Proceedings of the ASME 2023
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC/CIE2023
August 20-23, 2023, Boston, Massachusetts

DETC2023-XXXXXX

MITIGATING THE EFFECTS OF SOURCE-DEPENDENT BIAS AND NOISE ON MULTI-SOURCE BAYESIAN OPTIMIZATION: APPLICATION TO MATERIALS DESIGN

Zahra Zanjani Foumani, Amin Yousefpour, Mehdi Shishehbor, Ramin Bostanabad[1]
Department of Mechanical and Aerospace Engineering, University of California, Irvine
Irvine, CA, USA

ABSTRACT

Bayesian optimization (BO) is a sequential optimization strategy that is increasingly employed in a wide range of areas including materials design and drug discovery. In real world applications, acquiring high-fidelity (HF) data is the major cost component of BO, since most of the problems demand the use of expensive HF simulations. To alleviate this bottleneck, multi-fidelity (MF) methods are proposed to forgo the sole reliance on the expensive HF data and reduce the sampling costs by querying inexpensive low-fidelity (LF) sources whose data are correlated with HF samples. Existing multi-fidelity BO (MFBO) methods operate under the following two assumptions: (1) Leveraging global (rather than local) correlation between HF and LF sources, and (2) Associating all the data sources with the same noise process. These assumptions dramatically reduce the performance of MFBO when LF sources are only locally correlated with the HF source or when the noise variance varies across the data sources. To dispense with these incorrect assumptions, we propose an MF emulation method that learns a source-dependent noise process and also enables BO to leverage highly biased LF sources which are only locally correlated with the HF source. We illustrate the performance of our method through analytical examples and engineering problems on materials design.

Keywords: Bayesian optimization; multi-fidelity modeling; emulation; heterogenous noise modeling; interval score.

1. INTRODUCTION

Bayesian optimization (BO) is a sequential and sample-efficient global optimization technique that is increasingly used in the optimization of expensive-to-evaluate (and typically black-box) functions [1]. BO has two main ingredients: an emulator which is typically a Gaussian process (GP) and an acquisition function (AF) [2]. The first step in BO is to train an emulator on some initial data. Then, an auxiliary optimization is solved to determine the new sample that should be added to the training data. The objective function of this auxiliary optimization is the AF whose evaluation relies on the emulator. Given the new sample, the training data is updated and the entire emulation-sampling process is repeated until the convergence conditions are met [3].

Although BO is a highly efficient technique, the total cost of optimization can be substantial if it solely relies on the accurate but expensive high-fidelity (HF) data source. To mitigate this issue, multi-fidelity (MF) techniques are widely adopted [4-6] where one uses multiple data sources of varying levels of accuracy and cost in BO. The fundamental principle behind MF techniques is to exploit the correlation between low-fidelity (LF) and HF data to decrease the overall sampling costs [7, 8].

Over the past two decades many multi-fidelity BO (MFBO) strategies have been proposed which primarily differ in terms of their emulator and AF. Most existing strategies rely on the Co-Kriging method [9], Kennedy and O'Hagan's bi-fidelity approach [10], and the BoTorch package [11]. These MFBO methods have some major drawbacks such as inability to simultaneously leverage multiple LF sources, sensitivity to the sampling costs (where highly inexpensive LF sources cause numerical and convergence issues), and presuming simple bias forms for the LF sources.

Some of these limitations are recently addressed in [12] where the authors propose to (1) use latent map Gaussian processes (LMGPs) for emulation, and (2) quantify the information value of LF and HF samples differently. Their AF is cost-aware in that it considers the sampling cost in quantifying the value of HF and LF data points. Henceforth, we refer to this method as $MFBO_{\alpha}$.

While $MFBO_{\alpha}$ performs much better than competing MF approaches, it has two main limitations which are demonstrated

[1] Corresponding author. Email: Raminb@uci.edu

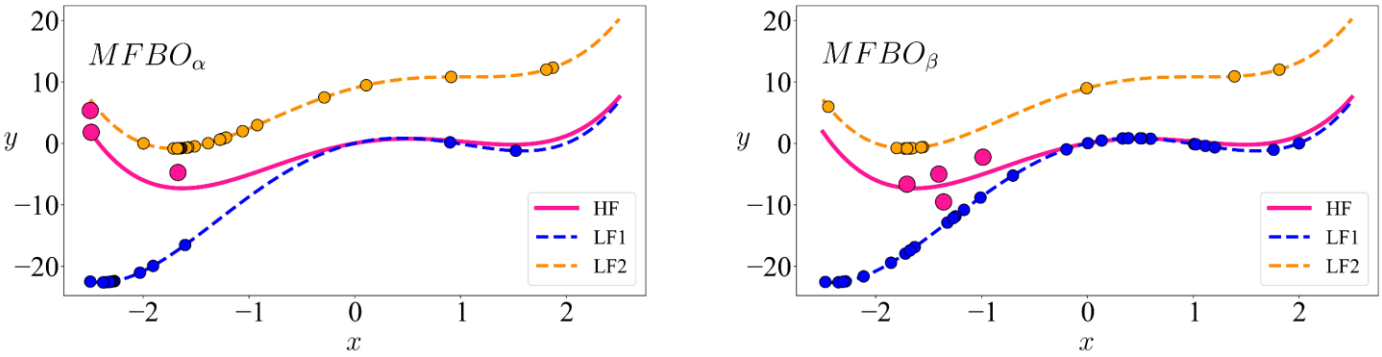

FIGURE 1 EFFECT OF HETEROGENOUS NOISE AND MODEL FIDELITY ON MFBO: HF data are noisy and expensive while the LF data are deterministic and cheap. In this example, LF1 is more correlated with the HF source for $x > 0$ while LF2 correlates better with HF for $x < 0$. The sampling cost of the HF and two LF sources are 1000/100/100, respectively. $MFBO_\beta$ is the approach we propose in this paper to increase sampling efficiency and solution accuracy. $MFBO_\beta$ more effectively explores the space (as it samples more points in $x < 0$) and better leverages LF1 in $x > 0$. As for LF2, $MFBO_\beta$ mostly samples from $x < 0$, since this region includes the optimum of LF2 and is more correlated with the HF. Initial data are not shown in these figures. $MFBO_\alpha$ uses both LF sources and learns a single noise process for the three sources.

with a simple 1D example in FIGURE 1. Firstly, $MFBO_\alpha$ excludes highly biased LF sources from BO with the rationale that they can steer the search process in the wrong direction. This exclusion is done before BO starts since the decision is made based on the fidelity manifold of LMGP that is trained on the *initial* data. However, this manifold only quantifies global accuracy of LF sources with respect to the HF source. That is, if an LF source is only correlated with the HF data in a small region, $MFBO_\alpha$ discards it. We argue that such an early exclusion is suboptimal since that small region may contain the global optimum of the HF source.

The second limitation of $MFBO_\alpha$ is that it assumes all sources are corrupted with the same noise process (with unknown noise variance). However, MF datasets typically have different levels of noise especially if some sources represent deterministic computer simulations while others are physical experiments [13, 14]. In such applications, $MFBO_\alpha$ overestimates the uncertainties which, in turn, reduces the performance of MFBO.

