ANZIAM J. **66** (2024), 35–61
doi:10.1017/S1446181124000087

METHODOLOGY AND CHALLENGES OF SURROGATE MODELLING METHODS FOR MULTI-FIDELITY EXPENSIVE BLACK-BOX PROBLEMS

NICOLAU ANDRÉS-THIÓ ✉[1,2], MARIO ANDRÉS MUÑOZ[2,3] and
KATE SMITH-MILES[1,2,†]

(Received 11 November, 2023; accepted 3 April, 2024; first published online 31 May, 2024)

Abstract

Many industrial design problems are characterized by a lack of an analytical expression defining the relationship between design variables and chosen quality metrics. Evaluating the quality of new designs is therefore restricted to running a predetermined process such as physical testing of prototypes. When these processes carry a high cost, choosing how to gather further data can be very challenging, whether the end goal is to accurately predict the quality of future designs or to find an optimal design. In the multi-fidelity setting, one or more approximations of a design's performance are available at varying costs and accuracies. Surrogate modelling methods have long been applied to problems of this type, combining data from multiple sources into a model which guides further sampling. Many challenges still exist; however, the foremost among them is choosing when and how to rely on available low-fidelity sources. This tutorial-style paper presents an introduction to the field of surrogate modelling for multi-fidelity expensive black-box problems, including classical approaches and open questions in the field. An illustrative example using Australian elevation data is provided to show the potential downfalls in blindly trusting or ignoring low-fidelity sources, a question that has recently gained much interest in the community.

[†] This is a contribution to the series of invited papers by past Tuck Medallists (Editorial, Issue 62(1)). Kate Smith-Miles was awarded the Tuck Medal in 2017.

[1] School of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria 3010, Australia; e-mail: nandres@student.unimelb.edu.au, smith-miles@unimelb.edu.au

[2] ARC Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Parkville, Victoria 3010, Australia

[3] School of Computer and Information Systems, The University of Melbourne, Parkville, Victoria 3010, Australia; e-mail: munoz.m@unimelb.edu.au

2020 *Mathematics subject classification*: 90-10.

Keywords and phrases: expensive black-box, surrogate modelling, bi-fidelity, Kriging, co-Kriging.

1. Introduction

Industrial design problems are often centred around the performance analysis of a new design under a variety of use case scenarios, or the search for design parameters that optimize the performance of a new product. These can be seen, for example, when characterizing the lift and drag of an aircraft under a wide range of flight conditions [38], or when optimizing the design of a ship's hull form to minimize its water drag [28]. In both cases, a set of inputs (that is, flight conditions or a hull's length, draught and so on) affect the investigated output (that is, aircraft lift and drag or a hull's water drag), but often no analytical expression of this relationship is known. Problems of this type are known as *black-box* problems, where the only way to gather information is to sample this black-box by following some predetermined procedure. Industrial black-box problems may also be *expensive*; that is, sampling the black-box carries a high-cost requirement such as building a prototype or running a lengthy computer simulation. This implies that the amount of available data is severely limited and new sampling sites must be chosen with care.

Surrogate modelling methods [22] have long been applied to these so-called *expensive black-box* (EBB) problems to surmount the limitations of working with very little available data. These techniques rely on constructing an accurate model of the black-box to help guide further sampling, whether to explore the input space to further learn the input-output relationship, or to optimize a design by exploring unknown regions and exploiting promising ones. Furthermore, a key assumption with EBB problems is that the cost to sample the black-box dominates almost any computational cost required to train a model. This presents a specific scenario where slow-to-train but highly accurate models can be constructed and relied upon.

It is often the case when dealing with a high-cost, highly accurate black-box that other cheaper, less accurate sources of information are available. Problems of this type are known as multi-fidelity expensive black-box (Mf-EBB) problems. The special cases where only one additional lower fidelity source is available are known as bi-fidelity expensive black-box (Bf-EBB) problems. These sources offer a trade-off between cost and accuracy, often providing a large amount of data which can enhance the understanding of the expensive source. When these sources are available, surrogate models often integrate them into the construction of a single model of the costly black-box. The fusion of multiple information sources into a single surrogate model has readily been applied to a very wide range of industrial design problems, from ship component [28, 41, 43] and civil infrastructure [13, 40] design to traffic state estimation [1] and inter-satellite calibration [10]. Aeroplane component design in particular often relies on multi-fidelity surrogate models [6, 26, 32, 37, 50, 52], where

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

expensive wind tunnel data might be supplemented with much more abundant but less reliable computational fluid dynamics (CFD) data.

When using surrogate modelling techniques in Mf-EBB problems, many high-level decisions need to be made. These include selecting the underlying surrogate model of the expensive black-box source, choosing criteria for selecting further sampling sites, dividing the available budget between an initial spread-out sample and further model-guided sampling, and choosing which source to sample at each iteration, among others. Furthermore, open questions still exist in the field, foremost among them how to distinguish between beneficial and harmful low-fidelity sources. Indeed, relying on this additional information might hinder the accuracy of a constructed model or lead to worse points being found when optimizing.

This tutorial-style paper presents an introduction for new practitioners to the usage of surrogate modelling techniques in Mf-EBB problems. This is achieved by going through the classical single-source technique known as *Kriging* [22, 25, 30] and its two-source variation known as *co-Kriging* [17]. Despite having been developed more than 20 years ago, these techniques are still relevant today due to their strong theory when training the model's hyperparameters and the model-based uncertainty metric they provide, which can guide exploration as well as a balance between exploration and exploitation of the design space. Furthermore, many newer approaches are variations of, or inspired by, these techniques. Methods used to train these models and to choose future samples based on the constructed model as well as other aspects to consider are presented here. An illustrative example is also used to illustrate these techniques, as well as the potential pitfalls of relying on low-fidelity sources without proper prior assessment.

The remainder of this paper is structured as follows. Section 2 formally defines the different variations of Mf-EBB problems one might consider, as different scenarios arise when the end goal is the construction of an accurate model or the optimization of a design. Section 3 presents the surrogate modelling techniques, Kriging and co-Kriging, their mathematical derivation, and the process followed to train them and choose sampling sites. Section 4 introduces an illustrative example using Australian elevation data and compares the performance of Kriging and co-Kriging models which rely on low-fidelity sources of varying quality. Finally, Section 5 concludes with existing open questions in the field.

2. Problem definitions

The common component to Mf-EBB problems is the existence of a high-fidelity source f_h and one or more low-fidelity sources $f_{l_1}, f_{l_2}, \ldots, f_{l_k}$. The term $f_h(\mathbf{x})$ denotes the true performance metric value of a design $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ obtained via the costly high-fidelity procedure. This is the source which will be modelled and potentially optimized as discussed in the following subsections. The term $f_{l_i}(\mathbf{x})$ denotes the approximation of the true value obtained from source l_i. It is assumed that the input space being investigated is a hypercube, namely, the sources are formally defined as

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

$$f_h, f_{l_1}, f_{l_2}, \dots, f_{l_k} : \Omega \to \mathbb{R},$$
$$\Omega = [x_1^{\perp}, x_1^{\top}] \times \cdots \times [x_d^{\perp}, x_d^{\top}] \subset \mathbb{R}^d,$$

where the hypercube Ω has bounds $[x_i^{\perp}, x_i^{\top}]$ for dimension i. It is often the case that each of the low-fidelity sources represents a progressive decrease both in accuracy and cost relative to f_h. This however is not a requirement and a source could be cheaper but more accurate than another, or more accurate in certain regions of the space. It is up to the surrogate modelling methods being employed to determine how to rely on these sources based on cost and apparent usefulness.

One of the underlying assumptions in this paper is that all information sources are deterministic and therefore sampling a source at a sampled location will only return an already known value. It is possible however for one or more of the sources to be stochastic, that is, sampling the same location repeatedly will lead to different values. Modelling techniques developed for stochastic sources can be quite different to the approaches presented in this paper. For the interested reader in surrogate models of stochastic EBB sources, the reader is directed to the work of Ankenman et al. [4]. Two further aspects of these sources might arise in industry but seldom appear in the literature. The first is the fact that the sources might not have the same domain, such as the high-fidelity source not being available for sampling in certain regions of the space. Another complex aspect is the changing trust that testing engineers will have in different sources, sometimes assuming that low-fidelity sources are more accurate than the labelled high-fidelity data. In a sense, however, this aspect can be analysed in a pre-processing step using engineering domain knowledge to label which source provides the "truth". Therefore, as is almost exclusively done in the literature, here we assume that f_h always provides the true performance value.

It is worth highlighting at this point that studies of Mf-EBB problems are conducted in one of two broad settings, namely, a synthetic and an applied setting. In a synthetic setting, all sources are only assumed to be expensive. In reality, they are almost always analytical (that is, they are based on a mathematical expression) and the low-fidelity sources are often a modification of the high-fidelity source via the addition of some terms or the modification of the coefficients of the expression [3, 45, 49]. A large amount of these synthetic instances have been implemented in C++ and made available on GitHub [2]. Studies that focus on this setting have large amounts of data from each of the sources available, despite using only limited samples when training models or optimizing. This large amount of data can be used to precisely assess the accuracy of constructed models, as well as to characterize the sources themselves to analyse their impact on algorithm performance [3, 45]. This type of study can provide insights into the inner workings of surrogate model methods and provide guidelines for their usage.

It is also possible for researchers to focus on the applied setting, that is, studies that focus on the limited data available and how to exploit it. This is particularly relevant to industrial problems, where the information sources are true expensive black-boxes

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

and very little is known about them. Studies of this type primarily propose new ways to identify useful low-fidelity sources [48] or ways to exploit low-fidelity information when it seems beneficial to do so [33]. Guidelines and insights developed in a synthetic setting can often impact the development of these applied techniques. Both settings therefore have their place and are beneficial to the literature. Mf-EBB problems can be further classified into three types of sub-problems. This division is characterized by the given resources, the aim and the success measure being used. These sub-problems are defined next and are later illustrated in Section 4.

2.1. Problem 1: model creation with fixed sample In this scenario, data from the available sources have already been gathered and no further sampling of any of the sources is allowed. That is, the sets $\mathbf{X}_h, \mathbf{X}_{l_1}, \ldots, \mathbf{X}_{l_k}$ and $\mathbf{y}_h, \mathbf{y}_{l_1}, \ldots, \mathbf{y}_{l_k}$ are defined as

$$\begin{aligned}
\mathbf{X}_h &= \{\mathbf{x}_1^h, \mathbf{x}_2^h, \ldots, \mathbf{x}_{n_h}^h\} \subset \Omega, \\
\mathbf{X}_{l_i} &= \{\mathbf{x}_1^{l_i}, \mathbf{x}_2^{l_i}, \ldots, \mathbf{x}_{n_{l_i}}^{l_i}\} \subset \Omega \quad \text{for } 1 \le i \le k, \\
\mathbf{y}_h &= \{f_h(\mathbf{x}_1^h), \ldots, f_h(\mathbf{x}_{n_h}^h)\}, \\
\mathbf{y}_{l_i} &= \{f_{l_i}(\mathbf{x}_1^{l_i}), \ldots, f_{l_i}(\mathbf{x}_{n_{l_i}}^{l_i})\} \quad \text{for } 1 \le i \le k.
\end{aligned}$$

As the data gathered come from an expensive source, the amount of high-fidelity samples is relatively small, that is, $n_h \le 20d$, where d is the problem dimension. Furthermore, if the sources $f_h, f_{l_1}, f_{l_2}, \ldots, f_{l_k}$ are assumed to be of decreasing quality, it may be that $\mathbf{X}_h \subseteq \mathbf{X}_{l_1} \subseteq \mathbf{X}_{l_2} \subseteq \cdots \subseteq \mathbf{X}_{l_k}$, although this is not always the case.

