Computers and Chemical Engineering 106 (2017) 171–182

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

An adaptive sampling approach for Kriging metamodeling by maximizing expected prediction error

Haitao Liu [a,*], Jianfei Cai [b], Yew-Soon Ong [b,c]

[a] Rolls-Royce@NTU Corporate Lab, Nanyang Technological University, Singapore 637460, Singapore
[b] School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore
[c] Data Science and Artificial Intelligence Research Center, Nanyang Technological University, 639798, Singapore

ARTICLE INFO

Article history:
Received 27 March 2017
Received in revised form 18 May 2017
Accepted 24 May 2017
Available online 8 June 2017

Keywords:
Adaptive sampling
Kriging metamodeling
Expected prediction error
Adaptive balance strategy

ABSTRACT

As a well-known approximation method, Kriging is widely used in process engineering design and optimization for saving computational budget. The Kriging model for a target function is fitted to a set of sample points, the responses of which are expensive to obtain in practice and the sample distribution of which has a great impact on the model prediction quality. Therefore, a main task in adaptive sampling for Kriging metamodeling is to gather informative points in order to build an accurate model with as few points as possible. To this end, we propose an adaptive sampling approach under the bias-variance decomposition framework. This novel sampling approach sequentially selects new points by maximizing an expected prediction error criterion that considers both the bias and variance information. Particularly, it presents an adaptive balance strategy to dynamically balance the local exploitation and global exploration via the error information from the previous iteration. Four benchmark cases and four engineering cases from low to high dimensions are used to assess the performance of the proposed approach. Numerical results reveal that this adaptive sampling approach is very promising for constructing accurate Kriging models for problems with diverse characteristics.

1. Introduction

As a statistical model, the Kriging metamodeling technique, also known as Gaussian process regression (GPR), has been extensively used in engineering design and optimization for relieving computational budget. For an expensive simulation-based function, Kriging can fit a mathematical model to a finite number of observed points as an approximation. The cheap-to-run Kriging model helps enhance the understanding of the target function by exploring the design space. Some variants of Kriging, e.g., co-Kriging (Kennedy and O'Hagan, 2000), blind-Kriging (Joseph et al., 2008), gradient-enhanced Kriging (Morris et al., 1993), and non-stationary Kriging (Xiong et al., 2007), have been developed for different purposes.

Given an expensive target function $f \in R^1$, the general Kriging metamodeling process consists of two parts: (1) generating a set of observed points by the design of experiments (DoE) techniques; and (2) fitting a Kriging model $\hat{f}$ to the observed data. It is found that the sample positions distinctly affect the prediction quality of Kriging. Considering the limited computational budget in practice, a key issue in Kriging metamodeling is how to gather informative points in order to build an accurate model with as few points as possible.

Suppose that we already have a set of initial points $\mathbf{X}_D$ in the domain $D \in R^n$, and maintain a large pool of candidate points $\mathbf{U} = D \setminus \mathbf{X}_D$. The main task of a sampling approach is to sequentially select informative points from $\mathbf{U}$ and evaluate their responses in order to efficiently refine the Kriging model. Recently, the adaptive sampling strategy, also known as active learning (Settles, 2010), has gained increasing attention. This sampling strategy sequentially selects new points based on the information of both the approximation model and the data itself from previous iterations. Recently, there have emerged various adaptive sampling strategies for global metamodeling (Mackman and Allen, 2010; Xu et al., 2014; Eason and Cremaschi, 2014; Garud et al., 2017; Wang and Ierapetritou, 2017). This article mainly focuses on the adaptive sampling developed under the Bayesian framework for Kriging metamodeling.

It is known that because of the Bayesian framework, the Kriging model provides not only the prediction response $\hat{f}(\boldsymbol{x})$ but also the prediction variance $s^2(\boldsymbol{x})$ (also known as mean square error, MSE) at an arbitrary point $\boldsymbol{x}$. Therefore, a straightforward adaptive sampling strategy for Kriging is to sequentially select a new point with

* Corresponding author.
E-mail addresses: htliu@ntu.edu.sg (H. Liu), ASJFCai@ntu.edu.sg (J. Cai), ASYSOng@ntu.edu.sg (Y.-S. Ong).