To address these two limitations, we introduce $MFBO_\beta$ for multi-fidelity cost-aware Bayesian optimization. $MFBO_\beta$ has the same AFs as $MFBO_\alpha$ and can leverage an arbitrary number of LF sources in optimizing an HF source. Unlike $MFBO_\alpha$, $MFBO_\beta$ never discards any LF sources (regardless of its bias with respect to the HF source) and estimates a noise process for each data source. We argue that $MFBO_\beta$ quantifies the uncertainties more accurately than $MFBO_\alpha$ and thus achieves a higher performance in MFBO. FIGURE 1 schematically demonstrates the advantages of $MFBO_\beta$ over $MFBO_\alpha$ in a 1D example where there are one HF and two LF sources.

The rest of the paper is organized as follows. We provide the methodological details in Section 2 and then evaluate the performance of $MFBO_\beta$ via multiple ablation studies in Section 3. We conclude the paper in Section 4 by summarizing our contributions and providing future research directions.

2. METHODS

In this section, we first provide some background on LMGP and MF modeling with LMGP in Section 2.1 and Section 2.2, respectively. We then propose our efficient mechanism for inversely learning a noise process for each data source in Section 2.3 . Next, we introduce the cost-aware AF of $MFBO_\beta$ in Section 2.4. Finally, in Section 2.5 we elaborate on our idea that improves the uncertainty quantification (UQ) capabilities of LMGPs and, in turn, benefits MFBO.

2.1 Latent Map Gaussian Process (LMGP)

Gaussian processes (GPs) are emulators which assume the training data come from a multivariate normal distribution with parametric mean and covariance functions. Following this assumption, the training data can be modeled as:

$$y(\boldsymbol{x}) = \beta + \xi(\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{x} = [x_1, x_2, \dots, x_{dx}]^T$ is the input vector, $y(\boldsymbol{x})$ is the output, β is an unknown coefficient, and $\xi(\boldsymbol{x})$ is a zero-mean GP with the covariance function:

$$cov\big(\xi(\boldsymbol{x}), \xi(\boldsymbol{x}')\big) = c(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 r(\boldsymbol{x}, \boldsymbol{x}') \tag{2}$$

where σ^2 is the variance of the process and $r(\cdot,\cdot)$ is the parametric correlation function. In this paper we use the Gaussian correlation function defined as:

$$r(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\sum_{i=1}^{dx} 10^{\omega_i} (x_i - x_i')^2\right\} \tag{3}$$

where $\boldsymbol{\omega} = [\omega_1, \omega_2, \dots, \omega_{dx}]^T$ are the scale parameters. GP modeling highly depends on the choice of the correlation

function which measures the distance between any two points. To directly use GPs in MF modeling, we follow [15] who convert MF modeling to a manifold learning problem via LMGPs which are extensions of GPs that can handle categorical data [16] while providing a visualizable manifold that can be used to interpret the correlation among data sources.

Denoting the categorical inputs by $\boldsymbol{t} = [t_1, t_2, \dots, t_{dt}]^T$ where variable t_i has l_i distinct levels, LMGP maps each combination of the categorical levels to a point in a learned quantitative manifold. To this end, LMGP assigns a unique vector to each combination of the categorical variables and then learns a linear transformation that maps these unique vectors into a compact manifold with dimensionality dz:

$$\boldsymbol{z}(\boldsymbol{t}) = \boldsymbol{\zeta}(\boldsymbol{t})\boldsymbol{A} \tag{4}$$

where $\boldsymbol{t}$ denotes a specific combination of the categorical variables, $\boldsymbol{z}(\boldsymbol{t})$ is the $1 \times dz$ posterior latent representation of $\boldsymbol{t}$, $\boldsymbol{\zeta}(\boldsymbol{t})$ is a unique prior vector representation of $\boldsymbol{t}$, and $\boldsymbol{A}$ is a rectangular matrix that maps $\boldsymbol{\zeta}(\boldsymbol{t})$ to $\boldsymbol{z}(\boldsymbol{t})$. In this paper, grouped one-hot encoding is used to generate the prior vectors and hence the dimensionality of $\boldsymbol{\zeta}(\boldsymbol{t})$ and $\boldsymbol{A}$ are $1 \times \sum_{i=1}^{dt} l_i$ and $\sum_{i=1}^{dt} l_i \times dz$, respectively. These mapped points encode the data source and can be directly embedded in the correlation function as:

$$\begin{aligned} r(\boldsymbol{u}, \boldsymbol{u}') = \exp\left\{-\sum_{i=1}^{dx} 10^{\omega_i} (x_i - x_i')^2\right\} \\ \times \exp\left\{-\sum_{i=1}^{dz} \big(z_i(\boldsymbol{t}) - z_i(\boldsymbol{t}')\big)^2\right\} \end{aligned} \tag{5}$$

where $\boldsymbol{u} = [\boldsymbol{x}; \boldsymbol{t}]$ and $\boldsymbol{z}(\boldsymbol{t}) = [z_1(\boldsymbol{t}), z_2(\boldsymbol{t}), \dots, z_{dz}(\boldsymbol{t})]$ is the location in the learned latent space corresponding to the specific combination of the categorical variables denoted by $\boldsymbol{t}$.

LMGP estimates the hyperparameters $(\beta, \boldsymbol{A}, \boldsymbol{\omega}, \sigma^2)$ via maximum a posteriori (MAP) and then uses the conditional distribution formulas to predict the response distribution at the arbitrary point $\boldsymbol{u}$ with the following mean and variance:

$$\mathbb{E}[y(\boldsymbol{u})] = \mu(\boldsymbol{u}) = \hat{\beta} + \boldsymbol{r}^T(\boldsymbol{u})\, \boldsymbol{R}^{-1}\big(\boldsymbol{y} - \boldsymbol{1}_{n\times 1}\hat{\beta}\big) \tag{6}$$

$$\begin{aligned} c\big(y(\boldsymbol{u}), y(\boldsymbol{u})\big) = \sigma^2(\boldsymbol{u}) \\ = \hat{\sigma}^2(1 - \boldsymbol{r}^T(\boldsymbol{u})\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{u}) \\ + (g(\boldsymbol{u}))^2(\boldsymbol{1}_{1\times n}\boldsymbol{R}^{-1}\boldsymbol{1}_{n\times 1})^{-1}) \end{aligned} \tag{7}$$

where n is number of training samples, $\mathbb{E}$ denotes expectation, $\boldsymbol{1}_{a\times b}$ is an $a \times b$ matrix of ones, $\boldsymbol{r}(\boldsymbol{u})$ is an $n \times 1$ vector with the i^{th}element $r(\boldsymbol{u}^i, \boldsymbol{u})$, $\boldsymbol{R}$ is an $n \times n$ matrix with $R_{ij} = r(\boldsymbol{u}^i, \boldsymbol{u}^j)$, and $g(\boldsymbol{u}) = 1 - \boldsymbol{1}_{1\times n}\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{u})$.

2.2 Multi-fidelity Emulation via LMGP

The first step to MF emulation with LMGP is to augment the inputs with the additional categorical variable s that indicates the sources of sample, i.e., $s = \{'1', '2', \dots, 'ds'\}$ where the j^{th} element corresponds to source j for $j = 1, \dots, ds$. Then, the training data from all sources are concatenated and used in LMGP to build an MF emulator. We refer the readers to [17] for more detail but note here that in case the input variables already contain some categorical features (see Section 3.2 for an example), we endow LMGP with two manifolds where one encodes the fidelity variable s while the other manifold encodes the rest of the categorical variables.