The aim in this problem is to construct a model $s_h : \Omega \to \mathbb{R}$ of f_h, which is as accurate as possible given the available data. The performance of the constructed model in a synthetic setting is measured using a large sample set $\mathbf{X}$ and $\mathbf{y}$ defined as

$$\begin{aligned}
\mathbf{X} &= \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \Omega, \\
\mathbf{y} &= \{f_h(\mathbf{x}_1), \ldots, f_h(\mathbf{x}_N)\},
\end{aligned}$$

where N should be a large number, say, $N = 1000d$. Given this sample, two accuracy measures are often used in the literature. The first is the root mean squared error (RMSE) between s_h and f_h, given by

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (s_h(\mathbf{x}_i) - f_h(\mathbf{x}_i))^2}{n}}$$

with a lower error indicating better performance. Another possible measure is the Pearson's correlation P_{corr} between s_h and f_h, given by

$$P_{\text{corr}} = \frac{1}{N-1}\left(\frac{\sum_{i=1}^{N} (f_h(\mathbf{x}_i) - \bar{y})(s_h(\mathbf{x}_i) - \bar{S})}{s_y s_S}\right),$$

$$\text{where } \bar{y} = \frac{1}{N}\sum_{i=1}^{N} f_h(\mathbf{x}_i),$$

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

$$s_y = \left[\frac{\sum_{i=1}^N (f_h(\mathbf{x}_i) - \bar{y})^2}{N-1}\right]^{1/2},$$

$$\bar{S} = \frac{1}{N}\sum_{i=1}^N s_h(\mathbf{x}_i),$$

$$s_S = \left[\frac{\sum_{i=1}^N (s_h(\mathbf{x}_i) - \bar{S})^2}{N-1}\right]^{1/2},$$

with a higher correlation indicating better performance. Whilst it is rare to find industrial examples that fit this problem definition, it is quite often used in studies that propose new surrogate modelling methods. Here the practitioner's decision-making is restricted to a single question: Given a set of data, which model should be trained? The performance of different choices, such as how to divide a total budget between different phases of an algorithm or the usage of different data acquisition strategies, does not need to be assessed. This allows for an easy comparison between models.

2.2. Problem 2: model creation with sample budget In this scenario, rather than being given already collected data from each of the sources, the user is given a total sampling budget of B to be used as desired. This budget indicates how many times the source f_h can be queried. Furthermore, a given set of cost ratios $0 \leq C_r^{l_1}, C_r^{l_2}, \ldots, C_r^{l_k} \leq 1$ indicates the cost of sampling the sources $f_{l_1}, f_{l_2}, \ldots, f_{l_k}$, respectively, relative to the cost of f_h. For instance, a cost ratio $C_r^{l_i} = 0.1$ indicates that sampling f_{l_i} is 10 times cheaper than sampling f_h and so one f_h sample can be replaced by $10\, f_{l_1}$ samples. The special case where $C_r^{l_i} = 0$ indicates that the source f_{l_i} is cheap, that is, it can be queried as often as desired with no associated cost.

The aim here is to gather the data $\mathbf{X}_h, \mathbf{X}_{l_1}, \ldots, \mathbf{X}_{l_k}$ and $\mathbf{y}_h, \mathbf{y}_{l_1}, \ldots, \mathbf{y}_{l_k}$ from each of the sources as defined in the previous subsection whilst satisfying

$$B \geq |\mathbf{X}_h| + C_r^{l_1}|\mathbf{X}_{l_1}| + \cdots + C_r^{l_k}|\mathbf{X}_{l_k}|.$$

These data are used to train a model s_h of f_h which is as accurate as possible. Once again, the performance is measured in terms of the accuracy of the constructed model, with either the RMSE or P_{corr} measures.

This type of problem is a step up from the previous one. The practitioner needs to choose not only what kind of surrogate model to use, but also how to divide the budget among fidelities and how to guide future data gathering. It is also much more realistic in terms of industrial problems, where the aim is not to optimize a design but rather to understand its performance under a range of conditions.

2.3. Problem 3: function optimization with sample budget In this final scenario, the user is once again given the sources $f_h, f_{l_1}, f_{l_2}, \ldots, f_{l_k}$, a total budget B and a set of cost ratios $0 \leq C_r^{l_1}, C_r^{l_2}, \ldots, C_r^{l_k} \leq 1$. The aim is to either minimize or maximize f_h, whilst sampling each of the sources at locations $\mathbf{X}_h, \mathbf{X}_{l_1}, \ldots, \mathbf{X}_{l_k}$ and satisfying

$$B \geq |\mathbf{X}_h| + C_r^{l_1}|\mathbf{X}_{l_1}| + \cdots + C_r^{l_k}|\mathbf{X}_{l_k}|.$$

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

The performance measure is simply the best objective function value of the sampled points $\mathbf{X}_h$. Whilst being very similar to the previous problem, the key difference here lies in the fact that one must balance the construction of an accurate surrogate to aid in optimization and the optimization process itself. Sampling low- and high-fidelity sources in unexplored regions can help generate more accurate models. Data gathering therefore needs to balance a need for exploration of the space to find new promising regions with the exploitation of already discovered good regions to find the best points possible.

3. Surrogate modelling techniques

As outlined in Section 1, one of the more prominent ways to approach Mf-EBB problems is through the construction of a surrogate model. This approach applies to all three of the Mf-EBB sub-problems defined in the previous section, whether the construction of a model is the end goal itself or when optimizing an expensive source. Practitioners need to choose a model that not only accurately models the expensive source with limited data, but also has a suitable mechanism to fuse information from low-fidelity sources. Furthermore, a suitable acquisition function must be chosen, a measure of how promising further sample locations are when exploring the space, exploiting known promising regions, and often balancing the two. Finally, when given a total budget with no initial data, an initial sample needs to be gathered to construct an initial surrogate model. Techniques that solve each of these aspects are presented here.

3.1. Surrogate model

Despite a very large number of data modelling techniques reported in the literature, dealing with expensive sources and small datasets often leads to the usage of models which might take some time to train, but which can be highly accurate even with limited information. Perhaps the two best-known underlying models used in the literature are Kriging [22, 25, 30] and radial basis functions (RBFs) [12, 18, 34, 39, 51]. These models provide predictions that are based both on an underlying simple model, often chosen to be constant or linear in the input variables, combined with an added term that represents the effect of known data points. This leads to a modelling surface that interpolates existing data, that is, the model prediction at already sampled locations is the known objective function value.

3.1.1. Kriging Kriging in particular provides very accurate models due to the combination of its high number of hyperparameters and an approach to choose them rooted in sound probability theory. Note that this method is used to model a single source; its adaptation to multiple sources is given in the next section. The formulation given by Kriging assumes that the function samples made so far at locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are realizations of random normal variables $Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n)$ with mean μ and variance σ^2. Further, the correlation between variables is based on the distance between them, that is,

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

$$\mathrm{corr}(Y(\mathbf{x}_i), Y(\mathbf{x}_j)) = \exp\left\{-\sum_{k=1}^{d} 10^{\theta_k} |\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)}|^{p_k}\right\}.$$

Thus, the multivariate random variable $\mathbf{Y} = [Y(\mathbf{x}_1)\cdots Y(\mathbf{x}_n)]$ has the distribution $\mathbf{Y} \sim N(\mathbf{1}\mu, \sigma^2 R)$, with

$$R_{i,j} = \exp\left\{-\sum_{k=1}^{d} 10^{\theta_k} |\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)}|^{p_k}\right\}.$$

Simple Kriging assumes constant mean μ across the space, but it is possible to fit a regression model to the data before adding the Kriging "layer". This process is known as Universal Kriging [31] and is not covered here.

The hyperparameters of a simple Kriging model are $\mu, \sigma^2, \theta_1, \ldots, \theta_d$ and $p_1, \ldots, p_d$. The values θ_k and p_k give an indication of the effect of moving along any of the dimensions (that is, changing the value of a single variable) on the objective function. The hyperparameter θ_k represents how the correlation changes with distance: large values mean there is no correlation even for close points in the kth dimension, but small values indicate that even relatively distant sample points (in the kth dimension) are correlated. The constant p_k allows the technique to model from smooth functions ($p_k = 2$) to rough, non-differentiable ones ($p_k \to 0$). The probability of having observed the values $\mathbf{y}$ given the process $\mathbf{Y}$ is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}|R|^{1/2}} \exp\left\{\frac{-(\mathbf{y} - \mathbf{1}\mu)^T R^{-1}(\mathbf{y} - \mathbf{1}\mu)}{2\sigma^2}\right\}.$$

The hyperparameter values are chosen to be the maximum likelihood estimators (MLEs). In practice, first the (monotonically increasing) log function is applied, leading to the maximization of

$$\log f_{\mathbf{Y}}(\mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log|R| - \frac{(\mathbf{y} - \mathbf{1}\mu)^T R^{-1}(\mathbf{y} - \mathbf{1}\mu)}{2\sigma^2}.$$

An analytical solution exists for the MLEs $\hat{\mu}$ and $\hat{\sigma}^2$ of μ and σ^2, namely,

$$\hat{\mu} = \frac{\mathbf{1}^T R^{-1}\mathbf{y}}{\mathbf{1}^T R^{-1}\mathbf{1}},$$

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{1}\hat{\mu})^T R^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu})}{n}.$$

Substituting this solution into the log-likelihood and removing constant terms leads to the concentrated log-likelihood

$$-\frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2}\log(|R|).$$

Finding the remainder of the hyperparameters must be done numerically as no analytical solution exists. This consists of an auxiliary optimization problem where

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

the concentrated log-likelihood is maximized with respect to $\theta_1, \ldots, \theta_d$ and $p_1, \ldots, p_d$. Finding good black-box algorithms to solve this problem is an active area of research in and of itself [46, 47], but in theory, any solver can be used, such as Accelerated Random Search [5]. It is recommended to linearly scale the data to lie within the hypercube $[0, 1]^d$ before fitting the model, and to use the bounds $-3 \le \theta_i \le 3$ and $0.1 \le p_i \le 2.0$ when finding the MLEs.