http://dx.doi.org/10.1016/j.compchemeng.2017.05.025
0098-1354/

maximal prediction variance, denoted as maximum mean square error (MMSE) (Jin et al., 2002), in order to reduce the generalization error of $\hat{f}$. The MMSE sampling approach is widely used due to the ease of implementation and the cost-efficient process. Based on Shannon's entropy theory, Shewry and Wynn (1987) proposed the maximum entropy (ME) criterion to select new points by maximizing the determinant of the correlation matrix. Particularly, the one-by-one ME approach is equivalent to the MMSE approach (Jin et al., 2002). To further improve the sampling performance, Morris et al. (1993) combined the maximum entropy sampling approach with the known first derivatives. Besides, under the expected error reduction framework, Sacks et al. (1989) selected new points by maximizing the expected reduction in integrated mean square error (IMSE) over the entire design space, which is equivalent to the active learning-Cohn (ALC) criterion (Cohn, 1996; Christen et al., 2011). However, due to the integral operation, the computational complexity of IMSE is much higher than that of MSE. A recent work (Beck and Guillas, 2016) suggested using the mutual information to adaptively select new points by maximizing the expected information gain.

The points generated by the above variance-based sampling approaches are found to primarily fill the domain evenly and have a slight adaption to the variability along each coordinate direction. This is because that they consider only the Kriging variance information that follows a stationary assumption wherein the correlation function is identical over the entire domain. As a result, the prediction variance solely depends on the sample locations. For effectively improving the model accuracy within limited computational budget, rather than fill the domain evenly, an interesting idea is to determine the sample positions according to the characteristics of target function, e.g., sampling more points in regions with large prediction errors.

To this end, Lin et al. (2004) used the prediction errors estimated by additional validation points to adjust the correlation matrix so that the correlation function is no longer identical over the domain. That is, it owns the ability to adapt to the function characteristics, which thereafter helps the sampling approach identify regions with large prediction errors. The validation points, however, are usually unavailable in practice. In a similar spirit, Farhang-Mehr and Azarm (2005) employed the locations of local optima on current Kriging model to adjust the correlation matrix in order to identify irregular regions. This adaptive sampling criterion heavily depends on the quality of the Kriging model. A poor model may guide an erroneous sampling direction. Recently, Liu et al. (2016a) proposed an adaptive maximum entropy (AME) sampling approach by using the cross-validation errors to adjust the correlation function, and moreover, employing a user-defined search pattern to circularly conduct sampling from global to local. Busby (2009) and Busby et al. (2007) decomposed the domain into cells with the edges being of the order of correlation lengths along different directions. Then, they adopted the cross-validation criterion and ME criterion to identify "bad cells" for sampling. Besides, Lam (2008) modified the expected improvement criterion developed for global optimization (Jones et al., 1998) to obtain a good global model fit. The expected improvement for global fit (EIGF) approach selects informative points that have a large expected improvement over the nearest observed points.

As has been pointed out by Liu et al. (2016a) and Deschrijver et al. (2011), an effective adaptive sampling approach should contain three parts:

(1) *Local exploitation*. This part accounts for the adaption of the sampling process by guiding the sampling in regions with large prediction errors. The local exploitation term can be represented in various ways, e.g., the prediction errors at validation points (Lin et al., 2004), the cross-validation errors (Jin et al., 2002; Liu et al., 2016a; Busby et al., 2007), the locations of local optima (Farhang-Mehr and Azarm, 2005), the local response variation (Lam, 2008) and more others.
(2) *Global exploration*. This part ensures the purpose of global metamodeling, and avoids the missing of undetected interesting regions by employing, e.g., some distance-based criteria (Jin et al., 2002).
(3) *Trade-off between local & global*. Last but not least, this part balances the local exploitation and global exploration, and has a great impact on the sampling performance. Most of current adaptive sampling approaches for Kriging, however, usually use a fixed balance rule, which is non-beneficial for sampling performance. Liu et al. (2016a) introduced a user-defined search pattern to balance the local exploitation and global exploration, but it is still inflexible. For NURBs-based metamodeling, Turner et al. (2007) proposed using two pure global criteria and two pure local criteria to formulate a cooling schedule wherein the Bernstein basis functions are adopted to decide the dominant criterion in different sampling stages. Singh et al. (2013) illustrated three conceptual balance strategies and recommended the adaptive balance strategy.

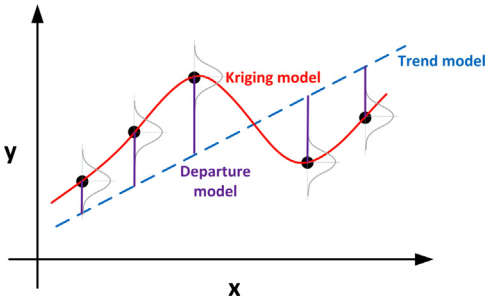

Fig. 1. Illustration of a Kriging model with linear regressions.