Oune, Bostanabad [18] show that LMGPs have the following primary advantages over other MF emulators: (1) they provide a more flexible and accurate mechanism to build MF emulators since they learn the relations between the sources in a nonlinear manifold, (2) they learn all the sources quite accurately rather than just emulating the HF source, and (3) they provide a visualizable global metric for comparing the relative discrepancies/similarities among the data sources.

2.3 Source-dependent Noise Modeling

The presence of noise significantly affects the performance of BO, and incorrectly modeling it can cause over-exploration or under-exploration of the search space. To mitigate the effects of noise in BO, we reformulate LMGPs to independently model a noise process for each data source. This reformulation can improve the accuracy of the model in noisy regions and, in turn, guide the search toward the global optimum when the modeled is deployed in MFBO.

To model noise in GPs, the nugget or jitter parameter, δ, is used [19] to replace $\boldsymbol{R}$ with $\boldsymbol{R}_\delta = \boldsymbol{R} + \delta\boldsymbol{I}$ where $\boldsymbol{I}$ is an $n \times n$ identity matrix. With this approach, the estimated stationary noise variance in the data is $\delta\sigma^2$ and the mean and variance formulations in Eq. (6) and Eq. (7) are modified by using $\boldsymbol{R}_\delta$ instead of $\boldsymbol{R}$.

Although incorporating this modification in the correlation matrix can enhance the performance of the emulator and BO in single-fidelity (SF) problems, it does not yield the same benefits in MF optimization. This is likely because of the dissimilar nature of the data sources and their corresponding noises. When dealing with multiple sources of data, each source may suffer from different levels and types of noise. Consider a bi-fidelity dataset where the HF data comes from an experimental setup and is subject to measurement noise, while the LF data is generated by a deterministic computer code which has a systematic bias due to missing physics. In this case, using only one nugget parameter in LMGP for MF emulation is obviously not an optimum choice.

To address this issue effectively, we propose to use multiple nugget parameters in the emulator. Specifically, we define the nugget vector $\boldsymbol{\delta} = [\delta_1, \delta_2, \dots, \delta_{ds}]$ and update the correlation matrix as follows:

$$\boldsymbol{R}_\delta = \boldsymbol{R} + \boldsymbol{N}_\delta \tag{8}$$

where $\boldsymbol{N}_\delta$ denotes an $n \times n$ diagonal matrix whose $(i,i)^{th}$ element is the nugget element corresponding to the data source of the i^{th} sample. For instance, suppose the i^{th} sample $(\boldsymbol{u}^i)$ is generated by source ds. Then, $(i,i)^{th}$ element of $\boldsymbol{N}_\delta$ is

δ_{ds}. When training the LMGP, we use Eq. (8) to build the correlation matrix and jointly estimate all the parameters via MAP as:

$$\begin{aligned}\left[\hat{\beta},\hat{\sigma},\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{A}},\widehat{\boldsymbol{\delta}}\right] &= \underset{\beta,\sigma,\boldsymbol{\omega},\boldsymbol{A},\boldsymbol{\delta}}{\operatorname{argmin}}\ L_{MAP} = \\ \underset{\beta,\sigma,\boldsymbol{\omega},\boldsymbol{A},\boldsymbol{\delta}}{\operatorname{argmin}} &\left(\frac{n}{2}\log(\sigma^2)+\frac{1}{2}\log\left(|\boldsymbol{R}_\delta|\right.\right. \\ &+\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{1}_{n\times 1}\beta)^T\boldsymbol{R}_\delta{}^{-1}(\boldsymbol{y} \\ &\left.\left.-\boldsymbol{1}_{n\times 1}\beta)+\log\left(\underset{\beta,\sigma,\boldsymbol{\omega},\boldsymbol{A},\boldsymbol{\delta}}{p(\cdot)}\right)\right)\right.\end{aligned} \tag{9}$$

where $p(\cdot)$ is the prior of the hyperparameters. We define independent priors for each parameter where $\boldsymbol{\omega}^i \sim N(-3,3), \beta \sim N(0,1), \boldsymbol{A}^{ij} \sim N(0,3), \sigma \sim LN(0,3)$[2], and $\boldsymbol{\delta}^i \sim LHS(0,0.01)$[3] [20].

2.4 Multi-source Cost-aware Acquisition Function

The choice of AF is crucial in MFBO since it must consider the biases of LF data and source-dependent sampling costs in addition to balancing exploration and exploitation. To capture these goals, separate AFs are defined in [17] for LF and HF sources with a focus on exploration and exploitation, respectively.

Following the idea of proposing an AF with a focus on exploration for the LF sources, the AF of the j^{th} LF source ($j \neq l$, l denotes the HF source) is defined as the exploration part of expected improvement (EI) in $MFBO_\alpha$:

$$\gamma_{LF}(\boldsymbol{u};j) = \sigma_j(\boldsymbol{u})\phi\left(\frac{y_j^* - \mu_j(\boldsymbol{u})}{\sigma_j(\boldsymbol{u})}\right) \tag{10}$$

where y_j^* is the best function value in the obtained dataset from source j and $\phi(\cdot)$ denotes the probability density function (PDF) of the standard normal variable. $\sigma_j(\boldsymbol{u})$ and $\mu_j(\boldsymbol{u})$ are the standard deviation and mean, respectively, of point $\boldsymbol{u}$ from source j which we estimate via:

$$\begin{aligned}\mu(\boldsymbol{u}) = \\ \hat{\beta} + \boldsymbol{r}^T(\boldsymbol{u})\,\boldsymbol{R}_\delta{}^{-1}(\boldsymbol{y}-\boldsymbol{1}_{n\times 1}\hat{\beta})\end{aligned} \tag{11}$$

$$\begin{aligned}c\big(y(\boldsymbol{u}),y(\boldsymbol{u})\big) &= \sigma^2(\boldsymbol{u}) \\ &= \hat{\sigma}^2\Big(1-\boldsymbol{r}^T(\boldsymbol{u})\boldsymbol{R}_\delta{}^{-1}\boldsymbol{r}(\boldsymbol{u}) \\ &+\big(g(\boldsymbol{u})\big)^2\big(\boldsymbol{1}_{1\times n}\boldsymbol{R}_\delta{}^{-1}\boldsymbol{1}_{n\times 1}\big)^{-1}\Big) \\ &+\hat{\delta}_j\end{aligned} \tag{12}$$

where $g(\boldsymbol{u}) = 1 - \boldsymbol{1}_{1\times n}\boldsymbol{R}_\delta{}^{-1}\boldsymbol{r}(\boldsymbol{u})$ and $\hat{\delta}_j$ is the estimated nugget parameter for source j.

As probability of improvement (PI) is computationally efficient and emphasizes exploitation, $MFBO_\alpha$ utilizes it as the AF for the HF data source. Accordingly, $MFBO_\beta$ uses PI for the HF source (source l) with the new standard deviation and mean calculated based on the Eq. (11) and Eq. (12):

$$\gamma_{HF}(\boldsymbol{u};l) = \psi\left(\frac{\mu_l(\boldsymbol{u}) - y_l^*}{\sigma_l(\boldsymbol{u})}\right) \tag{13}$$

where $\psi(\cdot)$ is the cumulative density function (CDF) of the standard normal distribution.

