{IMSE}^\star(\mathbf{X}_n; \theta)$, predicted by kernel $K(\cdot, \cdot; \theta)$. The three panels correspond to increasing design sizes $m = 5, 10, 20$ (from left to right).

The correct θ_0 can be identified in Figure 16 as the value at which the black solid and dashed lines intersect: $\mathsf{IMSE}^\star(\mathbf{X}_n; \theta_0) = \overline{\mathsf{ISE}}$. We can see that for the correct parameter value the unbiased estimates (red and green curves) both have the correct mean, while the biased estimator (blue line) has, as foreseen, a negative bias. For $\theta \neq \theta_0$ all three estimators have a non-zero bias, which decreases when m grows as the estimators become less dependent on the prior stochastic model for f. For large design sizes, the two linear estimates (blue and green curves) have nearly the same bias, showing that bias correction is mainly relevant for small validation designs. As anticipated, the unbiased estimates display a larger sensitivity with respect to model mismatch than the original $\widehat{\mathsf{ISE}}_{biased}(\theta)$, which displays a remarkably stable behaviour with respect to θ.

In Figure 17 wrong values of ν of $K(\cdot, \cdot)$ are assumed by the design algorithm. In the top row $\nu = 1/2$, less regular than Q, while in the bottom row $\nu = 5/2$, more regular than

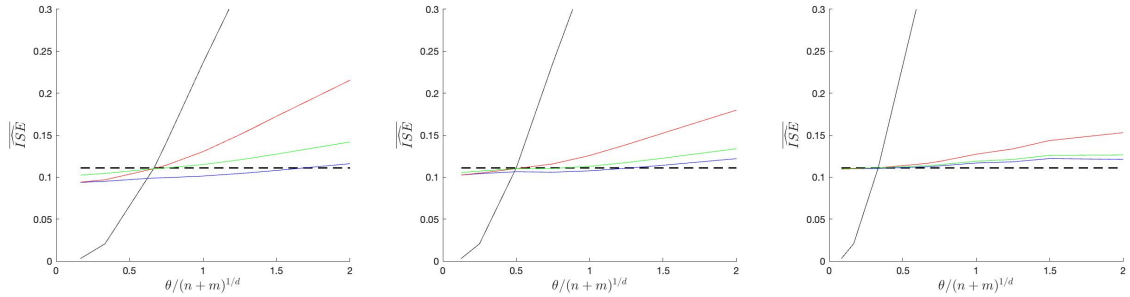

Figure 16: $\mathsf{IMSE}^{\star}(\mathbf{X}_n;\theta)$ (black), $\overline{\mathsf{ISE}}$ (dashed), $\widehat{\overline{\mathsf{ISE}}}_{biased}(\theta)$ (blue), $\widehat{\overline{\mathsf{ISE}}}_{affine}(\theta)$ (red) and $\widehat{\overline{\mathsf{ISE}}}_{linear}(\theta)$ (green). From left to right $m = 5, 10, 20$. Q and K are the Matérn 3/2 kernel.

