Multifidelity Data Fusion via Gradient-Enhanced Gaussian Process Regression

Yixiang Deng[1,2], Guang Lin[3,4], and Xiu Yang*[5]

[1]School of Engineering, Brown University, USA
[2]Division of Applied Mathematics, Brown University, USA
[3]Department of Mathematics, Purdue University, USA
[4]School of Engineering, Purdue University, USA
[5]Department of Industrial and Systems Engineering, Lehigh University, USA

August 4, 2020

Abstract

We propose a data fusion method based on multi-fidelity Gaussian process regression (GPR) framework. This method combines available data of the quantity of interest (QoI) and its gradients with different fidelity levels, namely, it is a Gradient-enhanced Cokriging method (GE-Cokriging). It provides the approximations of both the QoI and its gradients *simultaneously* with uncertainty estimates. We compare this method with the conventional multi-fidelity Cokriging method that does not use gradients information, and the result suggests that GE-Cokriging has a better performance in predicting both QoI and its gradients. Moreover, GE-Cokriging even shows better generalization result in some cases where Cokriging performs poorly due to the singularity of the covariance matrix. We demonstrate the application of GE-Cokriging in several practical cases including reconstructing the trajectories and velocity of an underdamped oscillator with respect to time simultaneously, and investigating the sensitivity of power factor of a load bus with respect to varying power inputs of a generator bus in a large scale power system. We also show that though GE-Cokriging method requires a little bit higher computational cost than Cokriging method, the result of accuracy comparison shows that this cost is usually worth it.

1 Introduction

Gaussian process (GP) is one of the most well studied stochastic processes in probability and statistics. Given the flexible form of data representation, GP is a powerful tool for classification and regression, and it is widely used in probabilistic scientific computing, engineering design, geostatistics, data assimilation, machine learning, etc. In particular, given a data set comprising input/output pairs of locations and quantity of interest (QoI), *GP regression* (GPR, also known as *Kriging*), can provide a prediction along with a mean squared error (MSE) estimate of the QoI at any location. Alternatively, from the Bayesian perspective, GPR identifies a Gaussian random variable at any location with a posterior mean (corresponding to the prediction) and variance (corresponding to the MSE). Generally speaking, the larger the given data set size is, the closer the GPR's posterior mean is to the ground truth and the smaller the posterior variance is.

In many practical problems, obtaining a large amount of data can be difficult because of the limitation of resources. There are several approaches to augment the data set in different manners. For example, the original Cokriging method exploits the correlation between multiple QoIs in the geostatistical study,

*xiy518@lehigh.edu

e.g., the correlation between temperature and precipitation [10, 29, 28], or that between near-surface soil density and the gravity-gradient [8], to improve the accuracy of prediction. Later, the Cokriging method was extended to utilizing correlation between the same QoI from models with different fidelities [13, 9, 11, 22]. This GP-based multi-fidelity method is very useful in scientific computing, because low-fidelity models, e.g., coarse-grained molecular dynamics [5, 27], Reynolds-average Navier-Stokes equations [25, 2], numerical simulations on coarse grids, are often used with high-fidelity models, e.g., molecular dynamics, full Navier-Stokes equations, numerical simulations on fine grids [18], in optimization, uncertainty quantification (UQ), control [21], variable-fidelity quantum mechanical calculations of bandgaps of solids [24], etc. In these tasks, the multi-fidelity method leverages low-fidelity models for speedup, while uses a high-fidelity model to establish accuracy and/or convergence guarantees. Moreover, the empirical statistics of simulation results from stochastic scientific computing models can be used to construct single- or multi-fidelity GP models [34, 33, 35]. In this work, Cokriging refers to the GP-based multi-fidelity approach.

Another important approach to enlarge the data set is to use gradient information of the QoI. This approach can be categorized as Cokriging because the QoI and its gradients are variables of different species. The idea of incorporating derivatives or gradients to optimize Bayesian prediction was proposed by Morris et al. [20]. The *gradient-enhanced Kriging* (GE-Kriging) method, also referred to as Gradient-based Kriging in some literature, has been widely investigated in areas such as computational fluid dynamics, especially in aerodynamics optimization problems [4, 32, 15, 3]. Incorporating gradient information in different ways, this method consists of direct and indirect approaches. The former uses the gradient information through an augmented covariance matrix [12], while the latter approximates the gradient via finite-difference method [3, 37]. The *gradient-enhanced Cokriging* (GE-Cokriging) method in [16] refers to a GE-Kriging method that uses a different covariance function between the QoI and its gradients other than that in conventional GE-Kriging. The GE-Cokriging method in [30] combines multi-fidelity information of the QoI and its gradients to predict the QoI only.

Most of the aforementioned works focus on enhancing the accuracy of predicting the QoI. Hence, when the gradient information is used, the method is a "gradient-enhanced" approach. However, in many applications, both the QoI and its gradient are important. For example, when studying the phase diagram of a dynamical system, one needs an accurate prediction of both location and velocity. Another example is the sensitivity analysis of a system, where the gradient information is critical. Therefore, in this work, we propose a comprehensive *multifidelity gradient-enhanced Cokriging* method to predict both QoI and its gradients *simultaneously* based on GE-Cokriging [30]. This method exploits the QoI and its gradient from models of different fidelities based on the combination of the GE-Kriging and the Cokriging to improve the prediction accuracy. In terms of predicting the QoI, this method can be considered as "gradient-enhanced", while from the perspective of estimating gradients, this method can be considered as "integral-enhanced". In this work, GE-Cokriging refers to our proposed multi-fidelity method, instead of the GE-Cokriging in [16].

In this paper, we firstly review GPR (Kriging) and its extension for a multi-fidelity study (Cokriging). Then, we describe the gradient-enhanced Kriging/Cokriging as well as the GE-Kriging/Cokriging method. Finally, we use four examples to demonstrate the efficacy of our approach.

2 Methodology

2.1 GPR framework

We present a brief review of the GPR method adopted from [1, 6, 34]. We denote the observation locations as $\boldsymbol{X} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ ($\boldsymbol{x}^{(i)} \in D, D \subseteq \mathbb{R}^d$) and the observed values of the QoI at these locations as $\boldsymbol{y} = (y^{(1)}, y^{(2)}, \ldots, y^{(N)})^\top$ ($y^{(i)} \in \mathbb{R}$). For simplicity, we assume that $y^{(i)}$ are scalars. The GPR method aims to identify a GP $Y(\boldsymbol{x}, \omega) : D \times \Omega \to \mathbb{R}$ based on the input/output data set $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, where Ω is the sample space of a probability triple. Here, $\boldsymbol{x}$ can be considered as parameters for this GP, such that $Y(\boldsymbol{x}, \cdot) : \Omega \to \mathbb{R}$ is a Gaussian random variable for any $\boldsymbol{x}$ in the set D. A GP $Y(\boldsymbol{x}, \omega)$ is usually denoted as

$$Y(\boldsymbol{x}) \sim \mathcal{GP}\left(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\right), \tag{2.1}$$

where ω is not explicitly listed for brevity, $\mu(\cdot) : D \to \mathbb{R}$ and $k(\cdot,\cdot) : D \times D \to \mathbb{R}$ are the mean and covariance functions (also called *kernel* function), respectively:

$$\mu(\boldsymbol{x}) = \mathrm{E}\left\{Y(\boldsymbol{x})\right\}, \tag{2.2}$$

$$k(\boldsymbol{x},\boldsymbol{x}') = \mathrm{Cov}\left\{Y(\boldsymbol{x}), Y(\boldsymbol{x}')\right\} = \mathrm{E}\left\{(Y(\boldsymbol{x}) - \mu(\boldsymbol{x}))(Y(\boldsymbol{x}') - \mu(\boldsymbol{x}'))\right\}. \tag{2.3}$$

The variance of $Y(\boldsymbol{x})$ is $k(\boldsymbol{x},\boldsymbol{x})$, and its standard deviation is $\sigma(\boldsymbol{x}) = \sqrt{k(\boldsymbol{x},\boldsymbol{x})}$. The covariance matrix, denoted as $\boldsymbol{C}$, is defined as $C_{ij} = k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$. Functions $\mu(\boldsymbol{x})$ and $k(\boldsymbol{x},\boldsymbol{x}')$ are obtained by identifying their hyperparameters via maximizing the log marginal likelihood [31]:

$$\ln L = -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\top \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) - \frac{1}{2}\ln|\boldsymbol{C}| - \frac{N}{2}\ln 2\pi, \tag{2.4}$$

where $\boldsymbol{\mu} = (\mu(\boldsymbol{x}^{(1)}), \ldots, \mu(\boldsymbol{x}^{(N)}))^\top$ and $|\boldsymbol{C}|$ is the determinant of matrix $\boldsymbol{C}$. For any $\boldsymbol{x}^* \in D$, the GPR posterior mean and variance are

$$\hat{y}(\boldsymbol{x}^*) = \mu(\boldsymbol{x}^*) + \boldsymbol{c}(\boldsymbol{x}^*)^\top \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}), \tag{2.5}$$

$$\hat{s}^2(\boldsymbol{x}^*) = \sigma^2(\boldsymbol{x}^*) - \boldsymbol{c}(\boldsymbol{x}^*)^\top \boldsymbol{C}^{-1}\boldsymbol{c}(\boldsymbol{x}^*), \tag{2.6}$$

where $\boldsymbol{c}(\boldsymbol{x}^*)$ is a vector of covariance: $(\boldsymbol{c}(\boldsymbol{x}^*))_i = k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^*)$. In practice, it is common to use $\hat{y}(\boldsymbol{x}^*)$ as the prediction, and $\hat{s}^2(\boldsymbol{x}^*)$ is also called the mean squared error (MSE) of the prediction because $\hat{s}^2(\boldsymbol{x}^*) = \mathrm{E}\left\{(\hat{y}(\boldsymbol{x}^*) - Y(\boldsymbol{x}^*))^2\right\}$ [6]. Consequently, $\hat{s}(\boldsymbol{x}^*)$, the posterior standard deviation, is called the root mean squared error (RMSE). Moreover, to account for the observation noise, one can assume that the noise is independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and variance δ^2, and replace $\boldsymbol{C}$ with $\boldsymbol{C} + \delta^2\boldsymbol{I}$. In this study, we assume that observations $\boldsymbol{y}$ are noiseless. If $\boldsymbol{C}$ is not invertible or its condition number is very large, one can add a small regularization term $\alpha\boldsymbol{I}$ (α is a small positive real number) to $\boldsymbol{C}$, which is equivalent to assuming there is an observation noise. In addition, $\hat{s}$ can be used in global optimization, or in the greedy algorithm to identify locations of additional observations.

2.2 Kriging and Cokriging with stationary kernel

In the widely used ordinary Kriging method, a stationary GP is assumed [14]. Specifically, μ is set as a constant $\mu(\boldsymbol{x}) \equiv \mu$, and $k(\boldsymbol{x},\boldsymbol{x}') = k(\boldsymbol{\tau})$, where $\boldsymbol{\tau} = \boldsymbol{x} - \boldsymbol{x}'$. Consequently, $\sigma^2(\boldsymbol{x}) = k(\boldsymbol{x},\boldsymbol{x}) = k(\boldsymbol{0}) = \sigma^2$ is a constant. The most widely used kernels in scientific computing is the Matérn functions, especially its two special cases, i.e., exponential and squared-exponential (Gaussian) kernels. For example, the Gaussian kernel can be written as $k(\boldsymbol{\tau}) = \sigma^2 \exp\left(-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}'\|_w^2\right)$, where the weighted norm is defined as $\|\boldsymbol{x}-\boldsymbol{x}'\|_w^2 = \sum_{i=1}^{d}\left(\frac{x_i - x_i'}{l_i}\right)^2$. Here, l_i ($i = 1, \ldots, d$), the correlation lengths in the i direction, are constants. Given a stationary covariance function, the covariance matrix $\boldsymbol{C}$ can be written as $\boldsymbol{C} = \sigma^2\boldsymbol{\Psi}$, where $\Psi_{ij} = \exp(-\frac{1}{2}\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\|_w^2)$. The estimators of μ and σ^2, denoted as $\hat{\mu}$ and $\hat{\sigma}^2$, are

$$\hat{\mu} = \frac{\boldsymbol{1}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{y}}{\boldsymbol{1}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{1}}, \qquad \hat{\sigma}^2 = \frac{(\boldsymbol{y}-\boldsymbol{1}\hat{\mu})^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{1}\hat{\mu})}{N}, \tag{2.7}$$

where $\boldsymbol{1}$ is a constant vector consisting of 1s [6]. It is also common to set $\mu = 0$ [31]. The hyperparameters σ and l_i are identified by maximizing the log marginal likelihood in Eq. (2.4). The terms $\hat{y}(\boldsymbol{x}^*)$ and $\hat{s}^2(\boldsymbol{x}^*)$ in Eq. (2.5) take the following form:

$$\hat{y}(\boldsymbol{x}^*) = \hat{\mu} + \boldsymbol{\psi}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{y}-\boldsymbol{1}\hat{\mu}), \tag{2.8}$$

$$\hat{s}^2(\boldsymbol{x}^*) = \hat{\sigma}^2\left(1 - \boldsymbol{\psi}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\psi}\right), \tag{2.9}$$

where $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{x}^*)$ is a (column) vector consisting of correlations between the observed data and the prediction, i.e., $\psi_i = \frac{1}{\sigma^2}k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^*)$.