In each iteration of BO, we first use the mentioned AFs to solve ds auxiliary optimizations to find the candidate points with the highest acquisition value from each source. We then scale these values by the corresponding sampling costs to obtain the following composite AF:

$$\begin{aligned}\gamma_{MFBO_\beta}(\boldsymbol{u};j) = \\ \begin{cases}\dfrac{\gamma_{LF}(\boldsymbol{u};j)}{O(j)} & j=1,\dots,ds \ \text{ and } \ j\neq l \\ \dfrac{\gamma_{HF}(\boldsymbol{u};l)}{O(l)} & j=l\end{cases}\end{aligned} \tag{14}$$

where $O(j)$ is the cost of acquiring one sample from source j. We determine the final candidate point (and the source that it should be sampled from) via:

$$[\boldsymbol{u}^{k+1}, j^{k+1}] = \underset{\boldsymbol{u},j}{\operatorname{argmax}}\ \gamma_{MFBO_\beta}(\boldsymbol{u};j) \tag{15}$$

2.5 Emulation for Exploration

The composite AF in Eq. (14quantifies the information value of LF samples via Eq. (10) whose value scales with the prediction uncertainties, i.e., $\sigma(\boldsymbol{u})$. The source-dependent noise modeling of Section 2.3 improves LMGP's ability in learning the uncertainty by introducing a few more hyper-parameters. However, the added hyperparameters may result into overfitting and, in turn, deteriorate the predicted uncertainties [21, 22].

A related issue is the effect of large *local* biases of LF sources which can inflate the uncertainty quite substantially and, as a result, increase $\gamma_{LF}(\boldsymbol{u};j)$. This increase causes MFBO to repeatedly sample from the biased LF sources (see FIGURE 1 where $MFBO_\alpha$ takes quite a lot of samples from LF1 in the $x < 0$ region while LF1 is quite biased for $x < 0$). Such repeated samplings reduce the efficiency of MFBO and may cause numerical issues or even convergence to a suboptimal solution.

To address the above issues simultaneously, we argue that the training process of the emulator should increase the importance of UQ which directly affects the exploration part of MFBO. To this end, we leverage strictly proper scoring rules while training LMGPs.

Scoring rules are standard methods for evaluating probabilistic predictions [23, 24]. In short, scoring rules evaluate

[2] Log-Normal
[3] Log-Half-Horseshoe with zero lower bound and scale parameter 0.01.

a *probabilistic* prediction by assigning the numerical score c to it. The scoring rule of an emulator is (strictly) proper if matching the predicted distribution with the underlying sample distribution (uniquely) maximizes the expected score for any sample [25].

The probabilistic nature of LMGP's prediction motivates us to use the negatively oriented interval score (hereafter denoted by IS) to evaluate the UQ capabilities of LMGPs. We choose IS since it is robust to outliers, rewards narrow prediction intervals, and is flexible in the choice of desired coverage levels [26, 27].

IS is a special case of quantile prediction that penalizes the model for each observation ($y(\boldsymbol{u}^i)$) that is not inside the $(1-\upsilon)*100\%$ prediction interval. The lower ($\mathcal{L}^i$) and upper ($\mathcal{U}^i$) endpoints of this prediction interval for the i^{th} observation are their predictive quantiles at levels $\frac{\upsilon}{2}$ and $1-\frac{\upsilon}{2}$, respectively. So, we calculate the IS as:

$$IS_\upsilon = \frac{1}{n}\sum_{i=1}^{n}(\mathcal{U}^i-\mathcal{L}^i)+\frac{2}{\upsilon}(\mathcal{L}^i-y(\boldsymbol{u}^i))\,\mathbb{1}\{y(\boldsymbol{u}^i)<\mathcal{L}^i\} + \frac{2}{\upsilon}(y(\boldsymbol{u}^i)-\mathcal{U}^i)\,\mathbb{1}\{y(\boldsymbol{u}^i)>\mathcal{U}^i\} \tag{16}$$

where $\mathbb{1}\{\cdot\}$ is an indicator function which is 1 if its condition holds and zero otherwise [28, 29]. We use $\upsilon = 0.05$ (95% prediction interval), so $\mathcal{U}^i = \mu(\boldsymbol{u}^i) + 1.96\,\sigma(\boldsymbol{u}^i)$ and $\mathcal{L}^i = \mu(\boldsymbol{u}^i) - 1.96\,\sigma(\boldsymbol{u}^i)$.

Having defined the IS, we now formulate the new objective function for training LMGPs where $IS_{0.05}$ is used as a penalty term during hyperparameter estimation to increase the focus on UQ. Since the effectiveness of this penalization mechanism depends on the value of the posterior, we introduce an adaptive coefficient whose magnitude depends on the posterior value. With this penalty term, we estimate the hyperparameters of LMGP via:

$$\left[\hat{\beta},\hat{\sigma},\widehat{\boldsymbol{\omega}},\widehat{\boldsymbol{A}},\widehat{\boldsymbol{\delta}}\right] = \underset{\beta,\sigma,\boldsymbol{\omega},\boldsymbol{A},\boldsymbol{\delta}}{\operatorname{argmin}}\; L_{MAP} + \varepsilon|L_{MAP}|\times IS_{0.05} \tag{17}$$

where $|\cdot|$ denotes the absolute function and ε is a user-defined scaling parameter. In this paper, we use $\varepsilon = 0.08$.

3. RESULTS AND DISCUSSION

We demonstrate the performance of $MFBO_\beta$ on two analytic examples (details in TABLE 1) and two real-world problems. In each case, we compare the results against those of $MFBO_\alpha$ and single-fidelity BO ($SFBO$). While $SFBO$ uses EI as its AF, $MFBO_\beta$ and $MFBO_\alpha$ use the AFs introduced in Section 2.4.

We assume that the cost of querying any of the data sources is much higher than the computational costs of BO (i.e., fitting LMGP and solving the auxiliary optimization problem). Therefore, we compare the methods based on their capability to identify the global optimum of the HF source and the overall data collection cost. By comparing these methods, we aim to demonstrate: (1) the advantages of estimating noise process for each data source, (2) that using IS improves the prediction which also enhances the convergence of BO (our defined AFs highly rely on the quality of the prediction), and (3) that deploying IS eliminates the need for excluding highly biased fidelity sources.

We use the same stop conditions across the three methods to clearly demonstrate the benefits of our two contributions. In particular, the optimization is stopped when either of the following happens: (1) the overall sampling cost exceeds a pre-determined maximum budget, or (2) the best HF sample does not change over 50 iterations. The maximum budget for the analytical examples is 40000 units, while it is 1000 and 1800 for the two real-world examples as their data collection cost is much lower than that of analytical ones.

3.1 Analytical Examples

We consider two analytical examples (*Wing*, *Borehole*) with the dimensionality of 10 and 8, respectively. To challenge the convergence and better illustrate the power of separate noise estimation, we only add noise to the HF data (the noise variance is defined based on the range of each function and is shown in TABLE 1). Both examples are single response and details regarding their formulation, initialization, and sampling cost is presented in TABLE 1. To assess the robustness of the results and quantify the effect of random initial data, we repeat the optimization process 20 times for each example with each of the three methods (all initial data are generated via Sobol sequence).

In each example, the relative root mean squared error (RRMSE) is calculated between LF sources and their corresponding HF source based on 1000 samples to show the relative accuracy of the LF sources (presented in TABLE 1). Based on these numbers, in *Borehole,* unlike *Wing*, the source ID, true fidelity level (based on the RRMSEs), and sampling cost are not related. For instance, although the first LF source is the most expensive one, it has the least accuracy.