Once the model has been trained, Kriging treats a prediction at a location $\mathbf{x}$ in the space as having the distribution $Y(\mathbf{x}) \sim N(s(\mathbf{x}), v^2(\mathbf{x}))$. That is, Kriging provides not only the most likely objective function value $s(\mathbf{x})$, but also a prediction error metric $v^2(\mathbf{x})$, defined as

$$s(\mathbf{x}) = \hat{\mu} + \mathbf{r}^T R^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}),$$

$$v^2(\mathbf{x}) = \hat{\sigma}^2\left[1 - \mathbf{r}^T R^{-1}\mathbf{r} + \frac{(1 - \mathbf{1}^T R^{-1}\mathbf{r})^2}{\mathbf{1}^T R^{-1}\mathbf{1}}\right],$$

$$\text{where } \mathbf{r} = \begin{bmatrix} \operatorname{corr}(Y(\mathbf{x}), Y(\mathbf{x}_1)), \\ \cdots \\ \operatorname{corr}(Y(\mathbf{x}), Y(\mathbf{x}_n)) \end{bmatrix}.$$

The prediction error metric $v^2(\mathbf{x})$, in particular, is a helpful metric used to guide further sampling. This is discussed below. For further details on the derivation of these expressions, the reader is directed to the work of Jones [22].

3.1.2. Co-Kriging The work of Kennedy and O'Hagan [24] presents a classical approach for the integration of multiple fidelity sources into a single model. Given the sources $f_{l_0}, f_{l_1}, \ldots, f_{l_k}$ of decreasing accuracy with $f_{l_0} = f_h$, the observed objective function value $f_{l_k}(\mathbf{x})$ at the location $\mathbf{x}$ is assumed to come from the random process $Y_{l_k}(\mathbf{x})$. For fidelities l_t with $0 \le t \le k-1$, the observed value $f_{l_t}(\mathbf{x})$ at the location $\mathbf{x}$ is assumed to come from the random process

$$Y_{l_t}(\mathbf{x}) = \rho_{l_{t+1}} Y_{l_{t+1}}(\mathbf{x}) + Y_{l_t}^{\text{diff}}(\mathbf{x}) \quad \text{for } 0 \le t \le k-1.$$

Here $\rho_{l_{t+1}}$ is kind of a regression parameter, $Y_{l_{t+1}}(\mathbf{x})$ is the random process representing the values obtained from source l_{t+1} and the difference term $Y_{l_t}^{\text{diff}}(\mathbf{x})$ is assumed to be independent of $Y_{l_{t+1}}(\mathbf{x}), \ldots, Y_{l_k}(\mathbf{x})$. The combination of this framework with Kriging in the special case where only two sources are available is known as co-Kriging [17]. Similarly to Kriging, the idea here is to model the responses of a single cheap objective function f_l at sample points $\mathbf{X}_l = (\mathbf{x}_1^l, \mathbf{x}_2^l, \ldots, \mathbf{x}_{n_l}^l)$ and the responses of the expensive objective function f_h at sample points $\mathbf{X}_h = (\mathbf{x}_1^h, \mathbf{x}_2^h, \ldots, \mathbf{x}_{n_h}^h)$ as the realization of a multivariate random variable:

$$\mathbf{Y} = (\mathbf{Y}_l(\mathbf{X}_l), \mathbf{Y}_h(\mathbf{X}_h)) = (Y_l(\mathbf{x}_1^l), \ldots, Y_l(\mathbf{x}_{n_l}^l), Y_h(\mathbf{x}_1^h), \ldots, Y_h(\mathbf{x}_{n_h}^h)).$$

Following the framework of Kennedy and O'Hagan [24], the multivariate random variable $\mathbf{Y}_l(\mathbf{X}_l)$, that is, the response of the cheap objective function, is treated as a multivariate normal random variable with distribution $N(\mu_l, \sigma_l^2 R_l)$. The multivariate

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

random variable $\mathbf{Y}_h(\mathbf{X}_h)$, that is, the response of the expensive objective function, is represented by a scaling of ρ of the response of the cheap expensive function $\mathbf{Y}_l(\mathbf{X}_h)$ plus a new Gaussian process $\mathbf{Y}_{\text{diff}}(\mathbf{X}_h)$, which models the difference between the cheap and expensive objective functions, that is,

$$\mathbf{Y}_h(\mathbf{X}_h) = \rho\mathbf{Y}_l(\mathbf{X}_h) + \mathbf{Y}_{\text{diff}}(\mathbf{X}_h).$$

The random variable $\mathbf{Y}_{\text{diff}}$ is also treated as a multivariate normal random variable with distribution $N(\mu_{\text{diff}}, \sigma^2_{\text{diff}}R_{\text{diff}})$. It is assumed that $\mathbf{Y}_l$ and $\mathbf{Y}_{\text{diff}}$ are independent random variables. Thus, the following covariance measures are given for $\mathbf{Y}_l$ and $\mathbf{Y}_h$:

$$\begin{aligned}
\operatorname{cov}(\mathbf{Y}_l(\mathbf{X}_l), \mathbf{Y}_l(\mathbf{X}_l)) &= \sigma_l^2 R_l(\mathbf{X}_l, \mathbf{X}_l),\\
\operatorname{cov}(\mathbf{Y}_h(\mathbf{X}_h), \mathbf{Y}_l(\mathbf{X}_l)) &= \rho\sigma_l^2 R_l(\mathbf{X}_h, \mathbf{X}_l),\\
\operatorname{cov}(\mathbf{Y}_h(\mathbf{X}_h), \mathbf{Y}_h(\mathbf{X}_h)) &= \rho^2\sigma_l^2 R_l(\mathbf{X}_h, \mathbf{X}_h) + \sigma^2_{\text{diff}}R_{\text{diff}}(\mathbf{X}_h, \mathbf{X}_h),\\
\text{where } R_l(\mathbf{x}_1, \mathbf{x}_2) &= \exp\Big\{-\sum_{k=1}^{d} 10^{\theta_k^l}|\mathbf{x}_1^{(k)} - \mathbf{x}_2^{(k)}|^{p_k^l}\Big\},\\
R_{\text{diff}}(\mathbf{x}_1, \mathbf{x}_2) &= \exp\Big\{-\sum_{k=1}^{d} 10^{\theta_k^{\text{diff}}}|\mathbf{x}_1^{(k)} - \mathbf{x}_2^{(k)}|^{p_k^{\text{diff}}}\Big\},\\
R_l(\mathbf{X}_l, \mathbf{X}_l)_{i,j} &= R_l(\mathbf{x}_i^l, \mathbf{x}_j^l) \qquad 1 \le i,j \le n_l,\\
R_l(\mathbf{X}_h, \mathbf{X}_l)_{i,j} &= R_l(\mathbf{x}_i^h, \mathbf{x}_j^l) \qquad 1 \le i \le n_h \quad 1 \le j \le n_l,\\
R_l(\mathbf{X}_h, \mathbf{X}_h)_{i,j} &= R_l(\mathbf{x}_i^h, \mathbf{x}_j^h) \qquad 1 \le i,j \le n_h,\\
R_{\text{diff}}(\mathbf{X}_h, \mathbf{X}_h)_{i,j} &= R_{\text{diff}}(\mathbf{x}_i^h, \mathbf{x}_j^h) \qquad 1 \le i,j \le n_h.
\end{aligned}$$

Training the hyperparameters of a co-Kriging model is a two-step process. First, the MLEs $\hat{\mu}_l, \hat{\sigma}_l^2, \hat{\theta}_1^l, \ldots, \hat{\theta}_d^l, \hat{p}_1^l, \ldots, \hat{p}_d^l$ are found similarly to the training of a single-source Kriging model. That is, a numerical solution is found to the auxiliary optimization problem

$$\begin{aligned}
\max_{\theta_1^l,\ldots,\theta_d^l,p_1^l,\ldots,p_d^l} & -\frac{n_l}{2}\log(\hat{\sigma}_l^2) - \frac{1}{2}\log(|R_l(\mathbf{X}_l, \mathbf{X}_l)|),\\
\text{where } \hat{\mu}_l &= \frac{\mathbf{1}^T R_l(\mathbf{X}_l, \mathbf{X}_l)^{-1}\mathbf{y}_l}{\mathbf{1}^T R_l(\mathbf{X}_l, \mathbf{X}_l)^{-1}\mathbf{1}},\\
\hat{\sigma}_l^2 &= \frac{(\mathbf{y}_l - \mathbf{1}\hat{\mu}_l)^T R_l(\mathbf{X}_l, \mathbf{X}_l)^{-1}(\mathbf{y}_l - \mathbf{1}\hat{\mu}_l)}{n_l},\\
\mathbf{y}_l &= (f_l(\mathbf{x}_1^l), \ldots, f_l(\mathbf{x}_{n_l}^l))^T.
\end{aligned}$$

The second step consists of finding the MLEs $\hat{\rho}, \hat{\mu}_{\text{diff}}, \hat{\sigma}^2_{\text{diff}}, \hat{\theta}_1^{\text{diff}}, \ldots, \hat{\theta}_d^{\text{diff}}, \hat{p}_1^{\text{diff}}, \ldots, \hat{p}_d^{\text{diff}}$. First, the vector $\mathbf{d}$ is defined as

$$\mathbf{d} = \mathbf{y}_h - \rho\mathbf{y}_l(\mathbf{X}_h),$$

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

where $\mathbf{y}_h = (f_h(\mathbf{x}_1^h), \ldots, f_h(\mathbf{x}_{n_h}^h))^T$, and $\mathbf{y}_l(\mathbf{X}_h)_i$ is $f_l(\mathbf{x}_i^h)$ if this value is known and otherwise, it is

$$s_l(\mathbf{x}_i^h) = \hat{\mu}_l + \mathbf{r}_l^T R_l(\mathbf{X}_l, \mathbf{X}_l)^{-1}(\mathbf{y}_l - \mathbf{1}\hat{\mu}_l) \quad \text{with } \mathbf{r}_l = (R_l(\mathbf{x}, \mathbf{x}_1^l), \ldots, R_l(\mathbf{x}, \mathbf{x}_{n_l}^l))^T.$$

That is, if a point has not been evaluated by f_l yet, the trained Kriging predictor of the low-fidelity source is used instead. The MLEs of the second set of hyperparameters are chosen to be the numerical solution of the auxiliary optimization function

$$\begin{aligned}
\max_{\rho, \theta_1^{\text{diff}}, \ldots, \theta_d^{\text{diff}}, p_1^{\text{diff}}, \ldots, p_d^{\text{diff}}} \; & -\frac{n_h}{2} \log(\hat{\sigma}_{\text{diff}}^2) - \frac{1}{2} \log(|(R_{\text{diff}}(\mathbf{X}_h, \mathbf{X}_h)|), \\
\text{where } \hat{\mu}_{\text{diff}} &= \frac{\mathbf{1}^T R_{\text{diff}}(\mathbf{X}_h, \mathbf{X}_h)^{-1} \mathbf{d}}{\mathbf{1}^T R_{\text{diff}}(\mathbf{X}_h, \mathbf{X}_h)^{-1} \mathbf{1}}, \\
\hat{\sigma}_{\text{diff}}^2 &= \frac{(\mathbf{d} - \mathbf{1}\hat{\mu}_{\text{diff}})^T R_{\text{diff}}(\mathbf{X}_h, \mathbf{X}_h)^{-1} (\mathbf{d} - \mathbf{1}\hat{\mu}_{\text{diff}})}{n_h}.
\end{aligned}$$