Next, we briefly review the formulation of the multifidelity Cokriging, and we use the two-fidelity model for demonstration. Suppose that we have high-fidelity data (e.g., accurate measurements of the QoI) $\boldsymbol{y}_H = (y_H^{(1)}, \ldots, y_H^{(N_H)})^\top$ at locations $\boldsymbol{X}_H = \{\boldsymbol{x}_H^{(i)}\}_{i=1}^{N_H}$, and low-fidelity data (e.g., measurements with lower accuracy or numerical approximations of the QoI) $\boldsymbol{y}_L = (y_L^{(1)}, \ldots, y_L^{(N_L)})^\top$ at locations $\boldsymbol{X}_L = \{\boldsymbol{x}_L^{(i)}\}_{i=1}^{N_L}$, where $y_H^{(i)}, y_L^{(i)} \in \mathbb{R}$ and $\boldsymbol{x}_H^{(i)}, \boldsymbol{x}_L^{(i)} \in D \subseteq \mathbb{R}^d$. We denote $\boldsymbol{X} = \{\boldsymbol{X}_L, \boldsymbol{X}_H\}$ and $\widetilde{\boldsymbol{y}} = (\boldsymbol{y}_L^\top, \boldsymbol{y}_H^\top)^\top$. Kennedy and O'Hagan [13] proposed a multifidelity formulation based on the auto-regressive model for GP $Y_H(\cdot)$ ($\sim \mathcal{GP}(\mu_H(\cdot), k_H(\cdot,\cdot))$):

$$Y_H(\boldsymbol{x}) = \rho Y_L(\boldsymbol{x}) + Y_d(\boldsymbol{x}), \tag{2.10}$$

where $Y_L(\cdot)$ ($\sim \mathcal{GP}(\mu_L(\cdot), k_L(\cdot,\cdot))$) regresses the low-fidelity data, $\rho \in \mathbb{R}$ is a regression parameter and $Y_d(\cdot)$ ($\sim \mathcal{GP}(\mu_d(\cdot), k_d(\cdot,\cdot))$) models the discrepancy between Y_H and ρY_L. This model assumes that

$$\mathrm{Cov}\left\{Y_H(\boldsymbol{x}), Y_L(\boldsymbol{x}') \mid Y_L(\boldsymbol{x})\right\} = 0, \quad \text{for all} \quad \boldsymbol{x}' \neq \boldsymbol{x},\ \boldsymbol{x}, \boldsymbol{x}' \in D. \tag{2.11}$$

The covariance of observations, $\widetilde{\boldsymbol{C}}$, is then given by

$$\widetilde{\boldsymbol{C}} = \begin{pmatrix} \boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L) & \rho \boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_H) \\ \rho \boldsymbol{C}_L(\boldsymbol{X}_H, \boldsymbol{X}_L) & \rho^2 \boldsymbol{C}_L(\boldsymbol{X}_H, \boldsymbol{X}_H) + \boldsymbol{C}_d(\boldsymbol{X}_H, \boldsymbol{X}_H) \end{pmatrix}, \tag{2.12}$$

where $\boldsymbol{C}_L$ and $\boldsymbol{C}_d$ are the covariance matrices computed from $k_L(\cdot,\cdot)$ and $k_d(\cdot,\cdot)$, respectively, i.e.,

$$\begin{aligned} &[\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)]_{ij} = k_L(\boldsymbol{X}_L^{(i)}, \boldsymbol{X}_L^{(j)}), \quad [\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_H)]_{ij} = k_L(\boldsymbol{X}_L^{(i)}, \boldsymbol{X}_H^{(j)}), \\ &[\boldsymbol{C}_L(\boldsymbol{X}_H, \boldsymbol{X}_L)]_{ij} = k_L(\boldsymbol{X}_H^{(i)}, \boldsymbol{X}_L^{(j)}), \quad [\boldsymbol{C}_L(\boldsymbol{X}_H, \boldsymbol{X}_H)]_{ij} = k_L(\boldsymbol{X}_H^{(i)}, \boldsymbol{X}_H^{(j)}), \\ &[\boldsymbol{C}_d(\boldsymbol{X}_H, \boldsymbol{X}_H)]_{ij} = k_d(\boldsymbol{X}_H^{(i)}, \boldsymbol{X}_H^{(j)}). \end{aligned} \tag{2.13}$$

One can assume parameterized forms for these kernels (e.g., Gaussian kernel) and employ the following two-step approach [7, 6] to identify hyperparameters:

1. Use Kriging to construct Y_L based on $\{\boldsymbol{X}_L, \boldsymbol{y}_L\}$.
2. Denote $\boldsymbol{y}_d = \boldsymbol{y}_H - \rho \boldsymbol{y}_L(\boldsymbol{X}_H)$, where $\boldsymbol{y}_L(\boldsymbol{X}_H)$ are the values of $\boldsymbol{y}_L$ at locations common to those of $\boldsymbol{X}_H$, then construct Y_d using $\{\boldsymbol{X}_H, \boldsymbol{y}_d\}$ via Kriging.

The posterior mean and variance of Y_H at $\boldsymbol{x}^* \in D$ are given by

$$\hat{y}(\boldsymbol{x}^*) = \mu_H(\boldsymbol{x}^*) + \widetilde{\boldsymbol{c}}(\boldsymbol{x}^*)^\top \widetilde{\boldsymbol{C}}^{-1}(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}}), \tag{2.14}$$

$$\hat{s}^2(\boldsymbol{x}^*) = \rho^2 \sigma_L^2(\boldsymbol{x}^*) + \sigma_d^2(\boldsymbol{x}^*) - \widetilde{\boldsymbol{c}}(\boldsymbol{x}^*)^\top \widetilde{\boldsymbol{C}}^{-1} \widetilde{\boldsymbol{c}}(\boldsymbol{x}^*), \tag{2.15}$$

where $\mu_H(\boldsymbol{x}^*) = \rho \mu_L(\boldsymbol{x}^*) + \mu_d(\boldsymbol{x}^*)$, $\sigma_L^2(\boldsymbol{x}^*) = k_L(\boldsymbol{x}^*, \boldsymbol{x}^*)$, $\sigma_d^2(\boldsymbol{x}^*) = k_d(\boldsymbol{x}^*, \boldsymbol{x}^*)$, and

$$\widetilde{\boldsymbol{\mu}} = \begin{pmatrix} \boldsymbol{\mu}_L \\ \boldsymbol{\mu}_H \end{pmatrix} = \begin{pmatrix} \left(\mu_L(\boldsymbol{x}_L^{(1)}), \ldots, \mu_L(\boldsymbol{x}_L^{(N_L)})\right)^\top \\ \left(\mu_H(\boldsymbol{x}_H^{(1)}), \ldots, \mu_H(\boldsymbol{x}_H^{(N_H)})\right)^\top \end{pmatrix}, \tag{2.16}$$

$$\widetilde{\boldsymbol{c}}(\boldsymbol{x}^*) = \begin{pmatrix} \rho \boldsymbol{c}_L(\boldsymbol{x}^*) \\ \boldsymbol{c}_H(\boldsymbol{x}^*) \end{pmatrix} = \begin{pmatrix} \left(\rho k_L(\boldsymbol{x}_L^{(1)}, \boldsymbol{x}^*), \ldots, \rho k_L(\boldsymbol{x}_L^{(N_L)}, \boldsymbol{x}^*)\right)^\top \\ \left(k_H(\boldsymbol{x}_H^{(1)}, \boldsymbol{x}^*), \ldots, k_H(\boldsymbol{x}_H^{(N_H)}, \boldsymbol{x}^*)\right)^\top \end{pmatrix}, \tag{2.17}$$

where $k_H(\boldsymbol{x}, \boldsymbol{x}') = \rho^2 k_L(\boldsymbol{x}, \boldsymbol{x}') + k_d(\boldsymbol{x}, \boldsymbol{x}')$. Alternatively, one can simultaneously identify hyperparameters in $k_L(\cdot,\cdot)$ and $k_d(\cdot,\cdot)$ along with ρ by maximizing the following log marginal likelihood:

$$\ln \widetilde{L} = -\frac{1}{2}(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{C}}^{-1}(\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}}) - \frac{1}{2} \ln \left|\widetilde{\boldsymbol{C}}\right| - \frac{N_H + N_L}{2} \ln 2\pi. \tag{2.18}$$

2.3 GE-Kriging/Cokriging

GE-Kriging uses the fact that under some condition, the derivative in physical space and the integral in the probability space are interchangable:

$$\begin{aligned}
\frac{\partial}{\partial x_i}\mu(\boldsymbol{x}) &= \frac{\partial}{\partial x_i}\mathrm{E}\left\{Y(\boldsymbol{x})\right\} = \mathrm{E}\left\{\frac{\partial}{\partial x_i}Y(\boldsymbol{x})\right\}, \\
\frac{\partial}{\partial x_i}k(\boldsymbol{x},\boldsymbol{x}') &= \frac{\partial}{\partial x_i}\mathrm{Cov}\left\{Y(\boldsymbol{x}),Y(\boldsymbol{x}')\right\} = \mathrm{Cov}\left\{\frac{\partial}{\partial x_i}Y(\boldsymbol{x}),Y(\boldsymbol{x}')\right\}, \\
\frac{\partial^2}{\partial x_i\partial x_j'}k(\boldsymbol{x},\boldsymbol{x}') &= \frac{\partial^2}{\partial x_i\partial x_j'}\mathrm{Cov}\left\{Y(\boldsymbol{x}),Y(\boldsymbol{x}')\right\} = \mathrm{Cov}\left\{\frac{\partial}{\partial x_i}Y(\boldsymbol{x}),\frac{\partial}{\partial x_j'}Y(\boldsymbol{x}')\right\}.
\end{aligned} \tag{2.19}$$

These formulas specify the covariance between the QoI and its gradient as well as the covariance between different components of the gradient. To simplify the notations, we use ∂_i and $\partial_{i'}$ to denote $\dfrac{\partial}{\partial x_i}$ and $\dfrac{\partial}{\partial x_i'}$, respectively, and $\nabla = (\partial_1, \partial_2, \ldots, \partial_d)^\top$, $\nabla' = (\partial_{1'}, \partial_{2'}, \ldots, \partial_{d'})$. Of note, for a scalar function z, ∇z is a column vector and $\nabla' z$ is a row vector. Since we use a stationary kernel in this work, i.e., $k(\boldsymbol{x},\boldsymbol{x}') = k(\boldsymbol{x}-\boldsymbol{x}')$, we have

$$\partial_i k(\boldsymbol{x},\boldsymbol{x}') = -\partial_{i'} k(\boldsymbol{x},\boldsymbol{x}'). \tag{2.20}$$

The analytical form of $\partial_i k(\boldsymbol{x},\boldsymbol{x}')$ and $\partial_i\partial_{j'} k(\boldsymbol{x},\boldsymbol{x}')$ can be found in the appendix of [30] for widely used kernel functions $k(\boldsymbol{x},\boldsymbol{x}')$, e.g., Matérn kernels with several specific selections of ν. Subsequently, GE-Kriging follows almost the same procedures as those in Kriging with the following modifications [16]:

1. The observation vector is augmented to include gradient data, i.e.,

$$\boldsymbol{y} = (y^{(1)}, y^{(2)}, \ldots, y^{(N)}, (\nabla y^{(1)})^\top, (\nabla y^{(2)})^\top, \ldots, (\nabla y^{(N)})^\top)^\top.$$

2. Given a constant posterior mean of the QoI, the posterior mean of the gradient is zero, hence, $\mathbf{1} = (\underbrace{1,1,\ldots,1}_{N},\underbrace{0,0,\ldots,0}_{N\times d})^\top$.