$MFBO_\alpha$ excludes the highly biased LF sources from BO before any new samples are obtained. This exclusion is done based on the latent map of the LMGP model that is trained on the *initial* data. FIGURE 2 shows the latent maps of *Wing* and *Borehole* examples where Source 1 represents the HF source, and the rest of the Sources represent LF ones. As shown in FIGURE 2, while all the fidelity sources of *Wing* are beneficial (since the points encoding the LF sources are very close to the HF point), the first two LF sources of *Borehole* are not correlated enough with the HF (their latent positions are distant from that of the HF) and hence are excluded in $MFBO_\alpha$. However, $MFBO_\beta$ does not require this exclusion because it leverages the biased LF sources merely in the regions that they are correlated with the HF source. In this paper, we do not exclude the biased sources in $MFBO_\alpha$ to better compare it with our proposed method.

FIGURE 3 summarizes the convergence history of each example by depicting the best HF sample found by each method versus its accumulated sampling cost. As we expect, MF methods ($MFBO_\beta$ and $MFBO_\alpha$) outperform *SFBO* in *Wing* (FIGURE 3a) by leveraging the inexpensive LF sources that are

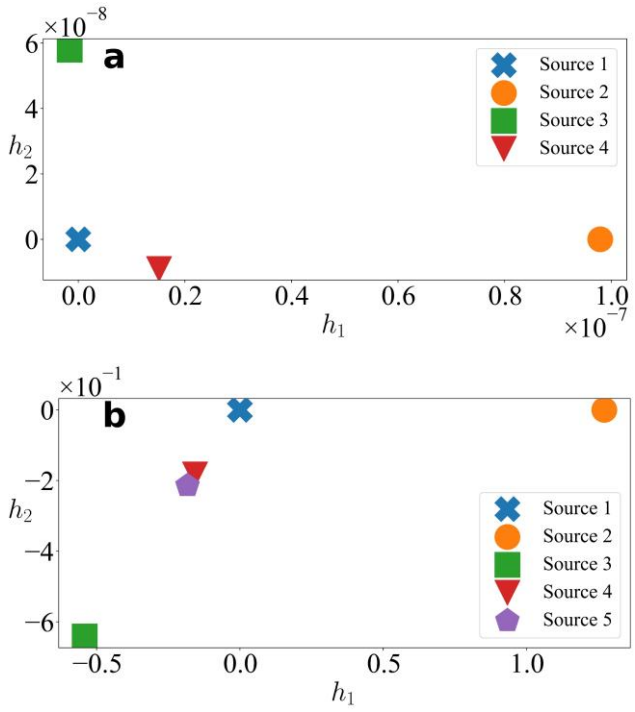

FIGURE 2 FIDELITY MANIFOLDS OF ANALYTIC EXAMPLES: The plots in **(a)** and **(b)** are obtained by fitting an LMGP to the initial data in the *Wing* and *Borehole* examples, respectively. Due to the consistency across the 20 repetitions, the plots are randomly chosen among them. In **(b)**, the HF source (Source 1) is encoded far from LF1 and LF2 which indicates that these two sources have large biases with respect to the HF source. $MFBO_\alpha$ excludes these two sources from the BO while $MFBO_\beta$ does not.

globally correlated with the HF, leading them to better convergence performance with lower cost. However, the superior performance of $MFBO_\beta$ is more obvious in *Borehole* with biased data sources.

In *Borehole* (FIGURE 3b), all the thin red curves ($MFBO_\alpha$) are straight lines, except for two curves. This means that for 18 repetitions, the optimization process fails to improve. The reason behind this lack of improvement is that $MFBO_\alpha$ fails to handle *local* correlation of the LF sources, and samples unvaluable points that steer the optimization to the wrong direction. Consequently, $MFBO_\alpha$ cannot find any efficient HF sample with large enough information value to compensate for its high sampling cost which results in the lack of improvement. Conversely, all the thin green curves ($MFBO_\beta$) converge to a value very close to the ground truth. In addition, $MFBO_\beta$ yields almost the same convergence value as *SFBO*, but with lower computational cost. This instance further demonstrates the effectiveness of our proposed AFs, since *SFBO* is very accurate due to only sampling from HF and not dealing with local biases.

FIGURE 4 illustrates the details of the accumulated convergence cost and values of each repetition. The pink boxes show the variations in convergence values through 20 repetitions and the blue ones demonstrate the convergence cost. The dashed pink line is the ground truth we aim to find. As also shown in FIGURE 3, in all the examples, $MFBO_\beta$ outperforms other methods considering convergence value and cost. As demonstrated in FIGURE 4a, unlike MF methods, the increase in dimensionality (8 in *Borehole* to 10 in *Wing*) adversely affects the performance of *SFBO*, as it cannot find the ground truth of *Wing* despite mere sampling from HF source. In addition, utilizing beneficial information provided by the inexpensive unbiased LF sources, causes the better performance of $MFBO_\beta$ and $MFBO_\alpha$ compared to *SFBO*. The lower variations in the convergence values of $MFBO_\beta$ through 20 repetitions (smaller pink boxes and whiskers) and lower convergence costs compared to $MFBO_\alpha$, further show the superiority of our proposed method; $MFBO_\beta$ outperforms $MFBO_\alpha$ even in the absence of biased sources.

The rationale behind the lower cost of $MFBO_\alpha$ in *Borehole* (FIGURE 4b) can be attributed to the observation made for FIGURE 3b; in 18 repetitions of *Borehole*, the optimization is not improved at all. So, in all those repetitions, the optimization meets the second stop condition and is terminated in the 50th iteration while it is not converged. This fact is also obvious in the long whiskers of the pink boxplots of $MFBO_\alpha$ (high variation on

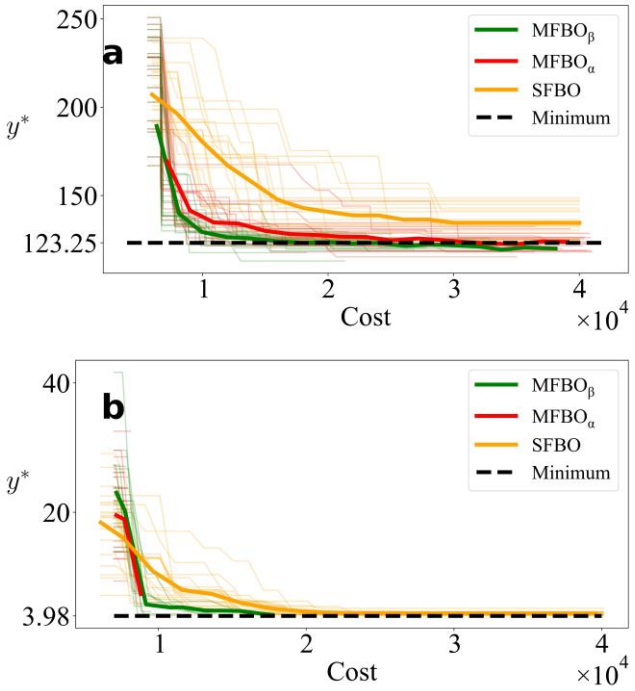

FIGURE 3 CONVERGENCE HISTORIES: The plots depict the best HF sample found by each approach versus their sampling costs accumulated during the BO iterations (the cost of initial data is included). **(a)** and **(b)** refer to *Wing* and *Borehole*, respectively. The thin curves show the convergence history of each repetition, and the solid thick ones indicate the average behavior across the 20 repetitions. In all the examples, $MFBO_\beta$ outperforms $MFBO_\alpha$ in terms of both convergence value and cost. In both examples, *SFBO* performs the worst. The ground truth is represented by the black dashed line.