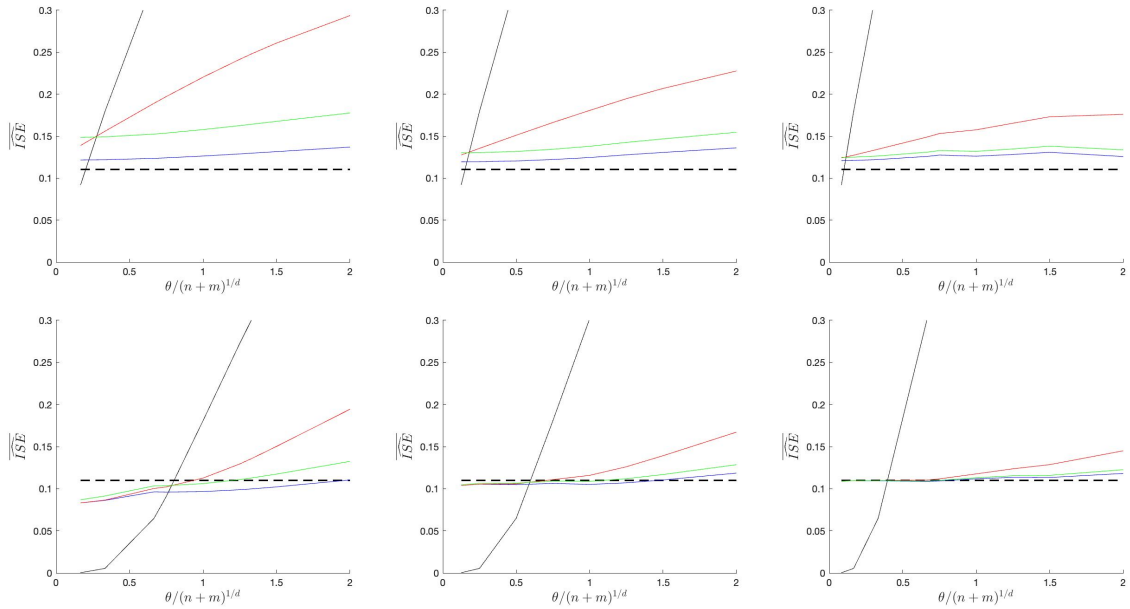

Figure 17: $\mathsf{IMSE}^{\star}(\mathbf{X}_n;\theta)$ (black), $\overline{\mathsf{ISE}}$ (dashed), $\widehat{\overline{\mathsf{ISE}}}_{biased}(\theta)$ (blue), $\widehat{\overline{\mathsf{ISE}}}_{affine}(\theta)$ (red), $\widehat{\overline{\mathsf{ISE}}}_{linear}(\theta)$ (green). From left to right $m = 5, 10, 20$. Q is the Matérn 3/2 kernel; K is the Matérn kernel with parameter $\nu = 1/2$ (top) and $\nu = 5/2$ (bottom).

the simulated model. While a much larger bias is observed for the exponential ($\nu = 1/2$) model in the top row, the curves in the bottom panels are similar to those in Figure 16), indicating that the estimator can accommodate a model that assumes a higher regularity. The robustness of BQ with respect to models assuming higher regularity than the true one has been previously noted in [9]. Finally, remark that $\widehat{\mathsf{ISE}}_{biased}$ has a remarkably stable behaviour, and that its bias is often the smallest amongst all three estimators.

Unless a high confidence can be given to the assumed GP model, including its scale parameter, the lack of robustness of the unbiased estimators prevents their use. For small design sizes, where bias correction could indeed be important, guaranteeing the fidelity of the assumed model is in general impossible, severely limiting the practical interest of the unbiased estimators discussed here.

APPENDIX B. POTENTIAL $P_{\overline{K}_{|n}}(\mathbf{z})$ FOR TENSOR-PRODUCT KERNELS ON $[0,1]^d$

B.1 Factorisation in the general case

A key difficulty for the algorithmic construction of a validation design by SBQ (see Section 3.2) or KH (see Section 4.3) is the calculation of $P_{\overline{K}_{|n}}(\mathbf{x}) = P_{\overline{K}_{|n},\mu}(\mathbf{x})$ for many $\mathbf{x}$ in order to choose $\mathbf{z}_{m+1}$. However, when K is a tensor-product kernel, $P_{\overline{K}_{|n},\mu}$ can be calculated explicitly.

Since μ is uniform on $\mathscr{X} = [0,1]^d$, we can write $\mu(\mathrm{d}\mathbf{x}) = \prod_{i=1}^{d} \mu_1(\mathrm{d}x_i)$ with μ_1 the uniform measure on $[0,1]$ and $\mathbf{x} =$

$(x_1,\ldots,x_d)^\top$. When $K(\mathbf{x},\mathbf{x}') = \prod_{i=1}^d K_i(x_i,x_i')$, with $\mathbf{x} = (x_1,\ldots,x_d)^\top$ and $\mathbf{x}' = (x_1',\ldots,x_d')^\top$, we have

$$P_{K,\mu}(\mathbf{x}) = \prod_{i=1}^d \int_{\mathscr{X}_i} K_i(x_i,x_i')\,\mu_1(\mathrm{d}x_i') = \prod_{i=1}^d P_{K_i,\mu_1}(x_i)\,.$$

One may refer to [18] for connections between positive-definiteness properties of the K_i and those of K. The expression of $P_{K_i,\mu_1}(\cdot)$ is available for many kernels K_i; see [14] and the references therein.