3. Covariance matrix $\boldsymbol{C} = \sigma^2\boldsymbol{\Psi}$, more specifically, the correlation matrix $\boldsymbol{\Psi}$ is expanded to include correlations between QoI and its gradient as well as correlations between components of the gradient, i.e.,

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{21} & \boldsymbol{\Psi}_{22} \end{bmatrix}, \tag{2.21}$$

where

$$\boldsymbol{\Psi}_{11} = \frac{1}{\sigma^2}\begin{bmatrix} k(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(1)}) & \cdots & k(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(N)}) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}^{(N)},\boldsymbol{x}^{(1)}) & \cdots & k(\boldsymbol{x}^{(N)},\boldsymbol{x}^{(N)}) \end{bmatrix},$$

$$\boldsymbol{\Psi}_{21} = \nabla\boldsymbol{\Psi}_{11} = \frac{1}{\sigma^2}\begin{bmatrix} \partial_1 k(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(1)}) & \cdots & \partial_1 k(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(N)}) \\ \vdots & \ddots & \vdots \\ \partial_d k(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(1)}) & \cdots & \partial_d k(\boldsymbol{x}^{(1)},\boldsymbol{x}^{(N)}) \\ \vdots & \ddots & \vdots \\ \partial_1 k(\boldsymbol{x}^{(N)},\boldsymbol{x}^{(1)}) & \cdots & \partial_1 k(\boldsymbol{x}^{(N)},\boldsymbol{x}^{(N)}) \\ \vdots & \ddots & \vdots \\ \partial_d k(\boldsymbol{x}^{(N)},\boldsymbol{x}^{(1)}) & \cdots & \partial_d k(\boldsymbol{x}^{(N)},\boldsymbol{x}^{(N)}) \end{bmatrix}, \quad \boldsymbol{\Psi}_{12} = \boldsymbol{\Psi}_{21}^\top,$$

$$\boldsymbol{\Psi}_{22} = \nabla'\nabla\boldsymbol{\Psi}_{11} = \begin{bmatrix} \boldsymbol{\psi}_{11} & \cdots & \boldsymbol{\psi}_{1N} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\psi}_{N1} & \cdots & \boldsymbol{\psi}_{NN} \end{bmatrix}, \quad \boldsymbol{\psi}_{lm} = \frac{1}{\sigma^2} \begin{bmatrix} \partial_1\partial_{1'}k(\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}) & \cdots & \partial_1\partial_{d'}k(\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}) \\ \vdots & \ddots & \vdots \\ \partial_d\partial_{1'}k(\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}) & \cdots & \partial_d\partial_{d'}k(\boldsymbol{x}^{(l)}, \boldsymbol{x}^{(m)}) \end{bmatrix}.$$

The posterior mean and variance of the QoI at a new location $\boldsymbol{x}^*$, denoted by $\hat{y}(\boldsymbol{x}^*)$ and $\hat{s}^2(\boldsymbol{x}^*)$, has the same form as in Kriging, i.e., Eqs. (2.8) and (2.9), except that $\boldsymbol{\psi} = \begin{pmatrix} \boldsymbol{\psi}(\boldsymbol{x}^*) \\ \nabla\boldsymbol{\psi}(\boldsymbol{x}^*) \end{pmatrix}$, where $\nabla\boldsymbol{\psi}(\boldsymbol{x}^*) = \frac{1}{\sigma^2}\begin{pmatrix} \nabla k(\boldsymbol{x}^{(1)}, \boldsymbol{x}^*) \\ \vdots \\ \nabla k(\boldsymbol{x}^{(N)}, \boldsymbol{x}^*) \end{pmatrix}$. Furthermore, the posterior mean and variance of the QoI's gradient at $\boldsymbol{x}^*$ are computed as

$$\widehat{\partial_i y}(\boldsymbol{x}^*) = (\partial_{i'}\boldsymbol{\psi})^\top \boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \mathbf{1}\hat{\mu}), \tag{2.22}$$

$$\widehat{s_i}^2(\boldsymbol{x}^*) = \hat{\sigma}^2\left[1 - (\partial_{i'}\boldsymbol{\psi})^\top \boldsymbol{\Psi}^{-1}\partial_{i'}\boldsymbol{\psi}\right], \tag{2.23}$$

where $\partial_{i'}\boldsymbol{\psi} = \begin{pmatrix} \partial_{i'}\boldsymbol{\psi}(\boldsymbol{x}^*) \\ \partial_{i'}(\nabla\boldsymbol{\psi}(\boldsymbol{x}^*)) \end{pmatrix}$ and $i = 1, 2, \ldots, d$.

Next, we introduce the details of GE-Cokriging method, which also shares a similar construction procedure as Cokriging except for some modifications to incorporate gradient information. Such modifications are as follows:

1. The observation vector is augmented to $\widetilde{\boldsymbol{y}} = \left(\boldsymbol{y}_L^\top, \boldsymbol{y}_H^\top, (\nabla\boldsymbol{y}_L)^\top, (\nabla\boldsymbol{y}_H)^\top\right)^\top$ and is of length $N_L + N_H + (N_L + N_H)d$.

2. The covariance matrix of the observation data, $\widetilde{\boldsymbol{C}}$ in Eq. (2.12), is augmented to include gradient information as well, i.e.,

$$\widetilde{\boldsymbol{C}} = \begin{pmatrix} \widetilde{\boldsymbol{C}}_{11} & \widetilde{\boldsymbol{C}}_{12} \\ \widetilde{\boldsymbol{C}}_{21} & \widetilde{\boldsymbol{C}}_{22} \end{pmatrix} \tag{2.24}$$

where $\widetilde{\boldsymbol{C}}_{11}$ takes the form of covariance matrix in Cokriging, see Eq. (2.12), and

$$\widetilde{\boldsymbol{C}}_{21} = \begin{bmatrix} \nabla\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L) & \rho\nabla\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_H) \\ \nabla\boldsymbol{C}_L(\boldsymbol{X}_H, \boldsymbol{X}_L) & \rho^2\nabla\boldsymbol{C}_L(\boldsymbol{X}_H, \boldsymbol{X}_H) + \nabla\boldsymbol{C}_d(\boldsymbol{X}_H, \boldsymbol{X}_H) \end{bmatrix}, \qquad \widetilde{\boldsymbol{C}}_{12} = \widetilde{\boldsymbol{C}}_{21}^\top,$$

$$\widetilde{\boldsymbol{C}}_{22} = \begin{bmatrix} \nabla'\nabla\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L) & \rho\nabla'\nabla\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_H) \\ \rho\nabla'\nabla\boldsymbol{C}_L(\boldsymbol{X}_H, \boldsymbol{X}_L) & \rho^2\nabla'\nabla\boldsymbol{C}_L(\boldsymbol{X}_H, \boldsymbol{X}_H) + \nabla'\nabla\boldsymbol{C}_d(\boldsymbol{X}_H, \boldsymbol{X}_H) \end{bmatrix}.$$

Here $\nabla\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)$ is a matrix constructed by replacing each element in $\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)$, i.e., $[\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)]_{ij}$, with its gradient $(\partial_1[\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)]_{ij}, \ldots, \partial_d[\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)]_{ij})^\top$. Similarly, $\nabla\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_H)$, $\nabla\boldsymbol{C}_L(\boldsymbol{X}_H, \boldsymbol{X}_L)$, $\nabla\boldsymbol{C}_L(\boldsymbol{X}_H, \boldsymbol{X}_H)$ and $\nabla\boldsymbol{C}_d(\boldsymbol{X}_H, \boldsymbol{X}_H)$ are constructed by replacing elements in corresponding matrices in Eq. (2.13) with their gradients, respectively. The matrix $\nabla'\nabla\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)$ is constructed by replacing each element in $\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)$, i.e., $[\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)]_{ij}$, with the matrix

$$\begin{pmatrix} \partial_1\partial_{1'}[\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)]_{ij} & \cdots & \partial_1\partial_{d'}[\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)]_{ij} \\ \vdots & \ddots & \vdots \\ \partial_d\partial_{1'}[\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)]_{ij} & \cdots & \partial_d\partial_{d'}[\boldsymbol{C}_L(\boldsymbol{X}_L, \boldsymbol{X}_L)]_{ij} \end{pmatrix}.$$

Other submatrices in $\widetilde{\boldsymbol{C}}_{22}$ are constructed in the same manner.

3. The posterior mean vector now becomes

$$\widetilde{\boldsymbol{\mu}} = \begin{pmatrix} \boldsymbol{\mu}_L \\ \boldsymbol{\mu}_H \\ \mathbf{0}_L \\ \mathbf{0}_H \end{pmatrix} = \begin{pmatrix} \left(\mu_L(\boldsymbol{x}_L^{(1)}), \ldots, \mu_L(\boldsymbol{x}_L^{(N_L)})\right)^\top \\ \left(\mu_H(\boldsymbol{x}_H^{(1)}), \ldots, \mu_H(\boldsymbol{x}_H^{(N_H)})\right)^\top \\ \underbrace{(0, \ldots, 0)}_{N_L \cdot d}{}^\top \\ \underbrace{(0, \ldots, 0)}_{N_H \cdot d}{}^\top \end{pmatrix}. \tag{2.25}$$

4. The covariance vector between the new observation location $\boldsymbol{x}^*$ and existing observation data $[\boldsymbol{X}_L, \boldsymbol{X}_H]$, denoted by $\widetilde{\boldsymbol{c}}(\boldsymbol{x}^*)$, is given by

$$\widetilde{\boldsymbol{c}}(\boldsymbol{x}^*) = \begin{pmatrix} \rho \boldsymbol{c}_L(\boldsymbol{x}^*) \\ \boldsymbol{c}_H(\boldsymbol{x}^*) \\ \rho \nabla \boldsymbol{c}_L(\boldsymbol{x}^*) \\ \nabla \boldsymbol{c}_H(\boldsymbol{x}^*) \end{pmatrix}, \tag{2.26}$$

where $\boldsymbol{c}_L(\boldsymbol{x}^*) = \left(k_L(\boldsymbol{x}_L^{(1)}, \boldsymbol{x}^*), \ldots, k_L(\boldsymbol{x}_L^{(N_L)}, \boldsymbol{x}^*)\right)^\top$ and $\boldsymbol{c}_H(\boldsymbol{x}^*) = \left(k_H(\boldsymbol{x}_H^{(1)}, \boldsymbol{x}^*), \ldots, k_H(\boldsymbol{x}_H^{(N_H)}, \boldsymbol{x}^*)\right)^\top$.

The estimators for the mean and standard deviation of QoI at the new observation location $\boldsymbol{x}^*$ in GE-Cokriging follow Eqs. (2.14) and (2.15) in Cokriging method with corresponding components updated as shown above.

We provide the formulas for the posterior mean and variance of the QoI's gradient at $\boldsymbol{x}^*$ as follows:

$$\widehat{\partial_i y}(\boldsymbol{x}^*) = (\partial_{i'} \widetilde{\boldsymbol{c}}(\boldsymbol{x}^*))^\top \widetilde{\boldsymbol{C}}^{-1} (\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{\mu}}), \tag{2.27}$$

$$\widehat{s_i}^2(\boldsymbol{x}^*) = \rho^2 \partial_i \partial_{i'} k_L(\boldsymbol{x}^*, \boldsymbol{x}^*) + \partial_i \partial_{i'} k_H(\boldsymbol{x}^*, \boldsymbol{x}^*) - [\partial_{i'} \widetilde{\boldsymbol{c}}(\boldsymbol{x}^*)]^\top \widetilde{\boldsymbol{C}}^{-1} \partial_{i'} \widetilde{\boldsymbol{c}}(\boldsymbol{x}^*), \tag{2.28}$$

where $i = 1, 2, \ldots, d$. The derivation of Eqs. (2.27) and (2.28) follow the same procedure as Eqs. (2.14) and (2.15) shown in [13, 6]. In other words, Eqs. (2.27) and (2.28) can be obtained by replacing $Y(\boldsymbol{x})$ in Eqs. (2.14) and (2.15) with $\partial_i Y(\boldsymbol{x})$. More specifically, $\mu_H(\boldsymbol{x}^*)$ is replaced with the mean of $\partial_i Y(\boldsymbol{x})$ (which is zero), $\widetilde{\boldsymbol{c}}$ is replaced with $\partial_i \widetilde{\boldsymbol{c}}$, and $\rho^2 \sigma_L^2(\boldsymbol{x}^*) + \sigma_d^2(\boldsymbol{x}^*)$ (i.e., $\rho^2 \mathrm{Var}\{Y_L(\boldsymbol{x}^*)\} + \mathrm{Var}\{Y_d(\boldsymbol{x}^*)\}$) is replaced with $\rho^2 \mathrm{Var}\{\partial_i Y_L(\boldsymbol{x}^*)\} + \mathrm{Var}\{\partial_i Y_d(\boldsymbol{x}^*)\} = \rho^2 \partial_i \partial_{i'} k_L(\boldsymbol{x}^*, \boldsymbol{x}^*) + \partial_i \partial_{i'} k_H(\boldsymbol{x}^*, \boldsymbol{x}^*)$.