Similarly to Kriging, a trained co-Kriging model treats a prediction at a location $\mathbf{x}$ as having the distribution $Y(\mathbf{x}) \sim N(s_h(\mathbf{x}), v_h^2(\mathbf{x}))$, providing both a prediction $s_h(\mathbf{x})$ and an error metric $v_h^2(\mathbf{x})$. These are defined as

$$\begin{aligned}
s_h(\mathbf{x}) &= \hat{\mu} + \mathbf{c}^T C^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}), \\
v_h^2(\mathbf{x}) &= \hat{\rho}^2 \hat{\sigma}_l^2 + \hat{\sigma}_{\text{diff}}^2 - \mathbf{c}^T C^{-1} \mathbf{c} + \frac{(1 - \mathbf{1}^T C^{-1} \mathbf{c})^2}{\mathbf{c}^T C^{-1} \mathbf{c}}, \\
\text{with } C &= \begin{bmatrix} \hat{\sigma}_l^2 R_l(\mathbf{X}_l, \mathbf{X}_l) & \hat{\rho}\hat{\sigma}_l^2 R_l(\mathbf{X}_l, \mathbf{X}_h) \\ \hat{\rho}\hat{\sigma}_l^2 R_l(\mathbf{X}_h, \mathbf{X}_l) & \hat{\rho}^2 \hat{\sigma}_l^2 R_l(\mathbf{X}_h, \mathbf{X}_h) + \hat{\sigma}_{\text{diff}}^2 R_{\text{diff}}(\mathbf{X}_h, \mathbf{X}_h), \end{bmatrix}, \\
\mathbf{c} &= \begin{bmatrix} \hat{\rho}\hat{\sigma}_l^2 R_l(\mathbf{X}_l, \mathbf{x}) \\ \hat{\rho}^2 \hat{\sigma}_l^2 R_l(\mathbf{X}_h, \mathbf{x}) + \hat{\sigma}_{\text{diff}}^2 R_{\text{diff}}(\mathbf{X}_h, \mathbf{x}) \end{bmatrix}, \\
\hat{\mu} &= \frac{\mathbf{1}^T C^{-1} \mathbf{y}}{\mathbf{1}^T C^{-1} \mathbf{1}}.
\end{aligned}$$

For further details on the derivation of these expressions, the reader is directed to the work of Forrester et al. [17]. It is worth pointing out that the correlation matrix C can grow very large as more low-fidelity sources are made available. A potential alternative is the use of *hierarchical Kriging* [19], a theoretical simplification of co-Kriging aimed at easier software implementation when a large number of sources are available. It is also possible to take a different approach entirely and to add a categorical variable that denotes from which source a point was sampled. This allows us to consider the data gathered from all sources as coming from a single source in $(d + 1)$-dimensional space and to train a single Kriging model on this $(d + 1)$-dimensional data [14].

3.2. Initial design of experiments In the case where no initial data are given, building an initial model first requires the sampling of some or all of the available

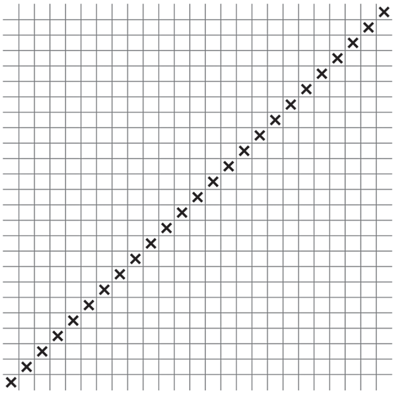

FIGURE 1. A two-dimensional sampling plan of 25 samples with clearly undesirable properties. The thin grey lines indicate the division of each of the two dimensions into 25 sections. Each cross represents a sample. Note that as each row and column contains only a single sample, this satisfies the definition of an LHS plan.

sources. Interestingly, no consensus exists in the literature on what proportion of the total budget should be spent on an initial sample and how this sample should be further spread among different sources. A recent 2019 survey [16] on the use of multi-fidelity models did not cover this question despite presenting different techniques to generate an initial sample plan. Forrester et al. [17] suggested a rule of thumb of $10d$ samples, where d is the problem dimension, to be further refined based on the total budget and the relative cost of the low-fidelity source. A very recent study [27] however which analysed (single-source) Kriging optimization on a large variety of functions found that a much smaller initial sample of size $d + 4$ worked best. When optimizing a bi-fidelity EBB problem using RBF surrogate models, Müller [33] found that taking an initial low-fidelity sample of size $2(d + 1)$ and a high-fidelity sample of size $d + 1$ outperformed sampling both the high- and low-fidelity sources $2(d + 1)$ times. Finally, despite not providing a suggested proportion of the total budget to be spent on an initial sample, Toal [45] suggested spending between 10% and 80% of the initial sampling budget on low-fidelity samples. Generating guidelines for the generation of an initial sample is one of the open questions in this field.

Having decided on the number of samples to collect from each of the sources, different approaches exist on how to place these samples within the space. A very common approach is to rely on latin hypercube sampling (LHS) [36]. Generating a sample of size n with this method consists of dividing each of the dimensions of the sample space into n intervals and then placing n samples in the space so that for every dimension, no two samples lie in the same interval. Whilst LHS in theory should provide a well-distributed sample within the space, simply relying on the generation of a random LHS plan can lead to a suboptimal design of experiments. Figure 1 in particular shows a sampling plan which, despite satisfying the LHS definition, is far from evenly sampling the design space. It is beneficial to generate an LHS

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

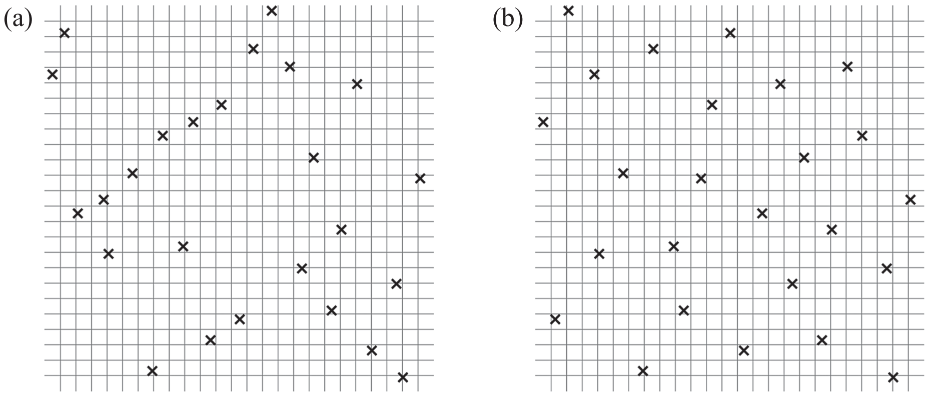

FIGURE 2. (a) A randomly generated LHS sampling plan of 25 samples inside the two-dimensional space $[0, 1]^2$. The minimum Euclidean distance between a pair of points is 0.0754. (b) Result of locally optimizing the sampling plan by swapping coordinates between pairs of points. This minimum distance between a pair of points is 0.1602 and the sample is more evenly spread out in the space.

plan that satisfies some kind of local optimality criteria, such as the maximization of the minimum Euclidean distance between any pair of points. This can be achieved by starting with a random LHS plan, and then iteratively taking two samples and swapping one of their coordinates as long as this increases the minimum distance between any pair of points. Once this can no longer be achieved, the obtained sample is locally optimal. The clear benefit of using such a procedure is shown in Figure 2. Note that this deterministic optimization can be quite lengthy at high dimensions and for large sample sizes, that is, $d \geq 10$ and $n \geq 100$, in which case, a heuristic approach such as simulated annealing [7] should be considered.

When generating a multi-fidelity sample, it is often a requirement that the sample plan of a particular fidelity is a subset of the sample plan of the previous fidelity. Note that this is not a requirement when constructing a co-Kriging model, although implementations of this technique often generate an initial design of experiments in this fashion. Given a sample plan of size n_l, the creation of a subset of size n_h can be achieved once again by starting with a random subset and then swapping a point in the subset with a point outside of it as long as this increases the minimum distance between a pair of points. Once this can no longer be done, the set is once again locally optimal. Figure 3 shows the benefit of this approach over simply choosing a random subset. Another approach when generating a subset is to start with a second LHS sampling plan and to map each of the samples to the larger sample set [35].

3.3. Acquisition function

Once an initial model has been trained, an acquisition function is a means by which to evaluate the benefit of taking a further sample in a particular location. New samples are chosen by finding locations that are the most promising. The definition of “promising” of course depends on the aim of the current

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

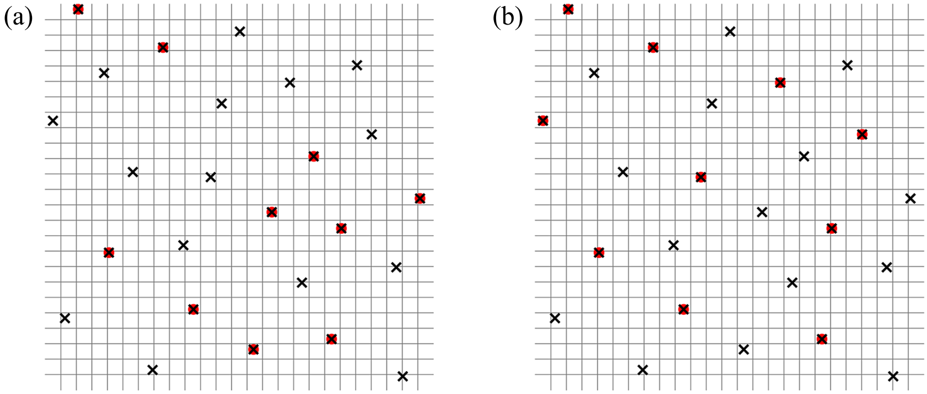

FIGURE 3. (a) Procedure when choosing a subset of an LHS plan. The crosses in both panels show an LHS plan of size 25 inside the two-dimensional space $[0, 1]^2$. The red circles represent a randomly chosen subset of size 10. The minimum distance between pairs of points is 0.1602. (b) Results of locally optimizing this subset in panel (a) by swapping points inside and outside of the set. The final subset is more evenly spread out, and the minimum distance between pairs of points is 0.2400.

process. When the aim is to create as accurate a model as possible, further sampling should rely on the model's error prediction. Kriging and co-Kriging particularly shine in this scenario as their error metric $v^2(\mathbf{x})$ is particularly well suited for searches of this type. An intuitive first option is to choose further sampling sites to be the location where the surrogate model is most unsure of its prediction, that is, to find

$$\arg\max_{\mathbf{x}\in\Omega} v^2(\mathbf{x}).$$

This approach has the benefit of being intuitive and easy to evaluate. It has been shown however that a large number of samples tend to be gathered on the boundary of the space. These samples might therefore have a low impact on the overall accuracy of the model across the whole space. A second metric which takes this into account can also be considered, namely,

$$\arg\min_{\mathbf{x}'\in\Omega} \int_{\mathbf{x}\in\Omega} v^2_{\mathbf{x}'}(\mathbf{x})\, d\mathbf{x}.$$

Here the term $v^2_{\mathbf{x}'}$ is the error prediction function of a surrogate model constructed with the available data with the addition of a sample at location $\mathbf{x}'$. The integral denotes the overall model error across the whole space if the point $\mathbf{x}$ were added to the dataset. This acquisition function takes a more global approach; however traditionally, this integral has been approximated using a technique such as Monte-Carlo integration [42]. Whilst the underlying assumption of Mf-EBB problems is that the computational time of the surrogate method is outweighed by sampling cost, it is worth stressing that this process can become computationally intensive. Promisingly, an analytical expression has been proposed for Kriging models with a reduced set of hyperparameters [8],

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

which might lead to this sampling strategy being applicable for relatively larger sample sizes in the multi-fidelity setting.