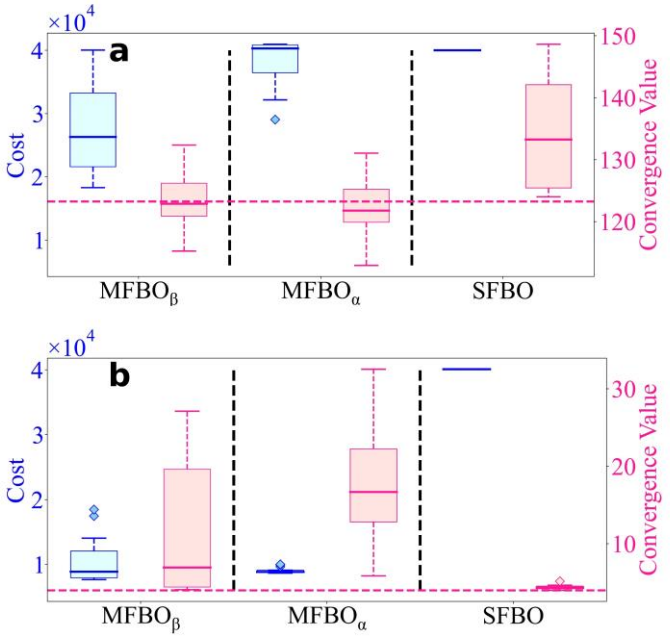

FIGURE 4 ACCUMULATED COSTS AND CONVERGENCE VALUES: The blue boxplots illustrate the accumulated costs of each repetition, and the pink ones demonstrate their convergence values for the three methods. The pink dashed line indicates the ground truth. **(a)** and **(b)** refer to *Wing* and *Borehole*, respectively. In both examples, the median of the convergence box of $MFBO_\beta$ is very closer to the ground truth than that of the $MFBO_\alpha$, while its cost is much lower which proves the superiority of our proposed approach.

the convergence values in each repetition) and the median which is far from the ground truth.

3.2 Real-world Datasets

In this section, we study two materials design problems where the aim is to find the composition that best optimizes the property of interest. We do not add noise to these two examples as they are noisy themselves. Both examples have categorical inputs and the HF and LF data are obtained via simulations (based on the density functional theory) with different fidelity levels.

The first problem is bi-fidelity where the goal is to find the member of the nanolaminate ternary alloy (*NTA*) family with the largest bulk modulus [30]. The HF and LF datasets are 10-dimensional (7 quantitative and 3 categorical where the latter have 10, 12, and 2 levels), single response with 224 samples each. We define the cost ratio of 10/1 for the fidelity sources and initialize the BO with 20 HF and 10 LF samples (the composition with the largest bulk modulus is excluded from the HF dataset). To quantify the robustness of the proposed method to the random initial data, we repeat this process 20 times for each BO method.

The second problem is hybrid organic–inorganic perovskite (*HOIP*) crystals that aim to find the compound with the smallest inter-molecular binding energy [31]. There are 3 fidelity sources for this example, 1 HF and 2 LFs, with the same dimensionality (1 output and 3 categorical inputs with 10, 3, and 16 levels) but different sizes. The HF dataset has 480 samples, while the first and second LF sources have 179 and 240 samples, respectively. We assign the cost ratio of 15/10/5 to the fidelity sources and initialize the BO with (15, 20, 15) samples for the HF and LF sources, respectively (the best compound is excluded). We repeat the BO process 20 times to assess the robustness.

As mentioned before, the first step in $MFBO_\alpha$ is to train an LMGP to the initial data in each problem to exclude the highly biased sources. As illustrated in FIGURE 5, the latent points of the fidelity sources of *NTA* are very close in the learned fidelity manifold which demonstrates a high correlation between its fidelity sources. However, both latent points of LF sources in *HOIP* are far from the HF one (Source 1), so they both should be excluded. By excluding both LF sources, the MF problem converts to the SF one, so $MFBO_\alpha$ is not applicable to it anymore. Therefore, we do not exclude the biased LF sources from *HOIP* in this paper to be able to compare the performance of $MFBO_\beta$ with $MFBO_\alpha$.

A summary of the convergence history of *NTA* and *HOIP* is depicted in FIGURE 6 by showing the best HF sample found by each method versus its accumulated sampling cost. In *NTA* (FIGURE 6a), the LF sources are globally correlated with the HF; consequently, MF methods perform better than *SFBO* by using inexpensive and informative LF sources. Additionally,

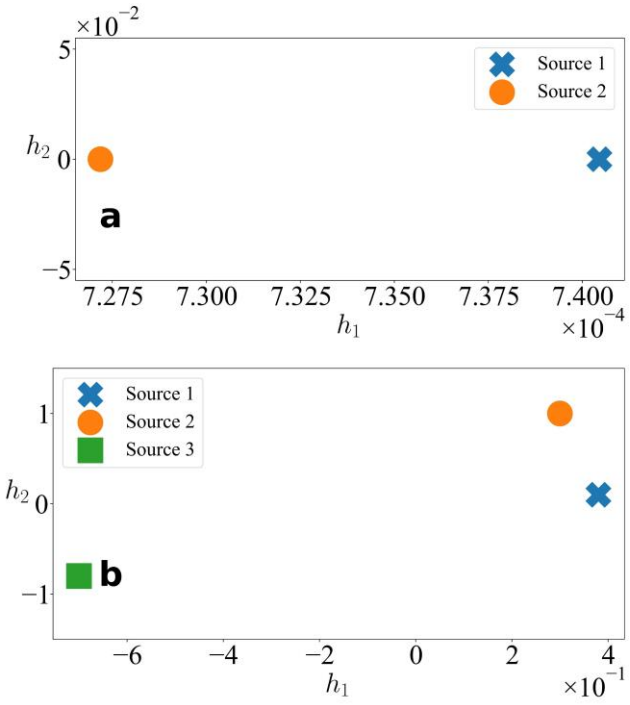

FIGURE 5 FIDELITY MANIFOLDS OF REAL-WORLD EXAMPLES: The plots in **(a)** and **(b)** are obtained by fitting an LMGP to the initial data in the *NTA* and *HOIP* examples, respectively. Due to the consistency across the 20 repetitions, the plots are randomly chosen among them. In **(b)**, $MFBO_\alpha$ excludes LF1 and LF2 as they are encoded far from the HF (Source 1).