This article derives an adaptive sampling approach for Kriging metamodeling under the bias-variance decomposition framework. The proposed approach sequentially selects the most informative points through maximizing the expected prediction error criterion that considers both the bias and variance information. Besides, it presents a novel adaptive balance strategy to gain benefits from effective local exploitation, while not hurting the performance through dynamically balancing the local exploitation and global exploration via the error information from the previous iteration.

The remaining of the article is organized as follows. Section 2 gives a brief introduction of the Kriging model. Section 3 describes the proposed adaptive sampling approach via maximizing expected prediction error. Thereafter, four benchmark cases and four engineering cases from low to high dimensions are employed in Section 4 to assess the sampling performance. Finally, Section 5 offers some concluding remarks.

2. Kriging model

Kriging was first introduced in the field of geology to estimate the properties of sampled minerals given a set of sampled sites (Journel and Huijbregts, 1978). Thereafter, Sacks et al. (1989) applied Kriging in the context of design and analysis of computer experiments (DACE), the meaning of which now has been extended to refer to the suite of all metamodeling techniques (Viana et al., 2014).

In the Kriging framework, the main assumption is that the deterministic output $f(\boldsymbol{x})$ is regarded as the realization of a stochastic process $y_K(\boldsymbol{x})$. As shown in Fig. 1, Kriging is composed of a global polynomial model, called *trend model*, over the entire domain and

a local spatial deviation model, called *departure model*. The trend model approximates the output response globally, while the departure model pulls the predicted responses through the observed points by accounting for the correlation in the residuals between the trend model and the observed responses.

The Kriging model can be expressed as

$$y_K(\boldsymbol{x}) = \boldsymbol{p}^T(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{x} \in R^n$ is a point with n input variables; the $k \times 1$ vector $\boldsymbol{p}(\boldsymbol{x}) = [p_1(\boldsymbol{x}), \ldots, p_k(\boldsymbol{x})]^T$ represents the regression functions with simple formulas such as 1, x, x^2, …; the $k \times 1$ vector $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_k]^T$ is the corresponding regression coefficients; $Z(\boldsymbol{x})$ is usually assumed to be a Gaussian random process with the mean and covariance as

$$\mathbb{E}(Z(\boldsymbol{x})) = 0 \tag{2}$$

and

$$Cov(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)) = \sigma^2 R(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{3}$$

where σ^2 is the process variance of $Z(\cdot)$, and $R(\cdot)$ is a spatial correlation function, which controls the smoothness of the Kriging model, and is usually expressed in a Gaussian form as

$$R(\boldsymbol{x}_i, \boldsymbol{x}_j) = \prod_{s=1}^{n} \exp\left(-\theta_s d_s^2\right) \tag{4}$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n]^T$ is the unknown correlation parameters, and $d_s = |x_i^s - x_j^s|$ is the distance between the sth components of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Other commonly used correlation functions can be found in Lophaven et al. (2002).

Suppose that we have m observed points $\mathbf{X}_D = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]^T$ and their output responses $\mathbf{y}_D = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_m)]^T$. To fit a Kriging model using $\mathbf{X}_D$ and $\mathbf{y}_D$, we first employ the maximum likelihood estimation (MLE) to estimate the regression coefficients $\boldsymbol{\beta}$, the process variance σ^2, and the correlation parameters $\boldsymbol{\theta}$. Given the model parameters $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$, the logarithm of the multivariate Gaussian likelihood function is expressed as

$$L[\boldsymbol{\gamma}|\mathbf{y}_D] = -\frac{m}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\ln(|\boldsymbol{R}|) - \frac{1}{2\sigma^2}(\mathbf{y}_D - \boldsymbol{F}\boldsymbol{\beta})^T\boldsymbol{R}^{-1}(\mathbf{y}_D - \boldsymbol{F}\boldsymbol{\beta}) \tag{5}$$

where $\boldsymbol{R}$ is a $m \times m$ correlation matrix with the element $R_{ij} = R(\boldsymbol{x}_i, \boldsymbol{x}_j)$; $\boldsymbol{F} = [\boldsymbol{p}(\boldsymbol{x}_1), \ldots, \boldsymbol{p}(\boldsymbol{x}_m)]^T$ is a $m \times k$ regression matrix with the element $F_{ij} = p_j(\boldsymbol{x}_i)$. By setting the gradients of the log-likelihood equation with respect to $\boldsymbol{\beta}$ and σ^2 to zeros, we obtain their estimations as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{R}^{-1}\mathbf{y}_D \tag{6}$$

and

$$\hat{\sigma}^2 = \frac{1}{m}(\mathbf{y}_D - \boldsymbol{F}\hat{\boldsymbol{\beta}})^T\boldsymbol{R}^{-1}(\mathbf{y}_D - \boldsymbol{F}\hat{\boldsymbol{\beta}}) \tag{7}$$