Before deriving the expression of $P_{\overline{K}_{|n},\mu}(\mathbf{x})$ we introduce some notation. Denote by $\overline{\mathbf{\Omega}}_{K,n}$ the $n\times n$ matrix with respective elements

$$\{\overline{\mathbf{\Omega}}_{K,n}\}_{j,k} = \prod_{i=1}^d \beta_{K_i}(x_{j\,i},x_{k\,i})\,,$$

and by $\overline{\omega}_{K,n}(\mathbf{x})$ the vector with j-th component

$$\{\overline{\omega}_{K,n}(\mathbf{x})\}_j = \prod_{i=1}^d \beta_{K_i}(x_{j\,i},x_i)\,,$$

where $x_{j\,i}$ (respectively, $x_{k\,i}$) is the i-th component of $\mathbf{x}_j$ (respectively, $\mathbf{x}_k$), and

$$\beta_{K_i}(r,s) \;=\; \int_{\mathscr{X}} K_i(r,t)K_i(s,t)\,\mu_1(\mathrm{d}t)\,,\;\; i=1,\ldots,d\,.$$

Then, using (2.6), direct calculation gives

$$\begin{aligned} P_{\overline{K}_{|n},\mu}(\mathbf{x}) \;&=\; 2\,P_{K^2,\mu}(\mathbf{x}) - 4\,\mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\overline{\omega}_{K,n}(\mathbf{x}) \\ &\quad +2\,\mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\overline{\mathbf{\Omega}}_{K,n}\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x}) \\ &\quad +\left[1-\mathbf{k}_n^\top(\mathbf{x})\mathbf{K}_n^{-1}\mathbf{k}_n(\mathbf{x})\right]\left[1-\mathrm{trace}(\mathbf{K}_n^{-1}\overline{\mathbf{\Omega}}_{K,n})\right]\,. \end{aligned}$$

The expressions of $P_{K^2,\mu_1}(x)$ and $\beta_K(u,v)$, $x,u,v\in[0,1]$, for μ_1 uniform on $[0,1]$ and $K_i(x,x')$ a Matérn 3/2 kernel (C.1) are given in Section B.2, making the expression of $P_{\overline{K}_{|n},\mu}(\mathbf{x})$ available in closed form when $K(\mathbf{x},\mathbf{x}')$ is the product of uni-dimensional Matérn 3/2 kernels and μ is uniform on $\mathscr{X}=[0,1]^d$. Similar calculations can be conducted for other kernels. The expression of $\mathscr{E}_{\overline{K}_{|n}}(\mu)$, which appears in the expansion of $\mathcal{R}(\zeta_m,\mathbf{X}_n)$, see (2.5), can be obtained in closed form in a similar way; see [13].

B.2 The Matérn 3/2 case

When $K_i(x,x') = K^{3/2}_{\text{Matérn}}(|x-x'|)$ given by (C.1) with $\theta = \gamma/\sqrt{3}$, we have [6]

$$P_{K_i,\mu_1}(x) = S_\gamma(x) + S_\gamma(1-x)\,,$$

with $S_\gamma(x) = \frac{1}{\gamma}\left[2-(2+\gamma x)\mathsf{e}^{-\gamma x}\right]$, $x\in[0,1]$. Straightforward but lengthy calculation gives

$$P_{K_i^2,\mu_1}(x) = T_\gamma(x) + T_\gamma(1-x)\,,$$

with $T_\gamma(x) = \frac{1}{4\,\gamma}\left[5-(5+6\,\gamma x+2\,\gamma^2x^2)\mathsf{e}^{-2\,\gamma x}\right]$, $x\in[0,1]$. Also, the expressions $\beta_{K_i}(u,v) = B_\gamma(u,v) - C_\gamma(u,v) - C_\gamma(1-u,1-v)$, $u,v\in[0,1]$, with