We note that the GE-Cokriging exploits the relation between QoI and its gradients, and once the hyper-parameters in the model are identified, we can compute the posterior mean and variance of the QoI and its gradients *simultaneously*. It has the potential to improve the accuracy of the prediction for both QoI and its gradients compared with predicting them separately. Also, in some cases, this approach can reduce computational cost compared to, for example, constructing Cokriging models for QoI and its gradients separately (see Section 3.5).

2.4 Integral-enhanced Kriging/Cokriging

In this section, we provide another perspective on using the QoI f and its gradients ∇f in GPR simultaneously. The aforementioned gradient-enhanced methods firstly assume a GP model $Y(\boldsymbol{x})$ for f, and the GP model for ∇f can be constructed accordingly by taking (partial) derivatives of $Y(\boldsymbol{x})$'s mean and covariance function. Alternatively, one can also assume a GP model for ∇f first, e.g., $\partial_i f$ is modeled by $Y(\boldsymbol{x})$, then the QoI f can be modeled by $\int Y(\boldsymbol{x}) \mathrm{d}x_i$, which is a GP because integral is a linear operator. Here we use the univariate function to further illustrate the concept. We model f' with GP $Y_{f'}(\boldsymbol{x}) \sim \mathcal{GP}(\mu_{f'}(\boldsymbol{x}), k_{f'}(\boldsymbol{x}, \boldsymbol{x}'))$,

then similar to Eqs. (2.19), the integrals in the physical space and in the probability space are interchangeable:

$$\begin{aligned}
\int \mu_{f'}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= \int \mathrm{E}\left\{Y_{f'}(\boldsymbol{x})\right\}\mathrm{d}\boldsymbol{x} = \mathrm{E}\left\{\int Y_{f'}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\},\\
\int k_{f'}(\boldsymbol{x},\boldsymbol{x}')\mathrm{d}\boldsymbol{x} &= \int \mathrm{Cov}\left\{Y_{f'}(\boldsymbol{x}), Y_{f'}(\boldsymbol{x}')\right\}\mathrm{d}\boldsymbol{x}\\
&= \int \mathrm{E}\left\{(Y_{f'}(\boldsymbol{x})-\mu_{f'}(\boldsymbol{x}))(Y_{f'}(\boldsymbol{x}')-\mu_{f'}(\boldsymbol{x}'))\right\}\mathrm{d}\boldsymbol{x}\\
&= \mathrm{E}\left\{\left[\int (Y_{f'}(\boldsymbol{x})-\mu_{f'}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\right](Y_{f'}(\boldsymbol{x}')-\mu_{f'}(\boldsymbol{x}'))\right\}\\
&= \mathrm{Cov}\left\{\int Y_{f'}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, Y_{f'}(\boldsymbol{x}')\right\},\\
\int\int k_{f'}(\boldsymbol{x},\boldsymbol{x}')\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' &= \int\int \mathrm{Cov}\left\{Y_{f'}(\boldsymbol{x}), Y_{f'}(\boldsymbol{x}')\right\}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}'\\
&= \int\int \mathrm{E}\left\{(Y_{f'}(\boldsymbol{x})-\mu_{f'}(\boldsymbol{x}))(Y_{f'}(\boldsymbol{x}')-\mu_{f'}(\boldsymbol{x}'))\right\}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}'\\
&= \mathrm{E}\left\{\int (Y_{f'}(\boldsymbol{x})-\mu_{f'}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\int (Y_{f'}(\boldsymbol{x}')-\mu_{f'}(\boldsymbol{x}'))\mathrm{d}\boldsymbol{x}'\right\}\\
&= \mathrm{Cov}\left\{\int Y_{f'}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \int Y_{f'}(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}'\right\}.
\end{aligned} \tag{2.29}$$

These formulas provide the mean and covariance of the GP $Y_f(\boldsymbol{x}) = \int Y_{f'}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ as well as the covariance between $Y_f(\boldsymbol{x})$ and $Y_{f'}(\boldsymbol{x})$. Of note, we use indefinite integral here and the constant associated with this integral needs identification via maximizing the log marginal likelihood. But this constant will not affect the covariance function, because $\mathrm{Cov}\left\{\int Y_{f'}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \int Y_{f'}(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}'\right\} = \mathrm{Cov}\left\{\int Y_{f'}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + a\int Y_{f'}(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}' + b\right\}$ for any constants a and b.

Then we can follow the same procedure in the gradient-enhanced Kriging in Section 2.3 to construct the covariance matrix $\boldsymbol{C}$ and compute the posterior mean and variance of f and f' at any location $\boldsymbol{x}^*$. Of note, this "integral-enhanced" GPR/Kriging is equivalent to the gradient-enhanced version. For example, if we set the mean of $Y_{f'}(\boldsymbol{x})$ to be zero, then the mean of $Y_f(\boldsymbol{x})$ is a constant μ, which needs identifying as in the gradient-enhanced version. Subsequently, the integral-enhanced Kriging is equivalent to the equivalence of the gradient-enhanced Kriging if the mean and covariance functions are selected appropriately. For example, if we assume zero mean and set $k_{f'}(\boldsymbol{x},\boldsymbol{x}') = \dfrac{\partial^2}{\partial x_i \partial x_j'} k_f(\boldsymbol{x},\boldsymbol{x}')$ for $Y_{f'}(\boldsymbol{x})$, where $k_f(\boldsymbol{x},\boldsymbol{x}')$ is the Gaussian kernel function, this integral-enhanced Kriging model is the same as the gradient-enhanced Kriging model that uses Gaussian kernel function and constant mean for $Y_f(\boldsymbol{x})$. In most cases, it is easier to compute the (partial) derivatives than to compute the integral. Therefore, it is more convenient to use the gradient-enhanced setting. The similar argument holds for Cokriging. In this work, we only show the results of gradient-enhanced Kriging/Cokriging.

3 Numerical examples

We present four numerical examples to demonstrate the performance of GE-Cokriging. The first two prototype examples show the capability GE-Cokriging's capability of approximating the QoI and its gradients of two 1D functions and a 2D function. The other two examples illustrate the high precision of GE-Cokriging in constructing the phase diagram of an underdamped oscillator and analyzing the sensitivity of power factor under varying power inputs in a large-scale power grid system. In all these examples, we assume that both the QoI and its gradients are collected at every observation locations. The hyperparameters in GP models are identified by maximizing associated log marginal likelihood function using genetic algorithm as in [6].

Lastly, we compare the prediction accuracy using Cokriging, GE-Kriging and GE-Cokriging in each case quantitatively. We also compare the computational cost of these methods in each case.

3.1 1D function

In this part, we compare the results of Cokriging and GE-Cokriging in approximating a 1D function. In this case, the target function to approximate is,

$$f_H(x) = (6x-2)^2 \sin(12x-4), \tag{3.1}$$

from which high-fidelity data are sampled. The low-fidelity data are sampled from the following function

$$f_L(x) = Af_H(x) + B(x-0.5) + C. \tag{3.2}$$

The observation locations of f_H are $X_H = \{0, 0.2, 0.6, 1.0\}$, and those for f_L are $X_L = \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. Here, the observation locations of data are chosen so that $X_H \subset X_L$.

3.1.1 1D Case 1: a classical case

We first show a well-studied case where parameters of low-fidelity function is given by $A = 0.5, B = 10, C = -5$ as in [6]. Hence, the low-fidelity function is

$$f_{L1}(x) = 0.5f_H(x) + 10(x-0.5) - 5. \tag{3.3}$$

Of note, we use fewer observation points in $\boldsymbol{X}_L$ than in [6].

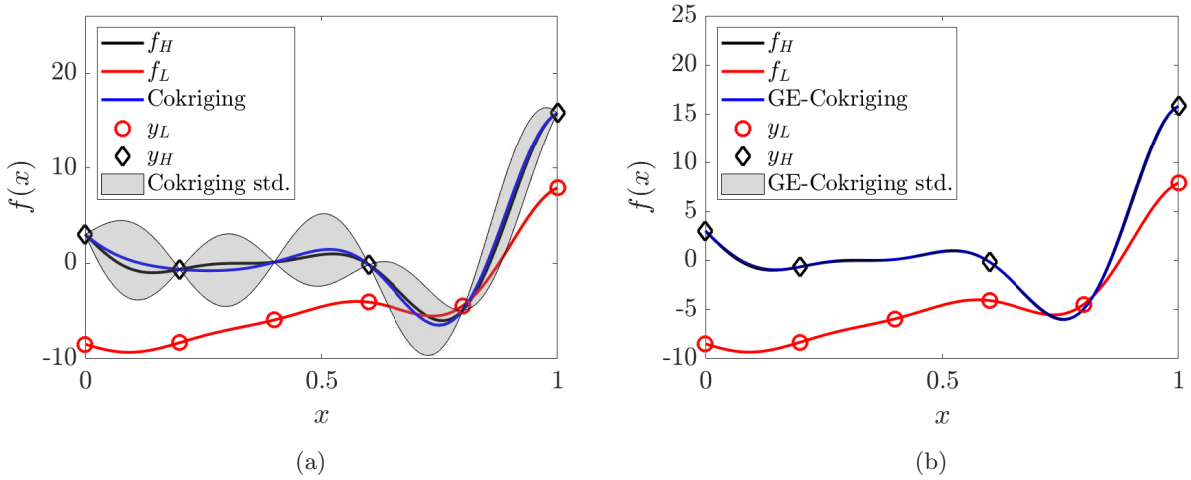

Figure 1: Prediction of the QoI for the 1D problem case 1. Prediction of posterior mean (black solid line) and standard deviation (grey shaded area) of QoI f_H by (a) Cokriging and (b) GE-Cokriging. The low-fidelity function f_{L1} is denoted by red solid lines, high-fidelity samples are denoted by black diamonds and low-fidelity samples by red circles. Colored online.

The results of Cokriging and GE-Cokriging for reconstructing f_H are shown in Fig. 1. Fig. 1a shows that Cokriging is able to capture f_H as the posterior mean is generally close to the high-fidelity function value. However, $\hat{s}$ of the prediction are large on most of the prediction locations, which indicates that Cokriging method yields considerable uncertainty at those locations, whereas this uncertainty is very small at X_c because a simple relation has been found between f_H and f_L based on available data [6]. As a comparison,

Fig. 1b illustrates that the posterior mean of GE-Cokriging coincides with f_H, and the uncertainty in the prediction is very small on the entire interval as the grey shaded area is almost invisible.

Next, we compare the performance of predicting the gradients of f_H, i.e., $\frac{\mathrm{d}f_H(x)}{\mathrm{d}x}$. Fig. 2 shows that Cokriging method suffers from the singularity of the covariance matrix in this setup, implied from sharp turning of predicted curvature between neighboring observations in Fig. 2a and large standard deviations in Fig. 2b on locations where observations are not available. As for GE-Cokriging method, the prediction of gradients is accurate both in terms of posterior mean illustrated in Fig. 2a and standard deviation illustrated Fig. 2b, which shows that the prediction uncertainty by Cokriging is almost 10 times greater than that by GE-Cokriging. We note that the performance of Cokriging is poor in this case because the covariance matrix $\tilde{\boldsymbol{C}}$ is close to a singular matrix. The reason for this phenomenon is that the value of $\frac{\mathrm{d}y_L}{\mathrm{d}x}$ is close at $x = 0.2$ and $x = 0.4$, as well as at $x = 0$ and $x = 0.6$. As we point out in Section 2.1, this singularity issue is common for GPR method in practice, and the typical approach to alleviate this is to add a diagonal matrix αI to the covariance matrix, which is equivalent to add noises in the collected data. In this paper, we set $\alpha = 10^{-14}$, which is much smaller than typical numbers used in practice, to demonstrate that the GE-Cokriging can help to alleviate the singularity issue without sacrificing accuracy of matching observation data.