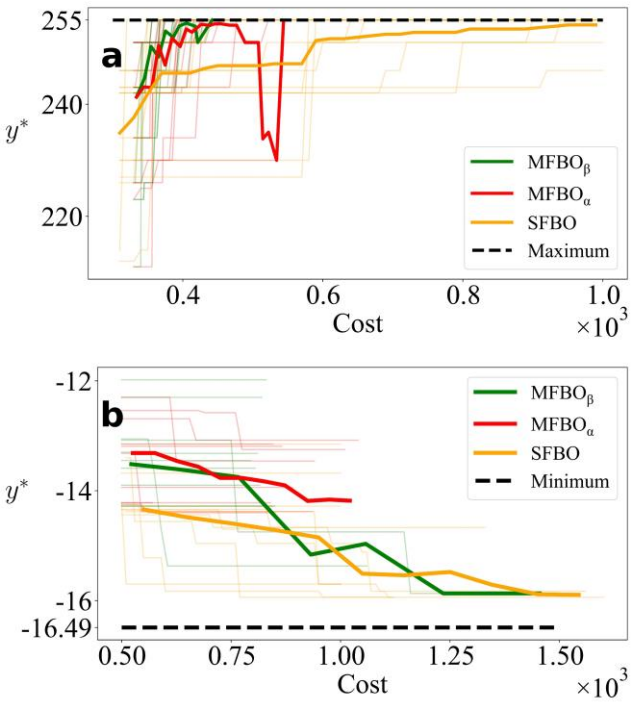

FIGURE 6 CONVERGENCE HISTORIES: The plots depict the best HF sample found by each approach versus their sampling costs accumulated during the BO iterations (the cost of initial data is included). **(a)** and **(b)** refer to *NTA* and *HOIP*, respectively. The solid thick curves indicate the average behavior across the 20 repetitions. $MFBO_\beta$ performs better than $MFBO_\alpha$ in both examples in case of convergence value and cost.

higher prediction accuracy of the emulator of $MFBO_\beta$ results in a more efficient sampling and faster convergence of BO in $MFBO_\beta$ compared to $MFBO_\alpha$.

The superiority of $MFBO_\beta$ is more obvious in *HOIP* (FIGURE 6b) with biased sources. In *HOIP* example, as we expect, $MFBO_\alpha$ converges to a sub-optimal compound since both LF sources are only locally correlated with the HF source. So, the AFs fail to sample valuable points to improve the optimization as they cannot find the region where the LF sources are beneficial and informative. Additionally, each data source is obtained from a distinct process, so it suffers from different types and levels of noise. Therefore, estimating a single noise for all the data sources in $MFBO_\alpha$, results in a poor emulation and further exacerbates the performance of AFs. $MFBO_\beta$ overcomes these issues by focusing more on UQ and estimating separate noise processes, resulting in better performance, and outperforming $MFBO_\alpha$.

Similar to Section 3.1, we also show the details of the accumulated convergence cost and values of each repetition in FIGURE 7. In this regard, convergence of all methods to the ground truth in all the repetitions in *NTA* (FIGURE 7a), makes the pink boxplots become straight lines (zero variation). However, the two outliers in $SFBO$ show its disability in finding the ground truth in two repetitions despite mere sampling from the accurate HF source. Additionally, the faster convergence of $MFBO_\beta$ compared to other methods which roots from efficient sampling, causes much lower convergence costs.

In *HOIP* (FIGURE 7b), as shown in FIGURE 6b, $MFBO_\alpha$ fails to manage the noise and biased sources, and converges to a sub-optimal compound. However, $MFBO_\beta$ leverages the biased LF sources only in the regions that they provide effective information by taking advantage of the penalized objective function which significantly improves the prediction. Furthermore, the large variation in the convergence values of $SFBO$ and its much higher cost, further highlights the superiority of $MFBO_\beta$ in efficiency and robustness, though the median of the convergence boxplot of $SFBO$ is closer to the ground truth.

4. CONCLUSION

In this paper, we develop a novel method to improve the performance of multi-fidelity cost-aware BO techniques. Our method enhances the accuracy and convergence rate of MFBO through two main contributions. Firstly, we enable the emulator to estimate separate noise processes for each source of data. This feature increases the accuracy of the trained model since different data sources may exhibit different types and levels of noise. Secondly, we define a new objective function penalized by IS to (1) further improve the accuracy of the prediction, (2) increase the focus on UQ, and (3) forgo the need to exclude biased data sources. The main advantages of our method are its efficient and superior performance in the presence of highly biased data sources, and no required prior knowledge about the

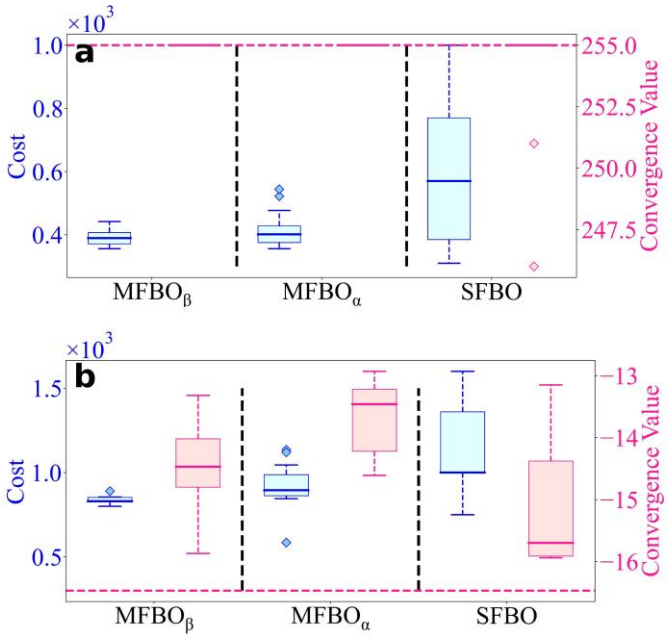

FIGURE 7 ACCUMULATED COSTS AND CONVERGENCE VALUES: The blue boxplots illustrate the accumulated costs of each repetition, and the pink ones demonstrate their convergence values. The pink dashed line indicates the ground truth. **(a)** and **(b)** refer to *NTA* and *HOIP*, respectively. **(a)** All the methods found the ground truth in all iterations, so there is no variation.

accuracy of the data sources. We illustrate these advantages through analytic and real-world examples.

We propose to use two fixed AFs in each iteration. However, one can also customize the choice of AFs for different iterations using adaptive approaches. The examples presented in this paper are limited to single-objective problems. In addition, in all the examples we report the noisy data as the result, but one can increase the accuracy by excluding the noise effects on the convergence value. We intend to explore these avenues further in future research, aiming to extend our proposed method to handle multi-objective problems with adaptive AFs.

ACKNOWLEDGEMENTS

We appreciate the support from National Science Foundation (award numbers OAC-2211908 and OAC-2103708) and the Early Career Faculty grant from NASA's Space Technology Research Grants Program (award number 80NSSC21K1809).

APPENDIX

A Table of Numerical

TABLE 1 LIST OF ANALYTIC FUNCTIONS:n denotes the number of initial samples. The relative root mean squared error (RRMSE) of an LF source is calculated by comparing its output to that of the HF source at 10000 random points. The cost column is the cost of obtaining a sample from the corresponding source. The last column denotes the variance of the added noise to the HF source.