$$\begin{aligned} B_\gamma(u,v) \;&=\; \frac{\mathsf{e}^{-\gamma|u-v|}}{6\,\gamma}\left[15\,(1+\gamma|u-v|)+6\,\gamma^2|u-v|^2\right. \\ &\quad \left.+\gamma^3|u-v|^3\right]\,, \\ C_\gamma(u,v) \;&=\; \frac{\mathsf{e}^{-\gamma(u+v)}}{4\,\gamma}\left[5+3\,\gamma(u+v)+2\,\gamma^2uv\right]\,, \end{aligned}$$

permit to calculate $P_{\overline{K}_{|n},\mu}(\mathbf{x})$ explicitly.

APPENDIX C. DETAILS ON NUMERICAL EXPERIMENTS

C.1 GP models

Let $\{f(\mathbf{x})\}_{\mathbf{x}\in\mathscr{X}}$, $f(\mathbf{x})\in\mathbb{R}$ be a real d-dimensional stochastic process defined over the compact index set $\mathscr{X}\subset\mathbb{R}^d$. $\{f(\mathbf{x})\}_{\mathbf{x}\in\mathscr{X}}$ is a Gaussian process with mean function $\mu(\cdot)$ and covariance kernel $K(\cdot,\cdot)$, noted $\{f(\mathbf{x})\}_{\mathbf{x}\in\mathscr{X}}\sim\mathcal{GP}(\mu(\cdot),K((\cdot,\cdot)$, if for any finite $n\in\mathbb{N}$, and any $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}\subset\mathscr{X}$, the collection of random variables $\{f(\mathbf{x}_i),i=1,\ldots,n\}$ is a n-dimensional normal random vector, i.e.

$$\{f(\mathbf{x}_i),i=1,\ldots,n\}\sim\mathcal{N}(\mu_{\mathbf{X}},\mathbf{K}_{\mathbf{X}})\;\;,$$

where $\mu_{\mathbf{X}}\in\mathbb{R}^d$ has i-th component $[\mu_{\mathbf{X}}]_i = \mu(\mathbf{x}_i)$, and the $n\times n$ matrix $\mathbf{K}_{\mathbf{X}}$ has generic (i,j) element $\left[\overline{K}_{\mathbf{X}}\right]_{(i,j)} = K(\mathbf{x}_i,\mathbf{x}_j)$.

In Section 4 we assume that $\{f(\mathbf{x})\}_{\mathbf{x}\in\mathscr{X}}\sim\mathcal{GP}(\mu(\cdot),K(\cdot,\cdot))$. All Gaussian models considered in the numerical experiments presented assume a zero mean, i.e., $\mu(\cdot)\equiv 0$, and are defined over $\mathscr{X}=[0,1]^d$. Besides, only stationary and isotropic processes are considered, i.e., all covariance kernels K satisfy $K(\mathbf{x},\mathbf{x}')=\Psi(\mathbf{x}-\mathbf{x}')=\psi\left(\|\mathbf{x}-\mathbf{x}'\|\right)$.

The experiments presented resort to several parametric families for the process kernel K, namely, the Cauchy kernel K_{Cauchy} as defined in [7], and the Matérn kernels $K^\nu_{\text{Matérn}}$ with regularity parameter $\nu\in\{1/2,3/2,5/2\}$, as defined below. For all kernels $\theta\in\mathbb{R}^+$ is the scale parameter, and for the Cauchy kernels (ρ,γ) are the long distance dependency and the shape parameters, respectively. Below, $\ell=\|\mathbf{x}-\mathbf{x}'\|$.

$$\begin{aligned} &\psi_{\text{Cauchy}}(\ell) = (1+(\theta\,\ell)^\gamma)^{-\rho/\gamma}\;\;, \\ &\psi^{1/2}_{\text{Matérn}}(\ell) = e^{-\theta\,\ell}\;, \\ &\psi^{3/2}_{\text{Matérn}}(\ell) = \left(1+\sqrt{3}\theta\,\ell\right)\,e^{-\sqrt{3}\theta\,\ell}\,, \\ &\psi^{5/2}_{\text{Matérn}}(\ell) = \left(1+\sqrt{5}\theta\,\ell+\frac{5}{3}(\theta\,\ell)^2\right)\,e^{\sqrt{5}\theta\,\ell}\,. \end{aligned} \tag{C.1}$$

For the Cauchy kernel, we set $\rho=\gamma=1$, and thus a rational kernel with bandwidth determined by θ.