When employing a surrogate model in optimization, perhaps an intuitive first choice is to sample where the model's prediction $s(\mathbf{x})$ is maximized/minimized. That is, in the case of a minimization problem, to sample at

$$\arg\min_{\mathbf{x}\in\Omega} s(\mathbf{x}).$$

This however tends to not be a good strategy. Samples gathered in this fashion can tend to cluster in the same region of the space, particularly for interpolating models. This is especially counterproductive with a limited sampling budget, as many samples might be wasted finding a local minimum before restarting the search in some other region, as is often done by black-box optimizers. A strategy is needed which, at least part of the time, searches the space for yet-to-be-explored regions that might contain good values. Essentially, a balance needs to be struck between exploration to find new promising regions and further exploitation of regions with good solutions.

Once again, Kriging and co-Kriging can rely on their model error prediction metric to provide such a strategy. Assume that the optimization consists of minimizing a function f and that the best objective function value found so far had the value $f^{\min}$. Recall that both Kriging and co-Kriging provide a prediction of $f(\mathbf{x})$ at location $\mathbf{x}$ as a normally distributed random variable $Y(\mathbf{x}) \sim N(s(\mathbf{x}), v^2(\mathbf{x}))$. Then the probability that sampling at location $\mathbf{x}$ yields an improvement of $I \geq 0$ over the best point found so far is the probability that $Y(\mathbf{x}) = f^{\min} - I$. For $I > 0$, this has the probability density function

$$\frac{1}{\sqrt{2\pi v^2(\mathbf{x})}} \exp\left\{-\frac{1}{2v^2(\mathbf{x})}(f^{\min} - I - s(\mathbf{x}))^2\right\}.$$

An analytical expression for the expected improvement exists and it is given by

$$\begin{aligned}\mathbb{E}[I(\mathbf{x})] &= \int_{I=0}^{I=\infty} I\left[\frac{1}{\sqrt{2\pi v^2(\mathbf{x})}} \exp\left\{-\frac{1}{2v^2(\mathbf{x})}(f^{\min} - I - s(\mathbf{x}))^2\right\}\right] dI \\ &= \sqrt{v^2(\mathbf{x})}\left[\frac{f^{\min} - s(\mathbf{x})}{\sqrt{v^2(\mathbf{x})}} \Phi\left(\frac{f^{\min} - s(\mathbf{x})}{\sqrt{v^2(\mathbf{x})}}\right) + \phi\left(\frac{f^{\min} - s(\mathbf{x})}{\sqrt{v^2(\mathbf{x})}}\right)\right],\end{aligned}$$

where Φ and ϕ are the cumulative density function and probability density function of the standard normal distribution, respectively. Choosing further samples by maximizing this expression, that is,

$$\arg\max_{\mathbf{x}\in\Omega} \mathbb{E}[I(\mathbf{x})],$$

is known as *efficient global optimization* [23] and has many beneficial properties. The first is that it requires no tuning of optimization parameters once training of the surrogate model is complete. The second is that it automatically balances exploration and exploitation. Note that both very low $s(\mathbf{x})$ values (that is, locations

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

with good predicted values) and very high $v^2(\mathbf{x})$ (that is, locations with high prediction uncertainty) lead to high expected improvement values. The third reason is the fact that locations that have already been sampled have an expected improvement of 0 and, therefore, will never be selected again for sampling.

Choosing the location at which the next sample is collected consists of an auxiliary optimization problem of the acquisition function of choice. Having chosen the next sampling site, however, the question remains as to which of the sources should be sampled. Classical co-Kriging simply samples both f_h and f_l at the chosen location. This simple approach is beneficial for the accuracy of further co-Kriging models, as the training of the intermediate model of $f_h - \rho f_l$ can rely on exact values for both f_h and f_l wherever f_h is known. As shown in the formulation of co-Kriging in Section 3.1.2, however, this is not a requirement for training the model. Intuitively, one can see that there might be a benefit in sometimes sampling only f_l if this source is highly accurate or choosing to not sample f_l if it is not reliable enough.

The question of when to sample low-fidelity sources is another very active area of research for Mf-EBB problems. A potential approach is to take into account the fidelity cost by taking the ratio of the acquisition function at each of the fidelities and the fidelity cost ratio [20, 21]. This leads to sampling of cheaper low-fidelity sources even if the potential gain is lower than sampling more expensive sources, as the cost-to-gain ratio is larger. This approach however can suffer from the assumption that cost is a proxy for the quality of a source. Indeed, it is possible for a source that is cheap to be very accurate and for a source that is relatively costly to be inaccurate. More recent approaches have taken this into account when choosing which source to sample, focusing on Bf-EBB problems. The work of van Rijn et al. [48] presents a framework that estimates the benefit of further sampling a single low-fidelity source when minimizing the surrogate model error. This benefit is used to guide the division of the remaining available budget between high- and low-fidelity sources. Müller [33] proposes an optimization framework in which two surrogate models are trained, the first with only high-fidelity data and the second with low-fidelity data. The high-fidelity model is used to choose further sample locations, whereas the low-fidelity model is used to conduct a preliminary assessment of a newly chosen sample site only when the low-fidelity source is deemed to be trustworthy. Whilst these approaches are relatively new, the need for techniques that conduct some kind of analysis on low-fidelity sources before choosing how to use them is clear. The potential downfalls of not doing so are illustrated in the next section.

4. Illustrative example

As discussed in Section 1, a multitude of real-life problems consider processes that perfectly fit the definition of an expensive black-box. The work of Forrester et al. [17], for example, focuses on the design of a transonic civil aircraft wing. The wing's design is impacted by the values of 11 design parameters although, in their work, only four of

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

them are investigated: area, aspect ratio, sweep and inboard taper ratio. Their objective function is defined as the performance of the wing's design, which is measured as the drag-to-dynamic pressure ratio for a fixed lift. Two simulation tools are available to assess the performance of a particular design: a linearized potential method denoted VSaero [29] and a simpler empirical drag estimation code denoted Tadpole [11].

Given the available design parameters and examined output in this example, each of the three problems defined in Section 2 might arise given certain conditions. Problems 1 and 2 (defined in Sections 2.1 and 2.2, respectively) arise when the aim is to accurately model the impact of the wing's area, aspect ratio, sweep and inboard taper ratio effect on the drag to dynamic pressure ratio. Problem 1 defines the setting where the data have been collected by a different engineering team and the only choice available is how to use them to train a model. Problem 2 defines the setting where engineers are given the freedom to choose which design values to assess and which simulator to query for this assessment. Problem 3 (defined in Section 2.3) is the setting investigated in the study of Forrester et al. [17], where the aim is to optimize the wing's performance by minimizing the drag-to-dynamic pressure ratio. The potential benefits of employing multiple information sources are clearly shown in the study of this particular industry problem, as the solutions found by co-Kriging are of a greater quality than those found by Kriging. As is the case with this example, many real-life expensive outputs can be approximated via alternative means: employing domain knowledge of similar designs and products, using existing industrial models, and the usage of simulation engines to emulate physical testing. Whilst relying on these additional sources can certainly aid in circumventing the cost of analysis, the accuracy of these sources might be put into question.

The study of aircraft design is just one popular example of industrial challenges in which various products and processes can be tackled by expensive black-box modelling and optimization methods, using multi-fidelity data sources. Other varied examples with these characteristics include civil infrastructure projects [13, 40], materials and drug screening [9, 15], and biomanufacturing process modelling [44]. An intuitive two-dimensional example is now introduced to illustrate how techniques such as co-Kriging, which rely on multiple sources of data, can be affected by the quality of the data. In this example, we assume we are map makers tasked with exploring the yet-unknown elevation of Australia. An elevation map of the country is shown in Figure 4. Here the only way to exactly measure the elevation at a particular location is to send a team of engineers to take an accurate measurement. Australian elevation can therefore be assumed here to be a two-dimensional EBB denoted f_h. Australian elevation makes for a hard black-box to model and optimize. As shown in the plot, the chosen coordinate window includes three very deep (that is, objective function values lower than −4000 meters) regions at the bottom, top-left and top-right, a very flat land above sea level, and only a few regions of high altitude, with the highest point being Mt Kosciuszko in New South Wales at 2228 m shown as the middle red triangle in the group of three on the bottom right of the map.

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

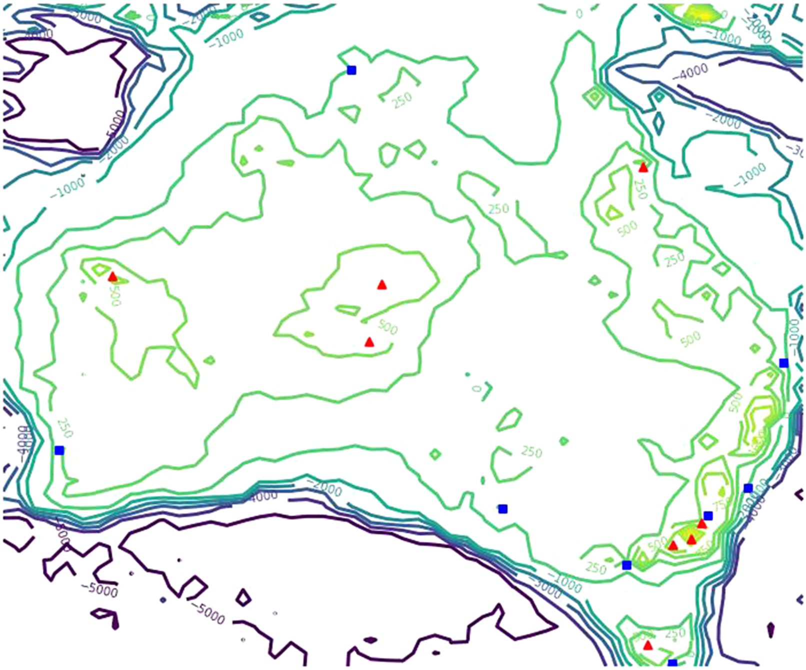

FIGURE 4. Elevation map of Australia. The blue squares mark the capital cities of the country and the red triangles mark the highest point in each state and territory.