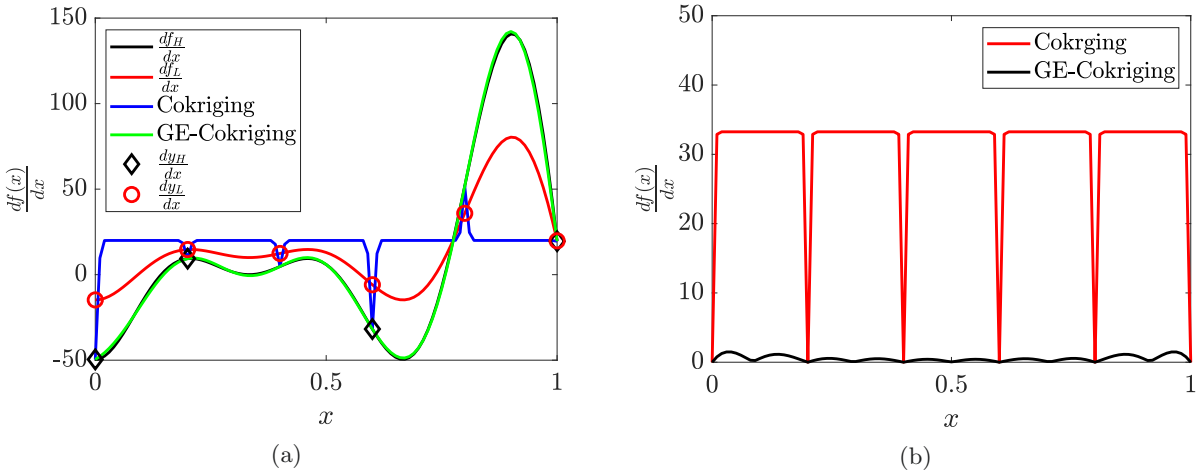

Figure 2: Prediction of the gradient of QoI for the 1D problem case 1. Prediction of posterior (a) mean by Cokriging (blue solid line) and GE-Cokriging (green solid line), where the gradient of high-fidelity function $\frac{\mathrm{d}f_H}{\mathrm{d}x}$ is denoted by black solid line, gradient of low-fidelity function $\frac{\mathrm{d}f_{L1}}{\mathrm{d}x}$ is denoted by red solid line, high-fidelity samples are denoted by black diamonds and low-fidelity samples by red circles and (b) standard deviation for gradient of QoI $\frac{\mathrm{d}f_H}{\mathrm{d}x}$ by Cokriging (red solid line) and GE-Cokriging (black solid line). Colored online.

3.1.2 1D Case 2: shifted f_{L1}

Next, we keep the sampling locations, i.e., X_H and X_L same as those in Section 3.1.1, and only modify the model parameters of the low-fidelity function in Eq. (3.3) by slightly shifting it, i.e., replace x with $x - 0.005$, resulting in the following form of low-fidelity function f_{L2},

$$f_{L2}(x) = f_{L1}(x - 0.005) = 0.5 f_H(x - 0.005) + 10(x - 0.005 - 0.5) - 5. \tag{3.4}$$

The posterior means and standard deviations of Cokriging and GE-Cokriging are shown in Fig. 3. It is shown in Fig. 3a that the Cokriging method is not able to obtain an accurate prediction of f_H, and the resulting uncertainty is large on the entire interval except for locations of X_H. On the contrary, as shown in Fig. 3b, the GE-Cokriging result is much closer to f_H and the uncertainty is very small.

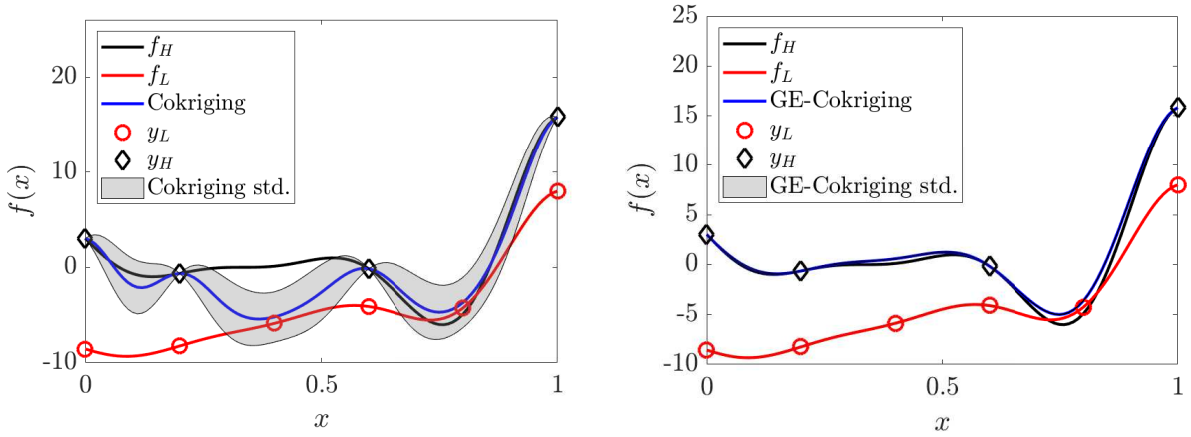

Figure 3: Prediction of the QoI for the 1D problem case 2. Prediction of posterior mean (black solid line) and standard deviation (grey shaded area) of QoI f_H by (a) Cokriging and (b) GE-Cokriging. The low-fidelity function f_{L2} is denoted by red solid lines, high-fidelity samples are denoted by black diamonds and low-fidelity samples by red circles. Colored online.

We present the prediction results of gradients by GE-Cokriging and Cokriging in Fig. 4. Similar to the observations from Fig. 2a, Cokriging in this case suffers from the singularity of the covariance matrix, with posterior mean deviating significantly from f_H (see Fig. 4a) and standard deviation being in the order comparable to its mean value (see Fig. 4b). In comparison, GE-Cokriging still yields a good result with posterior mean close to $\frac{\mathrm{d}f_H}{\mathrm{d}x}$ (see Fig. 4a) and low uncertainty, i.e., small standard deviations (see Fig. 4b). These contrasts between the Cokriging and GE-Cokriging suggest that the gradient information from high-fidelity function and low-fidelity function can help to improve the prediction accuracy of not only QoI but also the corresponding gradients.

3.2 Branin function

We extend the application of GE-Cokriging method in approximating a 2D function, namely a modified Branin function [6], given by

$$f_H(x,y) = a(\bar{x}_2 - b\bar{x}_1^2 + c\bar{x}_1 - r)^2 + g(1-p)\cos(\bar{x}_1) + g + qx, \tag{3.5}$$

where

$$\bar{x}_1 = 15x - 5, \bar{x}_2 = 15y, x \in [0,1], y \in [0,1],$$

with

$$a = 1, b = \frac{5.1}{4\pi^2}, c = \frac{5}{\pi}, r = 6, g = 10, p = \frac{1}{8\pi}, q = 5,$$

and the low-fidelity function is constructed as follows,

$$f_L(x,y) = Af_H(Bx + (1-B), Cy), \tag{3.6}$$

where A = 1.1, B = 0.95, C = 0.9. The contour of the modified Branin function f_H that we aim to approximate is shown in Fig. 5a and the contour for the low-fidelity function f_L is shown in Fig. 5d. The samples for high-fidelity observation locations $\boldsymbol{X}_H$ (black squares in Fig. 5a) and low-fidelity observation locations $\boldsymbol{X}_L$ (black circles in Fig. 5d) are randomly selected from the uniformly spaced grid of size 41×41 on the domain $[0,1] \times [0,1] \in \mathbb{R}^2$. We note that $\boldsymbol{X}_H \subset \boldsymbol{X}_L$ as before.

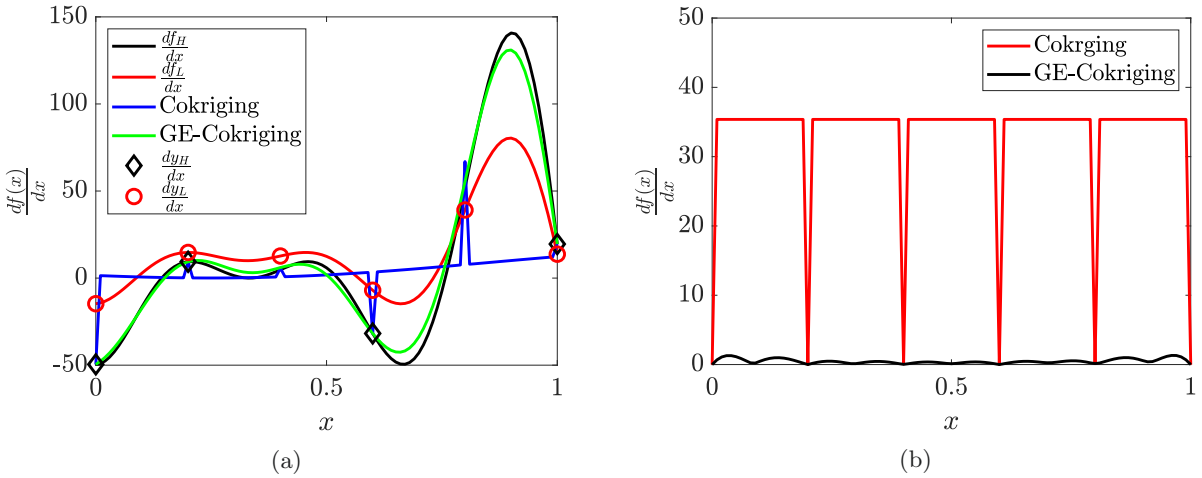

Figure 4: Prediction of the gradient of QoI for the 1D problem case 2. Prediction of posterior (a) mean by Cokriging (blue solid line) and GE-Cokriging (green solid line), where the gradient of high-fidelity function $\frac{\mathrm{d}f_H}{\mathrm{d}x}$ is denoted by black solid line, gradient of low-fidelity function $\frac{\mathrm{d}f_{L2}}{\mathrm{d}x}$ is denoted by red solid line, high-fidelity samples are denoted by black diamonds and low-fidelity samples by red circles and (b) standard deviation for gradient of QoI $\frac{\mathrm{d}f_H}{\mathrm{d}x}$ by Cokriging (red solid line) and GE-Cokriging (black solid line). Colored online.

We first compare the results of reconstructing f_H by Cokriging and GE-Cokriging shown in Fig. 5. It is clear that the posterior mean of GE-Cokriging (Fig. 5c) is closer to f_H than that of Cokriging (Fig. 5b). Also the degree of uncertainty is distinct as posterior standard deviation of Cokriging (Fig. 5e) is one order of magnitude larger than that in GE-Cokriging (Fig. 5f).

Next, we compare the prediction of gradients by Cokriging and GE-Cokriging. Fig. 6a and Fig. 6d profile contours of exact $\frac{\partial f_H}{\partial x}$ and $\frac{\partial f_H}{\partial y}$, respectively. For predicting $\frac{\partial f_H}{\partial x}$, GE-Cokriging (Fig. 6c) shows higher accuracy globally while Cokriging (Fig. 6b) can not result in accurate prediction in the lower left corner, where the available observation data is rare. As for $\frac{\partial f_H}{\partial y}$, since the target function is relatively smooth, both Cokriging (Fig. 6e) and GE-Cokriging (Fig. 6f), are capable of obtaining accurate prediction, while GE-Cokriging still outperforms Cokriging in the sense of the total RMSE recorded in Tab. 1.