The parameter θ_0 of the simulated GP model is dependent of the size of the learning design of $\mathscr{F}_n$: $\theta_0 = n^{1/d}$. This will guarantee the numerical stability of the KH algorithm used to define $\mathbf{X}_n$ (see below), and that the interpolator $\eta_{\mathscr{F}_n}$ will have a moderate error level.

C.2 Sampling from a GP processes

The material in Section 4 presents the average performance of the ISE estimators over $M = 500$ simulations from an assumed GP model. These simulations are supported in a dense finite subset of $\mathscr{X}$ of size $L = 2^{12}$: $\mathscr{X}_L \subset \mathscr{X}$. If $d = 1$ $\mathscr{X}_L$ is an uniform grid, and, for $d \geq 2$, a scrambled low-discrepancy Sobol sequence.

Generation of realisations from the GP model requires factorisation of the matrix $\mathbf{K}_{\mathscr{X}_L}$ collecting the values of kernel K over the pairs of points of $\mathscr{X}_L$:

$$f^{(i)}(\mathbf{t}) = \mathbf{K}^{-1/2}\mathbf{u}\,, \qquad \mathbf{u} \sim \mathcal{N}\left(0, I_{|\mathcal{T}|}\right)\,, \qquad \mathbf{t} \in \mathcal{T}\,.$$

When L is very large this may lead to numerical instabilities for some parameter values, due to near singularity of $\mathbf{K}$. In that case, our simulated signals are the optimal MSE estimate (under the simulated $\mathcal{GP}$) of samples obtained as above over a smaller dense subset $\mathscr{X}_M$ of $\mathscr{X}$, of size $M = 10^3\, d$:

$$\{\mathbf{u}^{(i)}(\mathbf{x})\}_{\mathbf{x}\in\mathscr{X}_M} \sim \mathcal{N}\left(0, I_M\right) \longrightarrow \{f^{(i)}(\mathbf{x})\}_{\mathbf{x}\in\mathscr{X}_M} \to \{\hat{f}^{(i)(\mathbf{x})}\}_{\mathbf{x}\in\mathscr{X}}\,.$$

The simulated functions are thus slightly smoother than the actual realisations from the assumed GP. We believe, however that this does not compromise the validity of our conclusions.

C.3 Learning design X_n, Interpolator $\eta_{\mathscr{F}_n}$, ISE estimates

In Section 4, for each GP kernel K and design size n, $\mathbf{X}_n$ is always the space-filling design obtained by standard KH for kernel K. For each realisation $f^{(i)}$, its interpolator $\eta_{\mathscr{F}_n^{(i)}}$ is the optimal interpolator for the assumed GP model using the learning data $\mathscr{F}_n^{(i)} = (\mathbf{X}_n, f^{(i)}(\mathbf{X}_n))$,

$$\eta_{\mathscr{F}_n^{(i)}}(\mathbf{x}) = \mathbf{k}_{|n}(\mathbf{x}, \mathbf{X}_n)^T K_{|n}(\mathbf{X}_n, \mathbf{X}_n)^{-1} f^{(i)}(\mathbf{X}_n)\,.$$

Simulated residuals are thus $\varepsilon^{(i)}(\mathbf{x}) = f^{(i)}(\mathbf{x}) - \eta_{\mathscr{F}_n^{(i)}}(\mathbf{x})$. For a validation measure $\zeta_m = (\mathbf{w}, \mathbf{Z}_m)$ the MSE of the corresponding estimate $\widehat{\mathsf{ISE}}$ is approximated as

$$\hat{\mathcal{R}}(\zeta) = \frac{1}{M}\sum_{i=1}^{M}\left(\widehat{\mathsf{ISE}}^{(i)} - \mathsf{ISE}^{(i)}\right)^2\,,$$

where

$$\widehat{\mathsf{ISE}}^{(i)} = \sum_{i=i}^{m} \mathbf{w}_i \varepsilon^2_{\eta_{\mathscr{F}_n^{(i)}}}(\mathbf{z}_i), \qquad \mathsf{ISE}^{(i)} = \frac{1}{L}\sum_{\mathbf{t}_i\in\mathscr{X}_L} (\varepsilon^{(i)}(\mathbf{t}_i))^2\,.$$

REFERENCES

[1] F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proc. 29th Annual International Conference on Machine Learning*, pages 1355–1362, 2012.