Name	Source ID	Formulation	n	RRMSE	Cost	Noise variance
Borehole	HF	$\frac{2\pi T_u (H_u - H_l)}{\ln\left(\frac{r}{r_w}\right)\left(1+\frac{2LT_u}{\ln\left(\frac{r}{r_w}\right) r_w^2 k_w}+\frac{T_u}{T_l}\right)}$	5	-	1000	16
	LF1	$\frac{2\pi T_u (H_u - 0.8 H_l)}{\ln\left(\frac{r}{r_w}\right)\left(1+\frac{1LT_u}{\ln\left(\frac{r}{r_w}\right) r_w^2 k_w}+\frac{T_u}{T_l}\right)}$	5	4.40	100	-
	LF2	$\frac{2\pi T_u (H_u + 3H_l)}{\ln\left(\frac{r}{r_w}\right)\left(1+\frac{8LT_u}{\ln\left(\frac{r}{r_w}\right) r_w^2 k_w}+0.75\frac{T_u}{T_l}\right)}$	50	1.54	10	-
	LF3	$\frac{2\pi T_u (1.1 H_u - H_l)}{\ln\left(\frac{4r}{r_w}\right)\left(1+\frac{3LT_u}{\ln\left(\frac{r}{r_w}\right) r_w^2 k_w}+\frac{T_u}{T_l}\right)}$	5	1.3	100	-
	LF4	$\frac{2\pi T_u (1.05 H_u - H_l)}{\ln\left(\frac{2r}{r_w}\right)\left(1+\frac{2LT_u}{\ln\left(\frac{r}{r_w}\right) r_w^2 k_w}+\frac{T_u}{T_l}\right)}$	50	1.3	10	-
Wing	HF	$0.36 s_w^{0.758} w_{fw}^{0.0035}\left(\frac{A}{cos^2(\Lambda)}\right)^{0.6} q^{0.006}\lambda^{0.04}\left(\frac{100t_c}{\cos(\Lambda)}\right)^{-0.3}(N_z W_{dg})^{0.49} + s_w w_p$	5	-	1000	9
	LF1	$0.36 s_w^{0.758} w_{fw}^{0.0035}\left(\frac{A}{cos^2(\Lambda)}\right)^{0.6} q^{0.006}\lambda^{0.04}\left(\frac{100t_c}{\cos(\Lambda)}\right)^{-0.3}(N_z W_{dg})^{0.49} + w_p$	5	0.19	100	-
	LF2	$0.36 s_w^{0.8} w_{fw}^{0.0035}\left(\frac{A}{cos^2(\Lambda)}\right)^{0.6} q^{0.006}\lambda^{0.04}\left(\frac{100t_c}{\cos(\Lambda)}\right)^{-0.3}(N_z W_{dg})^{0.49} + w_p$	10	1.14	10	-
	LF3	$0.36 s_w^{0.9} w_{fw}^{0.0035}\left(\frac{A}{cos^2(\Lambda)}\right)^{0.6} q^{0.006}\lambda^{0.04}\left(\frac{100t_c}{\cos(\Lambda)}\right)^{-0.3}(N_z W_{dg})^{0.49}$	50	5.75	1	-

References

1. Shahriari, B., et al., *Taking the human out of the loop: A review of Bayesian optimization.* Proceedings of the IEEE, 2015. **104**(1): p. 148-175.
2. Frazier, P.I.J.a.p.a., *A tutorial on Bayesian optimization.* 2018.
3. Li, S. *Multi-Fidelity Bayesian Optimization via Deep Neural Networks*. in *Neural Information Processing Systems*. 2020. Vancouver.
4. Song, J., Y. Chen, and Y. Yue. *A general framework for multi-fidelity bayesian optimization with gaussian processes*. in *The 22nd International Conference on Artificial Intelligence and Statistics*. 2019. PMLR.
5. Takeno, S., et al. *Multi-fidelity Bayesian optimization with max-value entropy search and its parallelization*. in *International Conference on Machine Learning*. 2020. PMLR.
6. Zhang, S., et al. *An efficient multi-fidelity bayesian optimization approach for analog circuit synthesis*. in *Proceedings of the 56th Annual Design Automation Conference 2019*. 2019.
7. Shu, L., P. Jiang, and Y. Wang, *A multi-fidelity Bayesian optimization approach based on the expected further improvement.* Structural and Multidisciplinary Optimization, 2021. **63**: p. 1709-1719.
8. Tran, A., T. Wildey, and S. McCann, *sMF-BO-2CoGP: A sequential multi-fidelity constrained Bayesian optimization framework for design applications.* Journal of Computing and Information Science in Engineering, 2020. **20**(3).
9. Xiao, M., et al., *Extended Co-Kriging interpolation method based on multi-fidelity data.* 2018. **323**: p. 120-131.
10. Kennedy, *Bayesian calibration of computer models.* Journal of the Royal Statistical Society: Series B (Statistical Methodology),, 2001.
11. Gardner, *Blackbox matrix-matrix gaussian process inference with gpu acceleration.* Advances in neural information processing systems, 2018.
12. Foumani, Z.Z., et al., *Multi-fidelity cost-aware Bayesian optimization.* 2023. **407**: p. 115937.
13. Escamilla-Ambrosio, P.J. and N. Mort. *Hybrid Kalman filter-fuzzy logic adaptive multisensor data fusion architectures*. in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*. 2003. IEEE.
14. Kreibich, O., J. Neuzil, and R.J.I.T.o.I.E. Smid, *Quality-based multiple-sensor fusion in an industrial wireless sensor network for MCM.* 2013. **61**(9): p. 4903-4911.
15. Eweis-Labolle, J.T., N. Oune, and R.J.J.o.M.D. Bostanabad, *Data Fusion With Latent Map Gaussian Processes.* 2022. **144**(9): p. 091703.
16. Oune, N., *Latent map Gaussian processes for mixed variable metamodeling.* Computer Methods in Applied Mechanics and Engineering 2021.
17. Zanjani Foumani, Z., *Multi-Fidelity Cost-Aware Bayesian Optimization.* arXiv preprint arXiv:2211.02732, 2022.
18. Oune, N., R.J.C.M.i.A.M. Bostanabad, and Engineering, *Latent map Gaussian processes for mixed variable metamodeling.* 2021. **387**: p. 114128.
19. Bostanabad, R., et al., *Leveraging the nugget parameter for efficient Gaussian process modeling.* International journal for numerical methods in engineering, 2018. **114**(5): p. 501-516.
20. Carvalho, C.M., N.G. Polson, and J.G.J.B. Scott, *The horseshoe estimator for sparse signals.* 2010. **97**(2): p. 465-480.
21. Gal, Y., M. Van Der Wilk, and C.E.J.A.i.n.i.p.s. Rasmussen, *Distributed variational inference in sparse Gaussian process regression and latent variable models.* 2014. **27**.
22. Mohammed, R.O. and G.C. Cawley. *Over-fitting in model selection with Gaussian process regression*. in *Machine Learning and Data Mining in Pattern Recognition: 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings 13*. 2017. Springer.
23. Lindley, D.V.J.I.S.R.R.I.d.S., *Scoring rules and the inevitability of probability.* 1982: p. 1-11.
24. Winkler, R.L., et al., *Scoring rules and the evaluation of probabilities.* 1996. **5**: p. 1-60.
25. Gneiting, T. and A.E.J.J.o.t.A.s.A. Raftery, *Strictly proper scoring rules, prediction, and estimation.* 2007. **102**(477): p. 359-378.
26. Bracher, J., et al., *Evaluating epidemic forecasts in an interval format.* PLoS computational biology, 2021. **17**(2): p. e1008618.
27. Mitchell, K. and C. Ferro, *Proper scoring rules for interval probabilistic forecasts.* Quarterly Journal of the Royal Meteorological Society, 2017. **143**(704): p. 1597-1607.
28. Gneiting, T. and A.E. Raftery, *Strictly proper scoring rules, prediction, and estimation.* Journal of the American statistical Association, 2007. **102**(477): p. 359-378.
29. Mora, C., et al., *Probabilistic Neural Data Fusion for Learning from an Arbitrary Number of Multi-fidelity Data Sets.* arXiv preprint arXiv:2301.13271, 2023.
30. Yeom, C.-U., *Performance evaluation of automobile fuel consumption us-*. Symmetry, 2019.
31. Herbol, H.C., et al., *Efficient search of compositional space for hybrid organic–inorganic perovskites via Bayesian optimization.* npj Computational Materials, 2018. **4**(1): p. 51.