3.3 Underdamped oscillator

We consider a driven harmonic oscillator described by the following second order ODE:

$$\begin{cases} m\ddot{x} + c\dot{x} + kx = F(t), \\ x(0) = 1, \quad \dot{x}(0) = 0, \end{cases} \tag{3.7}$$

where m is the mass, c is the damping coefficient, k is a constant (e.g., elasticity coefficient of a string), and $F(t)$ is the external force. We rewrite the ODE in Eq. (3.7) as

$$\ddot{x} + 2\zeta\omega_0\dot{x} + \omega_0^2 x = \frac{F(t)}{m}, \tag{3.8}$$

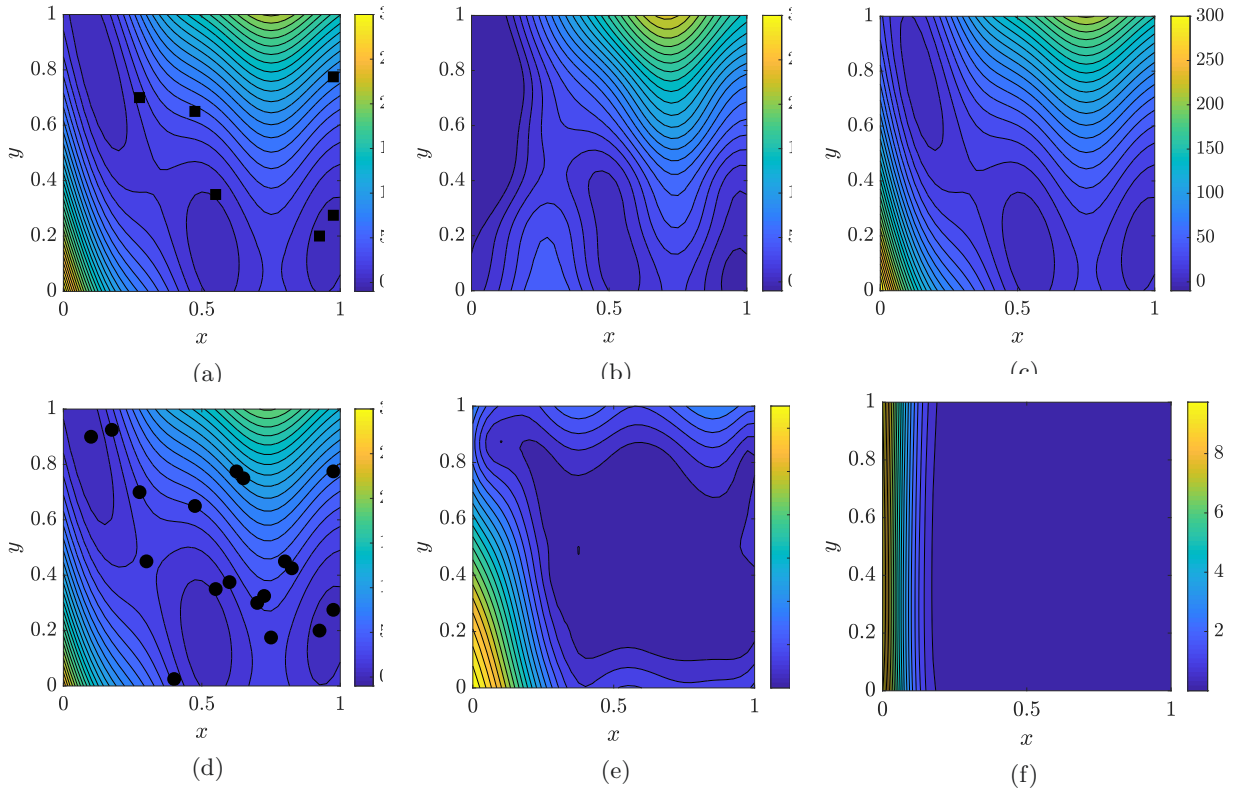

Figure 5: The high-fidelity and low-fidelity function of the 2D problem and the posterior prediction for the high-fidelity function. (a) The high-fidelity function, namely the modified Brainin function f_H (contour) and observation locations (black squares). Posterior mean of QoI prediction by (b) Cokriging and (c) GE-Cokriging. (d) Low-fidelity function f_L (contour) and observation locations (black dots). Posterior standard deviation of QoI by (e) Cokriging and (f) GE-Cokriging. Colored online.

where $\omega_0 = \sqrt{\dfrac{k}{m}}$ is the undamped angular frequency, and $\zeta = \dfrac{c}{2\sqrt{mk}}$ is the damping ratio. We set $\zeta = 1/\sqrt{37}$ and $\omega_0 = \dfrac{6}{\sqrt{1-\zeta^2}}$ in this study. The external force is set as the step response:

$$\frac{F(t)}{m} = \begin{cases} \omega_0^2, & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{3.9}$$

The analytical solution to Eq. (3.7) is

$$x_H(t) = \mathrm{e}^{-\zeta\omega_0 t}\frac{\sin(\sqrt{1-\zeta^2}\omega_0 t + \varphi)}{\sin\varphi}, \quad \varphi = \arccos\zeta, \tag{3.10}$$

and the velocity is

$$\dot{x}_H(t) = -\frac{\omega_0 \mathrm{e}^{-\zeta\omega_0 t}}{\sin\varphi}\left[\zeta\sin(\sqrt{1-\zeta^2}\omega_0 t + \varphi) - \sqrt{1-\zeta^2}\cos(\sqrt{1-\zeta^2}\omega_0 t + \varphi)\right]. \tag{3.11}$$

The low-fidelity model is a simple harmonic oscillator model:

$$\begin{cases} m\ddot{x} + kx = 0, \\ x(0) = 1, \quad \dot{x}(0) = 0, \end{cases} \tag{3.12}$$

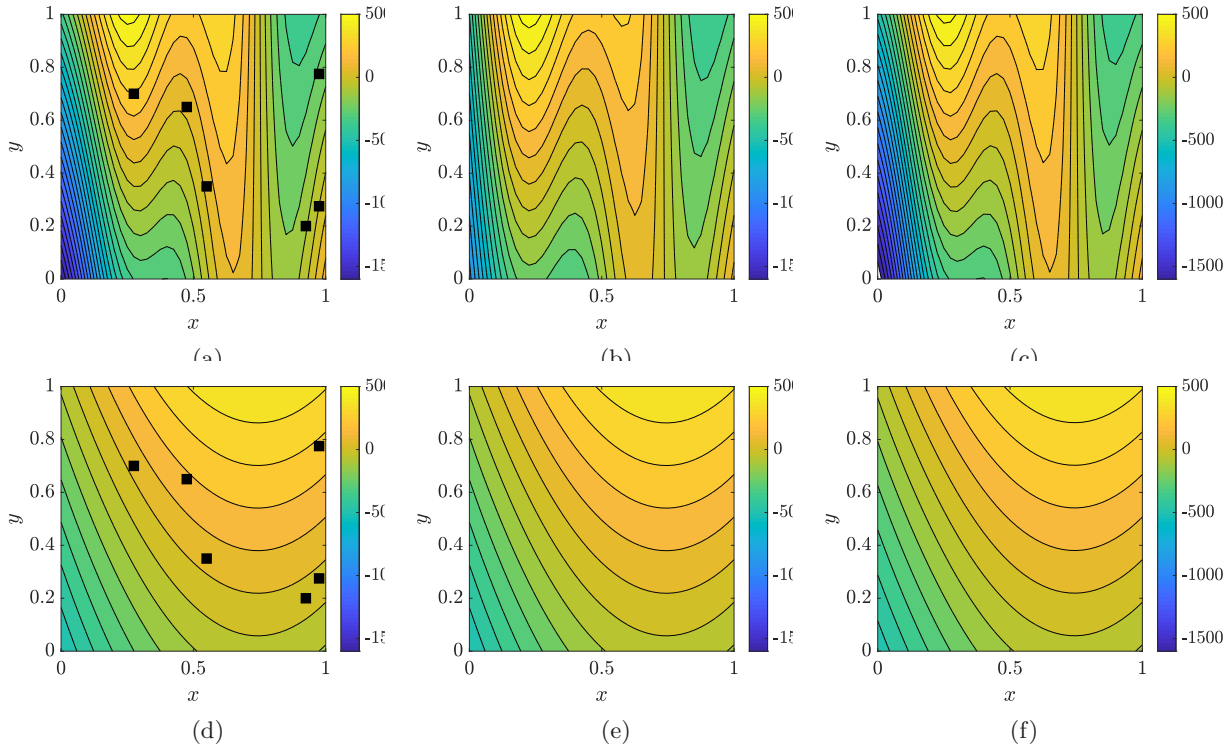

Figure 6: The high-fidelity gradients in x and y directions of the 2D problem and the corresponding posterior predictions. (a) The gradient of high-fidelity function in x direction, $\frac{df_H}{dx}$ (contour) and high-fidelity samples (black squares) of gradient in x direction. Posterior mean of gradient prediction in x direction by (b) Cokriging and (c) GE-Cokriging. (d) The gradient of high-fidelity function in y direction, $\frac{df_H}{dy}$ (contour) and high-fidelity samples (black squares) of gradient in y direction. Posterior mean of gradient prediction in y direction by (e) Cokriging and (f) GE-Cokriging. Colored online.

which is equivalent to setting $\zeta = 0$ and $F(t) = 0$ in Eq. (3.8). The analytical solution to the low-fidelity model is

$$x_L(t) = \cos(\omega_0 t), \tag{3.13}$$

and the velocity is

$$\dot{x}_L(t) = -\omega_0 \sin(\omega_0 t). \tag{3.14}$$

The observation locations for high- and low-fidelity models are set as $T_H = \{0.6j\}_{j=0}^{5}$ and $T_L = \{0.3j\}_{j=0}^{10}$, respectively. We compare the constructed trajectory $x(t)$ and velocity $\dot{x}(t)$ on $[0, 3]$ by Cokriging and GE-Cokriging in Fig. 7. Cokriging again shows worse performance both for prediction of QoI (Fig. 7a) and gradient (Fig. 7c) marked by significant deviations from the true values as well as large uncertainties at locations distant from observation locations, while GE-Cokriging manages to reconstruct the trajectory (Fig. 7b) and velocity (Fig. 7d) of the oscillator well with small standard deviations. The overlapping between trajectory-velocity phase diagram by GE-Cokriging and the exact phase diagram (Fig. 7e) emphasizes that GE-Cokriging can provide accurate predictions for QoI and the corresponding gradients simultaneously, while Cokriging failed to. We also note that Cokring suffers from singularity of the covariance matrix again, while GE-Cokriging doesn't have this concern.

3.4 Sensitivity of a power grid system

We now consider the relationship between the power input of a generator bus, denoted as x, and real-time power factor of a load bus, as $f(x)$, in a large-scale power system from IEEE 118 bus test case [26]. We use MATPOWER [36], which provides a model for the IEEE 118 bus test case, to run simulations and generate sample points. The $f_H(x)$ and $f_L(x)$ represent the alternative current (AC) and direct current (DC) models approximating $f(x)$, respectively.

The observation locations for Cokriging and GE-Cokriging consist of 51 low-fidelity samples from DC model on $X_L = \{20 + 2j\}_{j=0}^{50}$ and five samples from AC model on $X_H = \{40, 48, 72, 98, 116\}$ (again, $X_H \subset X_L$). In addition to reconstructing f_H accurately, estimating the change of power factor of a load bus in response to the change of power input of a generate bus, i.e., the sensitivity of f with respect to x, is important for safety or energy-efficiency consideration. This change is reflected by the derivative of $f(x)$, i.e., $\frac{\mathrm{d}f(x)}{\mathrm{d}x}$. Therefore, we aim to approximate both f_H and its derivative. Here we use finite-difference method to obtain $\frac{\mathrm{d}f_H}{\mathrm{d}x}$ and $\frac{\mathrm{d}f_L}{\mathrm{d}x}$ at X_H and X_L, respectively, and the step size is 0.25.

The results in Fig. 8 suggest that the Cokriging method can approximate f_H with noticeable standard deviations (Fig. 8a), but it fails to reconstruct $\frac{\mathrm{d}f_H}{\mathrm{d}x}$ (Fig. 8c). On the other hand, GE-Cokriging can reconstruct both f_H (Fig. 8b) and $\frac{\mathrm{d}f_H}{\mathrm{d}x}$ (Fig. 8d) accurately with rather small uncertainty, and the only noticeable discrepancy appears near the left boundary because that region is far from available data. Unlike other cases, here we notice the occurrence of wiggling in the high-fidelity gradient prediction by GE-Cokriging. This is caused by the aliasing error as we used finite-different method to approximate the gradient functions, recall that no wiggling is observed in previous examples where the gradients are observed directly. Again, reconstructing the gradient using Cokriging suffers from the sigularity of the covariance matrix as shown in Fig. 8c, whereas GE-Cokriging doesn't have this concern (see Fig. 8d).

3.5 Quantitative comparison and computational efficiency

To analyze and compare the accuracy and efficiency among Cokriging, GE-Kriging and GE-Cokriging, we run simulations for five times with random initial conditions for each numerical example, and list the relative mean squared errors for QoI prediction and gradient of QoI prediction in Tab. 1. The numerical simulations were performed on the same laptop with Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz. We recorded the time for each separate run and computed the corresponding mean and standard deviation from these 5 runs for each example (see Tab. 2).

The results in Tab. 1 show that GE-Cokriging outperforms Cokriging and GE-Kriging in terms of relative mean squared error for all examples presented. We note that in GE-Kriging, only high-fidelity QoI data (including high-fidelity gradient data) was used for training. GE-Cokriging improves accuracy in all cases compared to Cokriging, which is consistent to the visual observations shown in each numerical example. It is also worth noting that the relative mean squared errors by Cokriging are almost one order of magnitude higher than those by GE-Cokriging in most of the cases. The errors in the prediction of QoIs by GE-Kriging are several times larger than those by GE-Cokriging, and the prediction of gradients by GE-Kriging are even worse than those by GE-Cokriging, in all examples. Hence, among these three methods compared, GE-Cokriging is able to maintain a robust prediction result both in terms of QoI and in terms of the gradient of QoI simultaneously, while the other two methods can not obtain comparable results. This further verifies that the information of QoI and its gradients can be strongly correlated, and hence is of great help to improve the accuracy of GPR methods when used jointly.