[2] F. Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. 66:55–69, 2013.

[3] O. Dubrule. Cross validation of kriging in a unique neighborhood. *Journal of the International Association for Mathematical Geology*, 15(6):687–699, 1983.

[4] V.V. Fedorov. *Theory of Optimal Experiments*. New York, 1972.

[5] E. Fekhari, B. Iooss, J. Muré, L. Pronzato, and J. Rendas. Model predictivity assessment: incremental test-set selection and accuracy evaluation. In N. Salvati, C. Perna, S. Marchetti, and R. Chambers, editors, *Studies in Theoretical and Applied Statistics, SIS 2021, Pisa, Italy, June 21-25*. Springer, to appear, 2022. Preprint hal-03523695.

[6] D. Ginsbourger, O. Roustant, D. Schuhmacher, N. Durrande, and N. Lenz. On ANOVA decompositions of kernels and Gaussian random field paths. In *Monte Carlo and Quasi-Monte Carlo Methods*, pages 315–330. Springer, 2016.

[7] T. Gneiting and M. Schlather. Stochastic models that separate fractal dimension and the hurst effect. *SIAM Review*, 46(2):269–282, 2004.

[8] F. Huszár and D. Duvenaud. Optimally-weighted herding is Bayesian quadrature. In *Uncertainty in Artificial Intelligence*, pages 377–385, 2012.

[9] M. Kanagawa, B.K. Sriperumbudur, and K. Fukumizu. Convergence guarantees for kernel-based quadrature rules in misspecified settings. In *Advances in Neural Information Processing Systems*, pages 3288–3296, 2016.

[10] S. Mak and V.R. Joseph. Support points. *The Annals of Statistics*, 46(6A):2562–2592, 2018.

[11] A. O'Hagan. Bayes–hermite quadrature. *Journal of statistical planning and inference*, 29(3):245–260, 1991.

[12] L. Pronzato. Performance analysis of greedy algorithms for minimising a maximum mean discrepancy. *Statistics and Computing (to appear)*, 2022. Preprint hal-03114891, arXiv:2101.07564.

[13] L. Pronzato and M.-J. Rendas. Validation design I: construction of validation designs via kernel herding. 2021. Preprint hal-03474805, arXiv:2112.05583.

[14] L. Pronzato and A.A. Zhigljavsky. Bayesian quadrature, energy minimization and space-filling design. *SIAM/ASA J. Uncertainty Quantification*, 8(3):959–1011, 2020.

[15] C.E. Rasmussen and Z. Ghahramani. Bayesian monte carlo. *Advances in neural information processing systems*, pages 505–512, 2003.

[16] S. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

[17] M.L. Stein. *Interpolation of Spatial Data. Some Theory for Kriging*. Springer, Heidelberg, 1999.

[18] Z. Szabó and B. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18:1–29, 2018.

[19] M. Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1121–1128, 2009.

[20] H.P. Wynn. The sequential generation of D-optimum experimental designs. 41:1655–1664, 1970.

Luc Pronzato. Bât. Euclide, Les Algorithmes, 2000 route des lucioles, 06900 Sophia Antipolis cedex, CNRS, Université Côte d'Azur, Laboratoire I3S, France.
E-mail address: pronzato@i3s.unice.fr

Maria-João Rendas. Bât. Euclide, Les Algorithmes, 2000 route des lucioles, 06900 Sophia Antipolis cedex, CNRS, Université Côte d'Azur, Laboratoire I3S, France.
E-mail address: rendas@i3s.unice.fr