Two additional low-fidelity sources are available in this example to aid in both exploration and optimization. The first is denoted f_l^{error} and is assumed to consist of existing inaccurate elevation records, generated here by adding a deterministic error to the true elevation and shown in Figure 5. Low-fidelity sources of this kind are often used in benchmarks in the literature. The second low-fidelity source is denoted f_l^{approx} and is generated with a simple procedure. For locations below sea-level (that is, for $f_h(\mathbf{x}) < 0$), this approximation returns a value of 0 (that is, $f_l^{\text{approx}}(\mathbf{x}) = 0$). Assuming that the elevation of the capital cities is known, at every location above sea-level, the low-fidelity source provides an approximation of the elevation by taking a distance-based weighted sum of the elevation of the two nearest capital cities. This source is shown in Figure 6. For simplicity, here we assume that both low-fidelity sources are cheap to sample, that is, the cost ratio is $C_r = 0$ for both sources; the added effect of having a non-cheap source is discussed further below.

Let us now analyse different problem settings to illustrate the three problem types defined in Section 2. Assume first that the data have already been gathered, that is, engineering teams have already been sent to different locations in Australia and queries have been made of the inaccurate records. Problem 1 defined in Section 2.1 corresponds here to the task of creating a model for the elevation of Australia based on this fixed sample. Two scenarios are compared. In the first one, 25 engineering teams

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

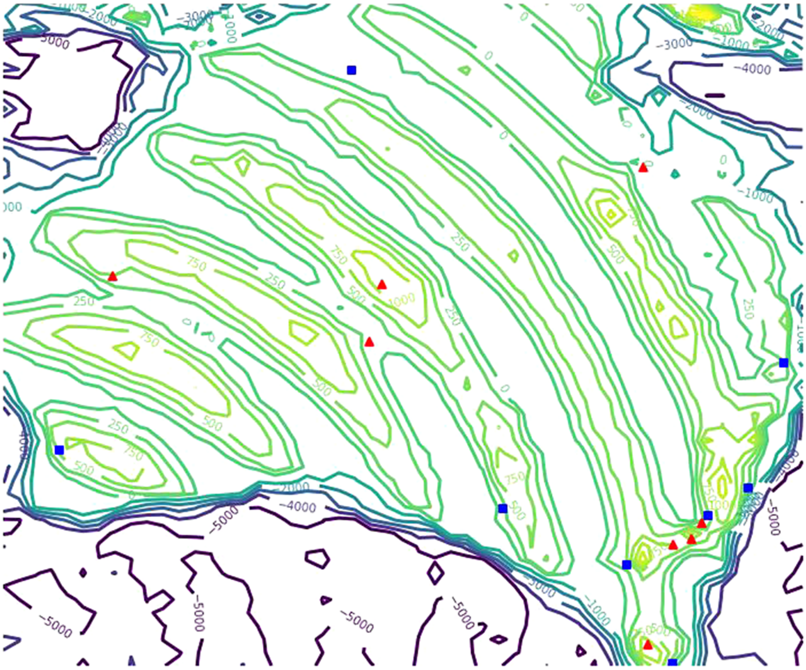

FIGURE 5. Low-fidelity source of the elevation of Australia, generated by adding a deterministic error to the true elevation. The blue squares mark the capital cities of the country, and the red triangles mark the highest point in each state and territory.

have been sent to collect data and 50 queries of the inexact records have been made. That is, 25 f_h samples and 50 f_l^{error} samples are available to train a model. In the second scenario, the same amount of data of both kinds is available, but all f_l^{error} data come from locations that are above sea level. This second scenario can be perhaps seen as more realistic if the source f_l^{error} is assumed to represent inaccurate records from city councils, which only exist on land. Figure 7 shows the locations at which these samples have been gathered for both scenarios. The accuracy of Kriging models trained only with high-fidelity data and co-Kriging models trained with high- and low-fidelity data in both scenarios is shown in Table 1. In this example, note that many factors can impact the benefit of relying on low-fidelity sources. The fact that relying on data from f_l^{error} can be beneficial is represented by the decreased error of co-Kriging models over Kriging models when these data have been collected evenly from the whole space. However, in the second scenario, co-Kriging models have a larger error than Kriging models, despite the same amount of low-fidelity data having been collected from the same source. As the only difference in the second scenario is the fact that low-fidelity data are restricted to lie above sea level, it is clear that when choosing whether to rely on a low-fidelity source, practitioners should carefully consider not only its quality but also the potential bias introduced by the sampling procedure.

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

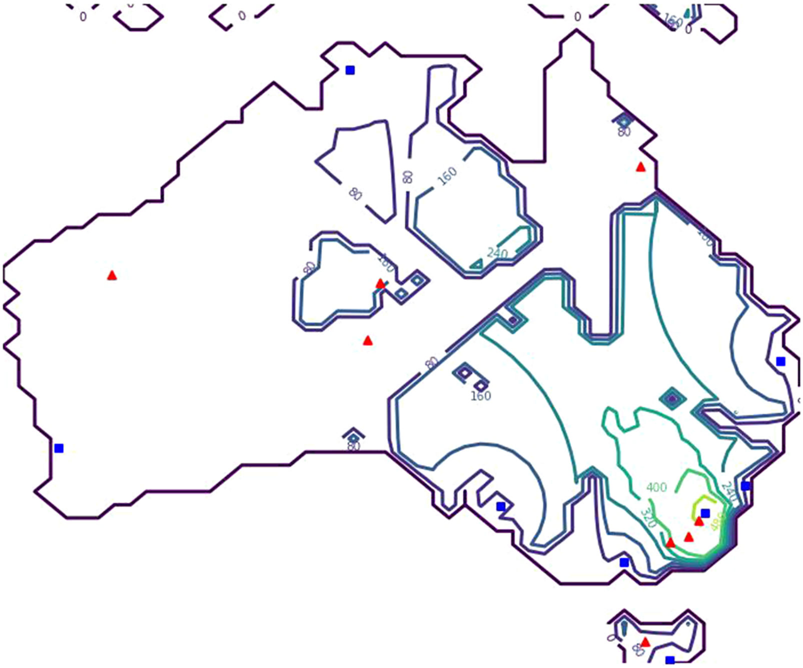

FIGURE 6. Low-fidelity source of the elevation of Australia, which is an approximation based on the known elevation of the capital cities. The blue squares mark the capital cities of the country, and the red triangles mark the highest point in each state and territory.

Let us now assume that we have been given more freedom of action when generating a map of Australia, that is, we are given the option to choose from where to gather data. The only given guideline is that we are restricted to sending a surveying team to at most 100 locations. This translates to Problem 2 defined in Section 2.2, namely, model creation with a sample budget. To assess the value of the low-fidelity sources, three approaches are compared. The first is to rely only on high-fidelity data via the use of a Kriging model. The second is to work with a co-Kriging model which relies on data from f_h and the low-fidelity source f_l^{error}. Finally, the third approach is to work with a co-Kriging model which relies on data from f_h and the low-fidelity source f_l^{approx}. Following the design of experiments outlined in Section 3.2, a set of 50 locations is chosen via a locally optimized LHS plan at which f_l^{error} and f_l^{approx} are sampled, and a subset of 25 locations is chosen at which f_h is sampled. The relevant data are used to train the Kriging model and the two co-Kriging models, and then further sampling locations are iteratively chosen by maximizing the model uncertainty, as described in Section 3.3. The process is stopped once the high-fidelity source has been sampled a total of 100 times. Repeating this approach 10 times and taking the average model error yields the results shown in Figure 8.

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

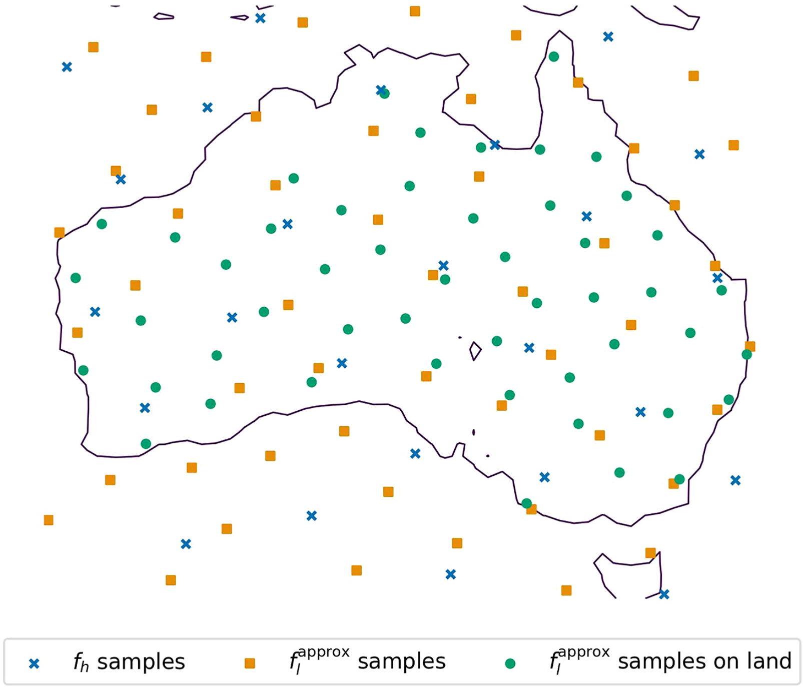
f_h samples
f_l^{approx} samples
f_l^{approx} samples on land

FIGURE 7. Locations at which samples have been gathered and made available to train a model. The blue crosses represent the 25 locations at which the high-fidelity source has been sampled. The orange squares represent 50 locations at which the low-fidelity source f_l^{approx} has been sampled and are spread out across the whole space. The green circles represent 50 locations at which the low-fidelity source f_l^{approx} has been sampled; this last set of samples has been restricted to lie on locations above sea level, marked by the shown outline.

TABLE 1. Error of constructed models for five repetitions. The second column shows the error of Kriging models constructed using 25 high-fidelity samples spread out across the space. The third column shows the error of co-Kriging models constructed with high-fidelity data and 50 samples from f_l^{error} spread out across the space. The fourth column shows the error of co-Kriging models constructed with high-fidelity data and 50 samples from f_l^{error} spread out across locations above sea level.