Tab. 2 shows that GE-Kriging and GE-Cokriging are more time-efficient compared to Cokriging, which is suggested by the fact that the prediction of gradients with GE-Kriging and GE-Cokriging take a rather small amount of time compared to Cokriging method. This is due to the fact that GE-Kriging and GE-Cokriging integrate both QoI data and the corresponding gradient data in the training step and hence provides prediction of QoI as well as the gradient on the new locations simultaneously in the predicting step. Whereas, Cokriging requires construction of a model for gradient data separately. Hence, the time for the prediction of the gradients by GE-Kriging and GE-Cokriging, i.e., the last two columns in Tab. 2, are for prediction only and is relatively short. It is also noticed that the time consumption of GE-Kriging

Case	Cokriging	GE-Kriging	GE-Cokriging	Cokriging (∇)	GE-Kriging (∇)	GE-Cokriging (∇)
1D1	$0.1146 \pm 1.49e\text{-}2$	$0.7534 \pm 5.17e\text{-}6$	$0.0138 \pm 9.98e\text{-}5$	$0.9565 \pm 1.84e\text{-}2$	$0.5985 \pm 3.85e\text{-}6$	$0.0221 \pm 1.49e\text{-}4$
1D2	$0.5325 \pm 1.47e\text{-}5$	$0.7534 \pm 6.30e\text{-}6$	$0.1254 \pm 4.18e\text{-}5$	$0.8964 \pm 2.97e\text{-}5$	$0.5986 \pm 4.83e\text{-}6$	$0.0973 \pm 3.13e\text{-}6$
2D*	$0.3152 \pm 2.21e\text{-}1$	$0.2471 \pm 1.60e\text{-}1$	$0.0292 \pm 8.86e\text{-}3$	$0.3101 \pm 1.88e\text{-}1$	$0.4062 \pm 2.03e\text{-}1$	$0.0798 \pm 3.26e\text{-}2$
2D**	-	-	-	$0.1341 \pm 1.78e\text{-}2$	$0.3342 \pm 1.39e\text{-}1$	$0.0114 \pm 5.81e\text{-}3$
Oscillator	$0.9224 \pm 2.15e\text{-}2$	$0.1259 \pm 3.27e\text{-}6$	$0.0926 \pm 2.40e\text{-}5$	$1.0771 \pm 3.92e\text{-}2$	$0.2639 \pm 1.78e\text{-}1$	$0.0993 \pm 2.49e\text{-}5$
Power	$0.2451 \pm 4.79e\text{-}4$	$0.1888 \pm 1.57e\text{-}5$	$0.0363 \pm 4.87e\text{-}6$	$0.7391 \pm 1.08e\text{-}2$	$0.2413 \pm 2.82e\text{-}5$	$0.0522 \pm 7.23e\text{-}6$

Table 1: Relative mean squared error (mean±standard deviation) of QoI and the corresponding gradients for each numerical example averaged over 5 separate runs with random parameters initialization by Cokriging, GE-Kriging and GE-Cokriging. $*$ denotes gradient in x direction and $**$ denotes gradient in y direction. ∇ denotes prediction of the gradient of QoI.

is smaller than that of GE-Cokriging, recall that GE-Kriging only used high-fidelity information while GE-Cokriging used both high-fidelity and low-fidelity information, which lead to a larger covariance matrix in GE-Cokriging compared to that in GE-Kriging. Although GE-Cokriging generally requires longer time in the training step, almost doubles Cokriging's training time, the total time cost of GE-Cokriging in QoI and gradients prediction is almost the same as that of Cokriging. Considering the significant improvement in accuracy and robustness, we can conclude that GE-Cokriging is an accurate and efficient approach to obtain prediction both QoI and its gradients simultaneously.

Case ID	Cokriging	GE-Kriging	GE-Cokriging	Cokriging (∇)	GE-Kriging (∇)	GE-Cokriging (∇)
1D1	$1.5702 \pm 1.33e\text{-}2$	$0.4193 \pm 4.196e\text{-}2$	$2.0945 \pm 1.28e\text{-}1$	$1.2128 \pm 1.42e\text{-}1$	$0.0042 \pm 4.42e\text{-}4$	$0.0156 \pm 1.56e\text{-}3$
1D2	$0.8337 \pm 1.06e\text{-}1$	$0.4452 \pm 3.06e\text{-}2$	$1.0767 \pm 8.82e\text{-}2$	$0.8329 \pm 9.54e\text{-}2$	$0.0036 \pm 6.76e\text{-}4$	$0.0142 \pm 1.18e\text{-}3$
2D*	$2.0945 \pm 7.08e\text{-}1$	$0.7623 \pm 5.26e\text{-}2$	$3.5935 \pm 7.42e\text{-}1$	$1.3417 \pm 2.96e\text{-}1$	$0.1245 \pm 2.93e\text{-}2$	$0.7502 \pm 1.72e\text{-}2$
2D**	-	-	-	$1.6846 \pm 3.71e\text{-}2$	-	-
Oscillator	$0.6402 \pm 3.75e\text{-}2$	$0.3321 \pm 1.84e\text{-}1$	$1.1046 \pm 9.08e\text{-}2$	$0.6926 \pm 3.34e\text{-}2$	$0.0095 \pm 8.79e\text{-}4$	$0.0074 \pm 1.96e\text{-}3$
Power	$0.4127 \pm 1.15e\text{-}2$	$0.4933 \pm 8.01e\text{-}2$	$2.4326 \pm 1.34e\text{-}1$	$0.9492 \pm 3.21e\text{-}3$	$0.0119 \pm 5.62e\text{-}3$	$0.0198 \pm 2.56e\text{-}3$

Table 2: Runtime (mean±standard deviation) of predicting QoI and its gradients for each numerical example averaged over 5 separate runs with random parameters initialization by Cokriging, GE-Kriging and GE-Cokriging. $*$ denotes gradient in x direction and $**$ denotes gradient in y direction. ∇ denotes prediction of the gradient of QoI.

4 Conclusion

In this work, we present a comprehensive gradient-enhanced multi-fidelity Cokriging method, namely GE-Cokriging, which incorporates available gradient information of multi-fidelity data, i.e., low-fidelity and high-fidelity observation of QoIs and its gradients. We present several numerical examples to study the performance of GE-Cokriging. Our results show that GE-Cokriging can accurately predict the QoI and its gradients simultaneously. We compare the performance of GE-Cokriging against GE-Kriging and multi-fidelity Cokriging, two popular GP-based prediction methods, and illustrate that GE-Cokriging is the most accurate, robust and efficient among these methods.

In particular, our result suggests that GE-Cokriging achieves better accuracy than GE-Kriging, this is because it exploits the information of the low-fidelity model. Also, GE-Cokriging yields more accurate results than using Cokriging for QoI and its gradients separately, because it takes advantage of the relation between these two quantities and makes use of corresponding data jointly. Even when some of the low-fidelity gradient information is misleading, for example, the gradient of low-fidelity data is negative while that of high-fidelity data is positive, the GE-Cokriging method may still be robust enough to predict accurately on target functions with less uncertainty compared to those by Cokriging and GE-Kriging. Moreover, the GE-Cokriging helps to alleviate the singularity issue of the covariance matrix, which is quite common in GPR methods. In terms of computational cost, the training of GE-Cokriging model, i.e., identifying

hyperparameters, could take longer time than Cokriging in solving a high-dimensional problem, given that the dimension of the covariance matrix is expanded due to the incorporation of gradient samples. However, once these hyperparameters are specified, the QoI and its gradients can be predicted simultaneously. This saves total computational time compared with Cokriging, which requires constructing models for QoI and its gradients separately, and hence needs training at least two models. Therefore, the overhead of training a model with a larger covariance matrix in GE-Cokriging is mitigated, and the overall time required to predict both QoI and its gradients for these three methods are comparable.

We note that our gradient-enhanced framework is also flexible for further extensions. In all of the numerical examples, we apply the commonly used stationary radial-basis function kernel. Other kernel functions, e.g., Matérn kernels with different smoothness, can be used to solve problems with desired regularity constraints. In addition, non-stationary kernels can be applied in this framework to model heterogeneous systems more accurately. Another extension can be to relax the constraints on the sample data to address the situation of missing data. More specifically, in the numerical examples presented, the gradient information is available with QoI at each observation location. Whereas in practice, it is possible that at some observation locations, either the QoI or its gradient is unavailable. In this scenario, modifications to the mean and covariance functions of the GP in our framework are needed. Moreover, we used the linear auto-regression form of the multi-fidelity Cokriging from [13], which can be replaced by more general nonlinear auto-regression forms, e.g., the methods used in [23, 9, 17], or even the deep neural network, e.g., [19]. Finally, as we point out in Section 2.4, our framework can also be built based on the "integral-enhanced" perspective, which can be useful in specific practical problems.

Acknowledgments

Yixiang Deng was supported by National Science Foundation (NSF) Award No. 1736088. Xiu Yang was supported by the U.S. Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research (ASCR) as part of Multifaceted Mathematics for Rare, Extreme Events in Complex Energy and Environment Systems (MACSER). Guang Lin gratefully acknowledges the support from National Science Foundation (DMS-1555072, DMS-1736364, and CMMI-1634832) and Brookhaven National Laboratory Subcontract 382247.

References

[1] Petter Abrahamsen. A review of gaussian random fields and correlation functions, 1997.

[2] Giancarlo Alfonsi. Reynolds-averaged navier–stokes equations for turbulence modeling. *Appl. Mech. Rev.*, 62(4), 2009.

[3] Hyoung Seog Chung and Juan Alonso. Design of a low-boom supersonic business jet using cokriging approximation models. In *9th AIAA/ISSMO symposium on multidisciplinary analysis and optimization*, page 5598, 2002.

[4] Richard Dwight and Zhong-Hua Han. Efficient uncertainty quantification using gradient-enhanced kriging. In *50th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference 17th AIAA/ASME/AHS Adaptive Structures Conference 11th AIAA No*, page 2276, 2009.

[5] Pep Espanol and Patrick Warren. Statistical mechanics of dissipative particle dynamics. *Europhys. Lett.*, 30(4):191, 1995.

[6] Alexander Forrester, Andy Keane, and Andràs Sòbester. *Engineering Design via Surrogate Modelling: A Practical Guide*. John Wiley & Sons, 2008.

[7] Alexander IJ Forrester, András Sóbester, and Andy J Keane. Multi-fidelity optimization via surrogate modelling. *Proc. R. Soc. A.*, 463(2088):3251–3269, 2007.

[8] Meixia Geng, Danian Huang, Qingjie Yang, and Yinping Liu. 3d inversion of airborne gravity-gradiometry data using cokriging. *Geophysics*, 79(4):G37–G47, 2014.

[9] Mark Girolami and Mingjun Zhong. Data integration for classification problems employing gaussian process priors. In *Adv. Neural. Inf. Process. Syst.*, pages 465–472, 2007.

[10] Pierre Goovaerts. Ordinary cokriging revisited. *Math. Geosci.*, 30(1):21–42, 1998.

[11] Loic Le Gratiet and Josselin Garnier. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *Int. J. Uncertain. Quan.*, 4(5):365–386, 2014.

[12] Zhong-Hua Han, Stefan Görtz, and Ralf Zimmermann. Improving variable-fidelity surrogate modeling via gradient-enhanced kriging and a generalized hybrid bridge function. *Aerosp. Sci. Technol.*, 25(1):177–189, 2013.

[13] Marc C Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

[14] Peter K Kitanidis. *Introduction to Geostatistics: Applications in Hydrogeology*. Cambridge University Press, 1997.

[15] J Laurenceau, M Meaux, M Montagnac, and P Sagaut. Comparison of gradient-based and gradient-enhanced response-surface-based optimizers. *AIAA J.*, 48(5):981–994, 2010.

[16] Luc Laurent, Rodolphe Le Riche, Bruno Soulier, and Pierre-Alain Boucard. An overview of gradient-enhanced metamodels with applications. *Arch. Comput. Methods Eng.*, 26(1):61–106, 2019.

[17] Seungjoon Lee, Felix Dietrich, George E Karniadakis, and Ioannis G Kevrekidis. Linking gaussian process regression with data-driven manifold embeddings for nonlinear data fusion. *Interface focus*, 9(3):20180083, 2019.