Repetition	Kriging	co-Kriging spread out	co-Kriging above sea-level
1	1308.55	1180.69	1465.15
2	1294.4	1195.26	1378.99
3	1294.44	1129.29	1590.07
4	1552.9	1175.21	1611.1
5	1366.34	1183.36	1399.48

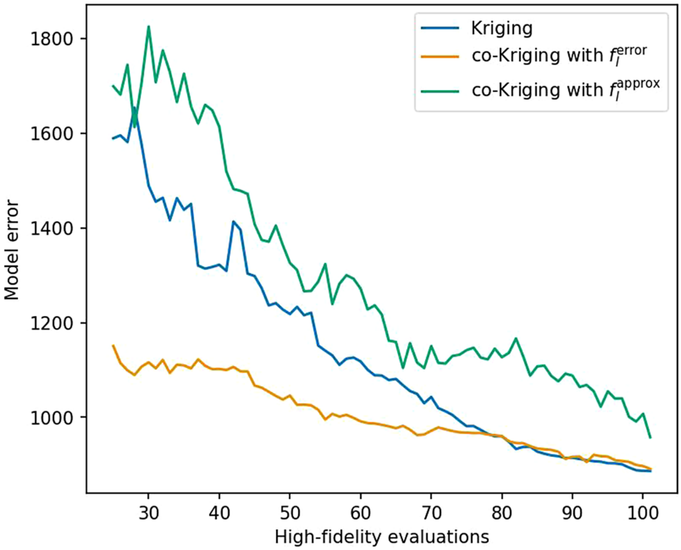

FIGURE 8. Average error of three surrogate models trying to model the elevation of Australia. The performance is shown for Kriging models which only use data from f_h (blue), co-Kriging models which use data from f_h and f_l^{error} (orange), and co-Kriging models which use data from f_h and f_l^{approx} (green).

It can be seen here that the low-fidelity source f_l^{error} is an asset when attempting to model the elevation of Australia, as co-Kriging models that rely on this source are the most accurate on average. Classical examples of this type have often been used in the literature to motivate the development of multi-fidelity techniques such as co-Kriging. It might be counter-productive to blindly rely on low-fidelity sources when they are available, however, as the same graph shows that co-Kriging models which rely on the source f_l^{approx} perform worse than Kriging models which only rely on high-fidelity data. This can be unsurprising, given that the source f_l^{approx} only gives an extremely rough outline of the elevation. This example nonetheless demonstrates that one must be wary of always using data simply because it is available. It is also worth noting that when much high-fidelity data is available (that is, towards the right of the graph), Kriging models and co-Kriging models trained with f_l^{error} data show similar performance. The plot's horizontal axis is based only on high-fidelity evaluations since so far, it has been assumed that the cost of sampling f_l^{error} was zero. It could be that this is not the case; assume for instance that obtaining the existing inaccurate elevation record of a particular location requires enquiring the archives and this is roughly ten times cheaper than sending a team of engineers to the location. In this case, the cost ratio C_r of f_l^{error} would be equal to 0.1 and the total budget spent at any given moment is impacted by the number of low-fidelity samples taken. If this were the case, the orange line should be "shifted" to the right, meaning that for large budgets, the best approach is to not rely on any low-fidelity data.

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

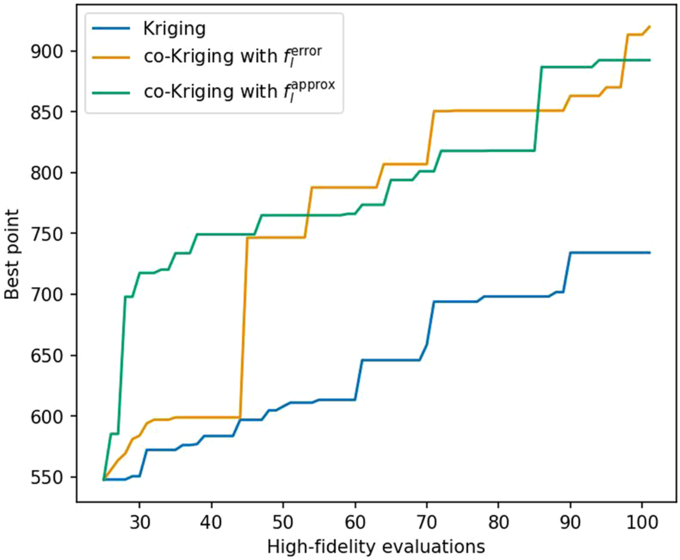

FIGURE 9. Average best point found during the optimization of three surrogate modelling techniques. The performance is shown for Kriging models which only use data from f_h (blue), co-Kriging models which use data from f_h and f_h^{error} (orange), and co-Kriging models which use data from f_h and f_h^{approx} (green).

Let us now focus on an optimization task. Assume that we have been tasked with finding the tallest point in Australia, once again with an overall budget of 100 samples. This translates to Problem 3 defined in Section 2.3, namely function optimization with sample budget. Once again, the performance of Kriging models, co-Kriging models trained with f_l^{error} data and co-Kriging models trained with f_l^{approx} data is compared using the same experimental setup. The only difference now is that further samples are iteratively chosen based on the expected improvement measure presented in Section 3.3. The results are shown in Figure 9. Perhaps unsurprisingly, relying on the source f_l^{error} is an improvement over using only high-fidelity data. Since this source helps generate more accurate models, it seems a natural conclusion that it should help find better points during optimization. What is perhaps unexpected is the fact that relying on the source f_l^{approx} leads to better points being found almost immediately, despite this source being counter-productive for surrogate model accuracy. This behaviour can be explained by the fact that the capital city Canberra is at an elevation of 554 m. As this low-fidelity approximation is based on the elevation of the capital cities and every other city is at sea level, the approximation leads to an almost flat objective function with only a high-elevation area in the bottom right of the map. This happens to coincide with the highest region of the true elevation of Australia. Because of this, co-Kriging models trained with these data choose to sample in that region, leading to good points being found very quickly. Note that the low-fidelity source does not think that the area near Canberra is very high (as it thinks

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

it has a maximum elevation of 500 m), but it thinks it is the highest area and this is sufficient. From a mathematical standpoint, this can be described as f_l^{approx} and f_h having a large difference, but being highly correlated in the highest regions of f_h. Once again, the fact that a low-fidelity source is harmful for model creation with a sample budget, but is beneficial for function optimization with a sample budget, speaks to the complexity of characterizing beneficial low-fidelity sources and of creating algorithms that intelligently exploit or ignore these sources. Finally, note that none of these examples involved the three sources being used at the same time. The question of reliance on low-fidelity sources only grows in difficulty when more of them are made available at the same time.

5. Conclusion

This tutorial-style paper has provided an introduction for new practitioners to surrogate modelling techniques for Mf-EBB problems. The focus has been on the definition of different problem variations and the use of traditional approaches, such as training co-Kriging models and using expected improvement as the acquisition function. As alluded to throughout the paper, many new techniques have been created since the multi-fidelity data fusion of Kennedy and O'Hagan [24] was first proposed over 20 years ago. However, a key aspect of methods for Mf-EBB problems, which still seems to lack a satisfying level of understanding, is how the quality of low-fidelity sources can impact multi-fidelity techniques and how this impact can be mitigated. The elevation example in particular has helped to highlight that characterizing and exploiting helpful low-fidelity sources is far from trivial. Indeed, relying on beneficial low-fidelity sources can become a sub-optimal choice if their domain is restricted to certain regions in the space or their cost is too great, and harmful low-fidelity sources when constructing accurate surrogate models can be very beneficial when used during an optimization process.

Strategic questions such as how to divide a total budget among an initial design of experiments and how to further divide the sample among sources also still lack a consensus from the literature. These questions are particularly important for new techniques that seek to learn whether a source is helpful or harmful before relying on them, as it is crucial to strike the right balance between spending long enough correctly characterizing a source and being left with enough budget to exploit this characterization. The recent work of Toal [45] and Andrés-Thió et al. [3] on the analysis of low-fidelity source quality impact on co-Kriging model performance has shown this area to require further study, particularly as it can provide insights into the creation of new adaptive techniques. Their work however has focused on the synthetic setting, using a large amount of data to characterize these sources which is not available in practice. A characterization based only on the data available to train these surrogate models appears to still be missing. Finally, it might be possible to create a methodology that does not choose what sources are worth using, but rather filters out only the harmful data from a particular source and relies only on localized helpful information. This finer approach would benefit from sources that are highly

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

accurate in some regions, but very inaccurate in others. Whether this is a feasible approach remains to be seen.

Acknowledgements

This research was supported by the Australian Research Council under grant number IC200100009 for the ARC Training Centre in Optimisation Technologies, Integrated Methodologies and Applications (OPTIMA). This research was also supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative. The first author is also supported by a Research Training Program Scholarship from The University of Melbourne. The authors would also like to thank Dr Francis Valentinis and Dr Vera Somers for useful discussions of industrial problems and the illustrative example.

References

[1] N. Alemazkoor and H. Meidani, "A data-driven multi-fidelity approach for traffic state estimation using data from multiple sources", *IEEE Access* **9** (2021) 78128–78137; doi:10.1109/ACCESS.2021.3081063.

[2] N. Andrés-Thió, "Bifidelity surrogate modelling benchmark problems", *GitHub repository* (2023); doi:10.5281/zenodo.8353690.

[3] N. Andrés-Thió, M. A. Muñoz and K. Smith-Miles, "Bifidelity surrogate modelling: showcasing the need for new test instances", *INFORMS J. Comput.* **34** (2022) 3007–3022; doi:10.1287/ijoc.2022.1217.

[4] B. Ankenman, B. L. Nelson and J. Staum, "Stochastic kriging for simulation metamodeling", in: *2008 Winter Simulation Conference, Miami, Fl, USA, 2008* (ed. M. Kam) (IEEE, New York, 2008) 362–370; doi:10.1109/WSC.2008.4736089.

[5] M. J. Appel, R. Labarre and D. Radulovic, "On accelerated random search", *SIAM J. Optim.* **14** (2004) 708–731; doi:10.1137/S105262340240063X.

[6] R. C. Arenzana, A. F. López-Lopera, S. Mouton, N. Bartoli and T. Lefebvre, "Multi-fidelity Gaussian process model for CFD and wind tunnel data fusion", in: *AeroBest 2021, Lisbonne, Portugal, 2021* (Centre national de la recherche scientifique, Paris, 2021) Article ID: hal-03346321; https://hal.science/hal-03346321/document.

[7] D. Bertsimas and J. Tsitsiklis, "Simulated annealing", *Statist. Sci.* **8** (1993) 10–15; doi:10.1214/ss/1177011077.

[8] M. Binois, J. Huang, R. B. Gramacy and M. Ludkovski, "Replication or exploration? Sequential design for stochastic simulation experiments", *Technometrics* **61** (2019) 7–23; doi:10.1080/00401706.2018.1469433.

[9] D. Buterez, J. P. Janet, S. J. Kiddle and P. Liò, "MF-PCBA: multifidelity high-throughput screening benchmarks for drug discovery and machine learning", *J. Chem. Inform. Model.* **63** (2023) 2667–2678; doi:10.1021/acs.jcim.2c01569.

[10] S. Cheng, B. A. Konomi, J. L. Matthews, G. Karagiannis and E. L. Kang, "Hierarchical Bayesian nearest neighbor co-Kriging Gaussian process models; an application to intersatellite calibration", *Spat. Stat.* **44** (2021) Article ID: 100516; doi:10.1016/j.spasta.2021.100516.

[11] J. Cousin and M. Metcalfe, "The BAe (commercial aircraft) LTD transport aircraft synthesis and optimisation program (TASOP)", in: *Aircraft Design, Systems and Operations Conference (Dayton, OH, USA, 17–19 September 1990)* (American Institute of Aeronautics and Astronautics, Reston, VA, 1990); doi:10.2514/6.1990-3295.