[18] Seungjoon Lee, Ioannis G Kevrekidis, and George Em Karniadakis. A general cfd framework for fault-resilient simulations based on multi-resolution information fusion. *J. Comput. Phys.*, 347:290–304, 2017.

[19] Xuhui Meng and George Em Karniadakis. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse pde problems. *J. Comput. Phys.*, 401:109020, 2020.

[20] Max D Morris, Toby J Mitchell, and Donald Ylvisaker. Bayesian design and analysis of computer experiments: use of derivatives in surface prediction. *Technometrics*, 35(3):243–255, 1993.

[21] Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Rev.*, 60(3):550–591, 2018.

[22] P Perdikaris, D Venturi, JO Royset, and GE Karniadakis. Multi-fidelity modelling via recursive co-kriging and Gaussian–Markov random fields. *Proc. R. Soc. A.*, 471(2179):20150018, 2015.

[23] Paris Perdikaris, Maziar Raissi, Andreas Damianou, ND Lawrence, and George Em Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proc. R. Soc. A*, 473(2198):20160751, 2017.

[24] Ghanshyam Pilania, James E Gubernatis, and Turab Lookman. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.*, 129:156–163, 2017.

[25] Osborne Reynolds. Iv. on the dynamical theory of incompressible viscous fluids and the determination of the criterion. *Philos. Trans. R. Soc. Lond. A*, (186):123–164, 1895.

[26] Christie Richard. Power systems test case archive, May 1993.

[27] Robert E Rudd and Jeremy Q Broughton. Coarse-grained molecular dynamics and the atomic limit of finite elements. *Phys. Rev. B*, 58(10):R5893, 1998.

[28] A Stein and LCA Corsten. Universal kriging and cokriging as a regression procedure. *Biometrics*, pages 575–587, 1991.

[29] A Stein, IG Staritsky, J Bouma, AC Van Eijnsbergen, and AK Bregt. Simulation of moisture deficits and areal interpolation by universal cokriging. *Water Resour. Res.*, 27(8):1963–1973, 1991.

[30] Selvakumar Ulaganathan, Ivo Couckuyt, Francesco Ferranti, Eric Laermans, and Tom Dhaene. Performance study of multi-fidelity gradient enhanced kriging. *Struct. Multidiscipl. Optim.*, 51(5):1017–1033, 2015.

[31] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[32] Ying Xuan, JunHua Xiang, WeiHua Zhang, and YuLin Zhang. Gradient-based kriging approximate model and its application research to optimization design. *Sci. China Technol. Sci.*, 52(4):1117–1124, 2009.

[33] Xiu Yang, David Barajas-Solano, Guzel Tartakovsky, and Alexandre M Tartakovsky. Physics-informed cokriging: A gaussian-process-regression-based multifidelity method for data-model convergence. *J. Comput. Phys.*, 395:410–431, 2019.

[34] Xiu Yang, Guzel Tartakovsky, and Alexandre Tartakovsky. Physics-informed kriging: A physics-informed gaussian process regression method for data-model convergence. *arXiv preprint arXiv:1809.03461*, 2018.

[35] Xiu Yang, Xueyu Zhu, and Jing Li. When bifidelity meets cokriging: An efficient physics-informed multifidelity method. *SIAM J. Sci. Comput.*, 42(1):A220–A249, 2020.

[36] Ray Daniel Zimmerman, Carlos Edmundo Murillo-Sánchez, and Robert John Thomas. Matpower: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Trans. Power Syst.*, 26(1):12–19, 2011.

[37] Ralf Zimmermann. On the maximum likelihood training of gradient-enhanced spatial gaussian processes. *SIAM J. Sci. Comput.*, 35(6):A2554–A2574, 2013.

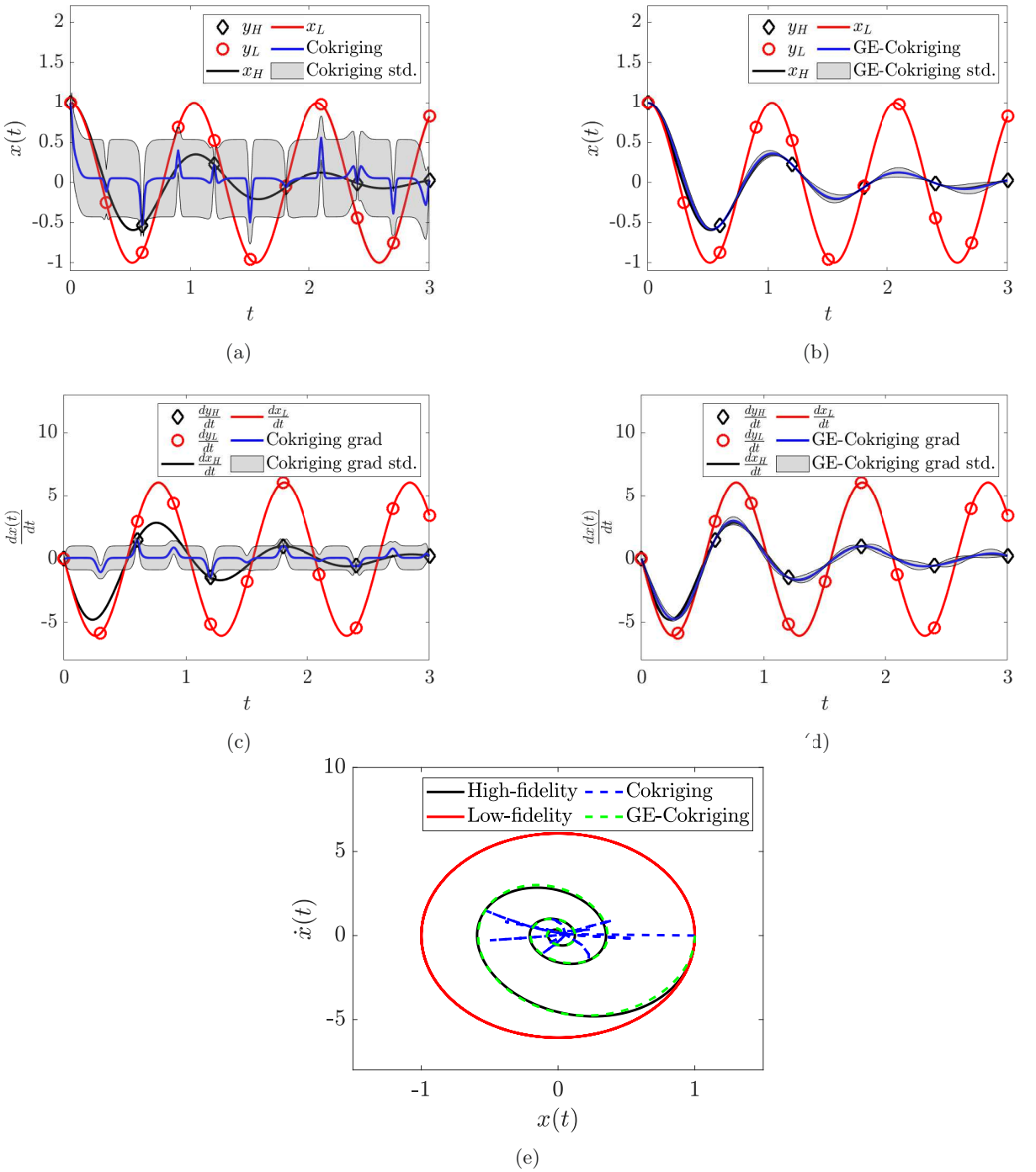

Figure 7: Prediction of the trajectory (QoI), velocity (gradient of QoI) and the phase diagram of an underdamped oscillator. Prediction of the posterior mean (blue solid lines) and standard deviation (grey shaded area) of the trajectory $x_H(t)$ by (a) Cokriging and (b) GE-Cokriging. Prediction of the posterior mean (blue solid lines) and standard deviation (grey shaded area) of the velocity $\frac{dx_H(t)}{dt}$ by (c) Cokriging and (d) GE-Cokriging. (e) Prediction of phase diagram by Cokriging (blue dashed line) and that by GE-Cokriging (black dashed line). Black diamonds denote high-fidelity observations, red circles denote low-fidelity observations, black solid lines denote the high-fidelity models and red solid lines denote low-fidelity models. Colored online.

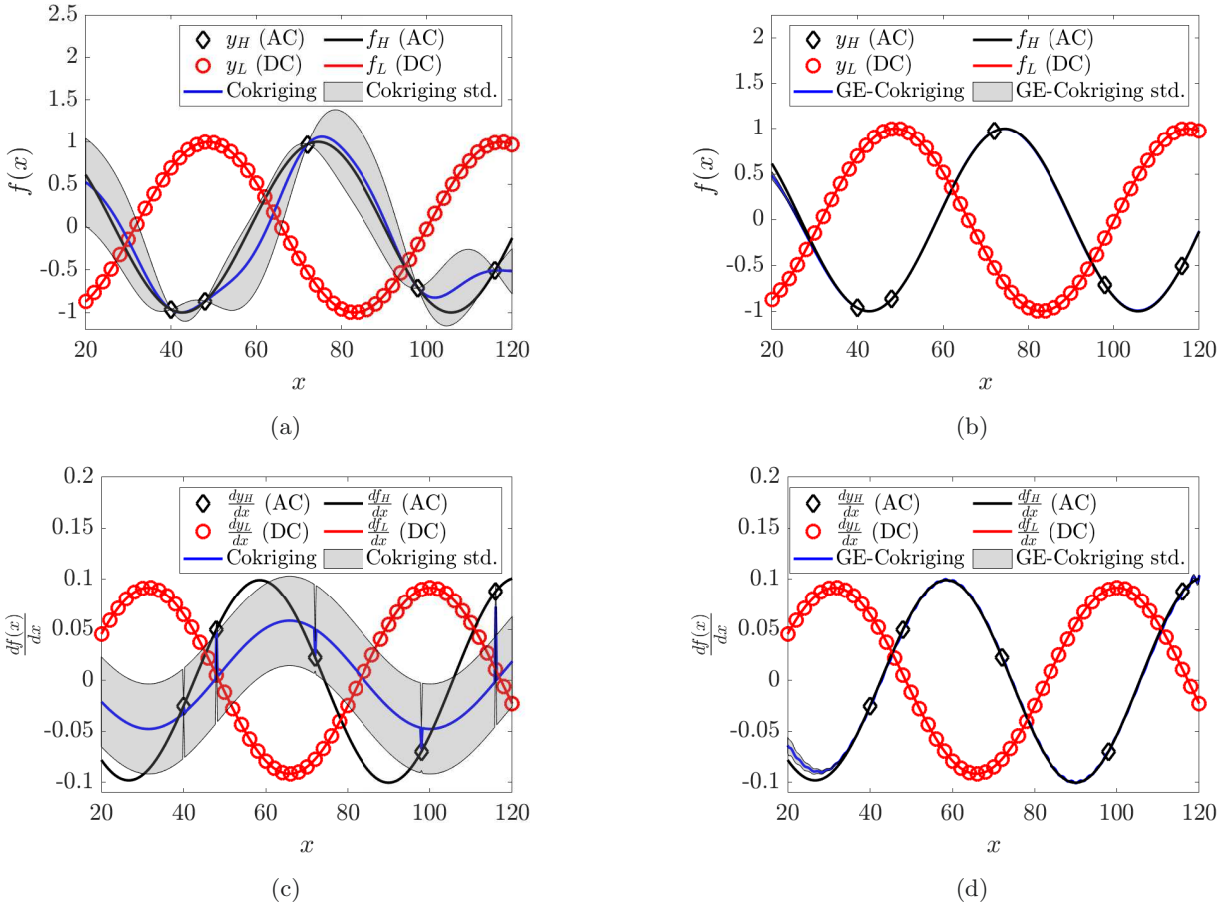

Figure 8: Prediction of the relationship between the power input of a generator bus x and real-time power factor of a load bus $f_H(x)$ by an AC model. Prediction of the posterior mean (blue solid lines) and standard deviation (grey shaded area) of $f_H(x)$ by (a) Cokriging and (b) GE-Cokriging. Prediction of the posterior mean (blue solid lines) and standard deviation (grey shaded area) of gradient of QoI $\frac{df_H(x)}{dx}$ by (c) Cokriging and (d) GE-Cokriging. Black diamonds denote high-fidelity observations, red circles denote low-fidelity observations, black solid lines denote the high-fidelity models and red solid lines denote the low-fidelity models. Colored online.