[12] J. Duchon, "Splines minimizing rotation-invariant semi-norms in Sobolev spaces", in: *Constructive theory of functions of several variables* (eds. W. Schempp and K. Zeller) (Springer, Berlin–Heidelberg, 1977) 85–100; doi:10.1007/BFb0086566.

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

[13] B. Elshafei, A. Peña, D. Xu, J. Ren, J. Badger, F. M. Pimenta, D. Giddings and X. Mao, "A hybrid solution for offshore wind resource assessment from limited onshore measurements", *Appl. Energy* **298** (2021) Article ID: 117245; doi:10.1016/j.apenergy.2021.117245.

[14] J. T. Eweis-Labolle, N. Oune and R. Bostanabad, "Data fusion with latent map Gaussian processes", *J. Mech. Des.* **144** (2022) Article ID: 091703; doi:10.1115/1.4054520.

[15] C. Fare, P. Fenner, M. Benatan, A. Varsi and E. O. Pyzer-Knapp, "A multi-fidelity machine learning approach to high throughput materials screening", *npj Comput. Mater.* **8** (2022) Article ID: 257; doi:10.1038/s41524-022-00947-9.

[16] M. G. Fernández-Godino, C. Park, N. H. Kim and R. T. Haftka, "Issues in deciding whether to use multifidelity surrogates", *AIAA J.* **57** (2019) 2039–2054; doi:10.2514/1.J057750.

[17] A. I. J. Forrester, A. Sóbester and A. J. Keane, "Multi-fidelity optimization via surrogate modelling", *Proc. Roy. Soc. A* **463**(2088) (2007) 3251–3269; doi:10.1098/rspa.2007.1900.

[18] H.-M. Gutmann, "A radial basis function method for global optimization", *J. Global Optim.* **19** (2001) 201–227; doi:10.1023/A:1011255519438.

[19] Z.-H. Han and S. Görtz, "Hierarchical Kriging model for variablefidelity surrogate modeling", *AIAA J.* **50** (2012) 1885–1896; doi:10.2514/1.J051354.

[20] X. He, R. Tuo and C. J. Wu, "Optimization of multi-fidelity computer experiments via the EQIE criterion", *Technometrics* **59** (2017) 58–68; doi:10.1080/00401706.2016.1142902.

[21] D. Huang, T. T. Allen, W. I. Notz and R. A. Miller, "Sequential Kriging optimization using multiple-fidelity evaluations", *Struct. Multidiscip. Optim.* **32** (2006) 369–382; doi:10.1007/s00158-005-0587-0.

[22] D. R. Jones, "A taxonomy of global optimization methods based on response surfaces", *J. Global Optim.* **21** (2001) 345–383; doi:10.1023/A:1012771025575.

[23] D. R. Jones, M. Schonlau and W. J. Welch, "Efficient global optimization of expensive black-box functions", *J. Global Optim.* **13** (1998) 455–492; doi:10.1023/A:1008306431147.

[24] M. C. Kennedy and A. O'Hagan, "Predicting the output from a complex computer code when fast approximations are available", *Biometrika* **87** (2000) 1–13; doi:10.1093/biomet/87.1.1.

[25] D. G. Krige, "A statistical approach to some basic mine valuation problems on the Witwatersrand", *J. Southern African Inst. Min. Metallurgy* **52** (1951) 119–139; https://hdl.handle.net/10520/AJA0038223X˙4792.

[26] H. B. Kurt, M. Millidere, F. S. Gomec and O. Ugur, "Multi-fidelity aerodynamic dataset generation of a fighter aircraft", in: *AIAA SciTech 2021 forum* (American Institute of Aeronautics and Astronautics, Reston, VA, 2021) Article ID: AIAA 2021-0544; doi:10.2514/6.2021-0544.

[27] R. Le Riche and V. Picheny, "Revisiting Bayesian optimization in the light of the COCO benchmark", *Struct. Multidiscip. Optim.* **64** (2021) 3063–3087; doi:10.1007/s00158-021-02977-1.

[28] X. Liu, W. Zhao and D. Wan, "Multi-fidelity co-Kriging surrogate model for ship hull form optimization", *Ocean Eng.* **243** (2022) Article ID: 110239; doi:10.1016/j.oceaneng.2021.110239.

[29] B. Maskew, "Prediction of subsonic aerodynamic characteristics: a case for low-order panel methods", *J. Aircraft* **19** (1982) 157–163; doi:10.2514/3.57369.

[30] G. Matheron, "Principles of geostatistics", *Econ. Geol.* **58** (1963) 1246–1266; doi:10.2113/gsecongeo.58.8.1246.

[31] G. Matheron, "The intrinsic random functions and their applications", *Adv. in Appl. Probab.* **5** (1973) 439–468; doi:10.2307/1425829.

[32] N. Mourousias, A. Malim, B. G. Marinus and M. Runacres, "Assessment of multi-fidelity Surrogate models for high-altitude propeller optimization", in: *AIAA AVIATION 2022 Forum, Chicago, IL, USA, 27 June–1 July, 2022* (American Institute of Aeronautics and Astronautics, Reston, VA, 2022) Article ID: AIAA 2022-3752; doi:10.2514/6.2022-3752.

[33] J. Müller, "An algorithmic framework for the optimization of computationally expensive bi-fidelity black-box problems", *INFOR Inf. Syst. Oper. Res.* **58** (2020) 264–289; doi:10.1080/03155986.2019.1607810.

[34] J. Müller and C. A. Shoemaker, "Influence of ensemble surrogate models and sampling strategy on the solution quality of algorithms for computationally expensive black-box global optimization problems", *J. Global Optim.* **60** (2014) 123–144; doi:10.1007/s10898-014-0184-0.

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press

[35] C. Park, R. T. Haftka and N. H. Kim, “Remarks on multi-fidelity surrogates”, *Struct. Multidiscip. Optim.* **55** (2017) 1029–1050; doi:10.1007/s00158-016-1550-y.

[36] J.-S. Park, “Optimal Latin-hypercube designs for computer experiments”, *J. Statist. Plann. Inference* **39** (1994) 95–111; doi:10.1016/0378-3758(94)90115-5.

[37] X. Peng, J. Kou and W. Zhang, “Multi-fidelity nonlinear unsteady aerodynamic modeling and uncertainty estimation based on hierarchical Kriging”, *Appl. Math. Model.* **122** (2023) 1–21; doi:10.1016/j.apm.2023.05.031.

[38] K. R. Quinlan, J. Movva, E. V. Stein and A. Kupresanin, “Multi-fidelity aerodynamic databases for efficient representation of hypersonic design spaces”, Technical report (Lawrence Livermore National Lab (LLNL), Livermore, CA, USA, 2021) https://www.osti.gov/biblio/1831394.

[39] R. G. Regis and C. A. Shoemaker, “A stochastic radial basis function method for the global optimization of expensive functions”, *INFORMS J. Comput.* **19** (2017) 497–509; doi:10.1287/ijoc.1060.0182.

[40] T. Savage, N. Basha, J. McDonough, O. K. Matar and E. A. del Rio Chanona, “Multi-fidelity data-driven design and analysis of reactor and tube simulations”, *Comput. Chem. Eng.* **179** (2023) Article ID: 108410; doi:10.1016/j.compchemeng.2023.108410.

[41] T. Scholcz and J. Klinkenberg, “Hull-shape optimisation using adaptive multi-fidelity Kriging”, in: *NATO-AVT-354 workshop on multi-fidelity methods for military vehicle design, Varma, Bulgaria* (NATO Science and Technology Organisation, Paris, 2022); https://www.researchgate.net/publication/364346359_Hull-Shape_Optimisation_Using_Adaptive_Multi-Fidelity_Kriging.

[42] S. Seo, M. Wallat, T. Graepel and K. Obermayer, “Gaussian process regression: active data selection and test point rejection”, in: *Mustererkennung 2000: 22. DAGM-Symposium, Kiel, 13–15 September 2000* (eds. G. Sommer, N. Krüger and C. Perwass) (Springer, Berlin–Heidelberg, 2000) 27–34; doi: 10.1007/978-3-642-59802-9_4.

[43] H. P. Solak et al., “Hydrofoil optimization via automated multi-fidelity surrogate models”, in: *10th International Conference on Computational Methods in Marine Engineering, Madrid, Spain, 2023* (eds. J. García-Espinosa, L. González, J. E. Gutiérrez and B. Serván-Camas) (Marine, HAL Open Science, 2023) Article ID: hal-04089604; https://hal.science/hal-04089604/.

[44] Y. Sun, W. Nathan-Roberts, T. D. Pham, E. Otte and U. Aickelin, “Multi-fidelity Gaussian process for biomanufacturing process modeling with small data”, Preprint, 2022, arXiv.2211.14493.

[45] D. J. J. Toal, “Some considerations regarding the use of multifidelity Kriging in the construction of surrogate models”, *Struct. Multidiscip. Optim.* **51** (2015) 1223–1245; doi:10.1007/s00158-014-1209-5.

[46] D. J. J. Toal, N. W. Bressloff and A. J. Keane, “Kriging hyperparameter tuning strategies”, *AIAA J.* **46** (2008) 1240–1252; doi:10.2514/1.34822.

[47] D. J. J. Toal, N. W. Bressloff, A. J. Keane and C. M. E. Holden, “The development of a hybridized particle swarm for kriging hyperparameter tuning”, *Eng. Optim.* **43** (2011) 675–699; doi:10.1080/0305215X.2010.508524.

[48] S. van Rijn, S. Schmitt, M. van Leeuwen and T. Bäck, “Finding efficient trade-offs in multi-fidelity response surface modelling”, *Eng. Optim.* **55** (2023) 946–963; doi:10.1080/0305215X.2022.2052286.

[49] H. Wang, Y. Jin and J. Doherty, “A generic test suite for evolutionary multifidelity optimization”, *IEEE Trans. Evol. Comput.* **22** (2018) 836–850; doi:10.1109/TEVC.2017.2758360.

[50] R. Wang, Y. Yang, X. Wang, B. Wang and G. Zhang, “Co-Kriging based multi-fidelity aerodynamic optimization for flying wing UAV with multi-shape wingtip design”, in: *2021 IEEE International Conference on Unmanned Systems (ICUS), Beijing, China* (ed. A. Ephremides) (IEEE, New York, 2021) 93–98; doi:10.1109/ICUS52573.2021.9641491.

[51] S. M. Wild, R. G. Regis and C. A. Shoemaker, “Orbit: optimization by radial basis function interpolation in trust-regions”, *SIAM J. Sci. Comput.* **30** (2008) 3197–3219; doi:10.1137/070691814.

[52] S. Yang and K. Yee, “Design rule extraction using multi-fidelity surrogate model for unmanned combat aerial vehicles”, *J. Aircraft* **59** (2022) 977–991; doi:10.2514/1.C036489.

https://doi.org/10.1017/S1446181124000087 Published online by Cambridge University Press