Besides, the estimation $\hat{\boldsymbol{\theta}}$ of the optimal correlation parameters can be obtained by solving the following auxiliary optimization problem as

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{argmax}\ \psi(\boldsymbol{\theta}) = -\frac{m}{2}\ln(\hat{\sigma}^2) - \frac{1}{2}\ln(|\boldsymbol{R}|) \tag{8}$$

Finally, after obtaining the estimated model parameters, the Kriging prediction $\hat{f}(\boldsymbol{x})$ at point $\boldsymbol{x}$ is the mean of the posterior Gaussian distribution as

$$\hat{f}(\boldsymbol{x}) = \mathbb{E}[y_K(\boldsymbol{x})|\mathbf{X}_D, \mathbf{y}_D, \boldsymbol{\gamma}, \boldsymbol{x}] = \boldsymbol{p}^T(\boldsymbol{x})\hat{\boldsymbol{\beta}} + \boldsymbol{r}^T(\boldsymbol{x})\boldsymbol{R}^{-1}(\mathbf{y}_D - \boldsymbol{F}\hat{\boldsymbol{\beta}}) \tag{9}$$

where the $m \times 1$ vector $\boldsymbol{r}(\boldsymbol{x}) = [R(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, R(\boldsymbol{x}, \boldsymbol{x}_m)]^T$ represents the correlation between point $\boldsymbol{x}$ and the m observed points. In addition, the prediction variance $s^2(\boldsymbol{x})$ is

$$\begin{aligned} s^2(\boldsymbol{x}) &= Var[y_K(\boldsymbol{x})|\mathbf{X}_D, \mathbf{y}_D, \boldsymbol{\gamma}, \boldsymbol{x}] \\ &= \hat{\sigma}^2[1 - \boldsymbol{r}^T(\boldsymbol{x})\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x}) \\ &\quad + (\boldsymbol{F}\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{p}(\boldsymbol{x}))^T(\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}(\boldsymbol{F}\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{p}(\boldsymbol{x}))] \end{aligned} \tag{10}$$

It is notable that Kriging is an interpolation model. This can be verified by substituting the ith observed point $\boldsymbol{x}_i$ into Eq. (9). In this context, $r(\boldsymbol{x}_i)$ is the ith column of $\boldsymbol{R}$, and consequently, $r^T(\boldsymbol{x}_i)\boldsymbol{R}^{-1}$ becomes a vector $\boldsymbol{I}_i$ with a 1 at the ith position. Then, we can see

$$\begin{aligned} \hat{f}(\boldsymbol{x}_i) &= \boldsymbol{p}^T(\boldsymbol{x}_i)\hat{\boldsymbol{\beta}} + \boldsymbol{I}_i(\mathbf{y}_D - \boldsymbol{F}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{p}^T(\boldsymbol{x}_i)\hat{\boldsymbol{\beta}} + f(\boldsymbol{x}_i) - \boldsymbol{p}^T(\boldsymbol{x}_i)\hat{\boldsymbol{\beta}} = f(\boldsymbol{x}_i) \end{aligned} \tag{11}$$

For more information about the Kriging model, please refer to Martin (2009), Kleijnen (2009) and the documentation of the DACE Toolbox (Lophaven et al., 2002).

3. An adaptive sampling approach for Kriging metamodeling

3.1. Maximum expected prediction error

It is known that in the Kriging framework, the posterior distribution of the response at a point $\boldsymbol{x}$ follows $y_K(\boldsymbol{x}) \sim \mathcal{N}(\mu_{\boldsymbol{x}}, \sigma_{\boldsymbol{x}}^2)$ where $\mu_{\boldsymbol{x}} = \hat{f}(\boldsymbol{x})$ and $\sigma_{\boldsymbol{x}}^2 = s^2(\boldsymbol{x})$. Here, we consider a prediction error function (also known as loss function) $\mathcal{L}$ defined in a squared form as

$$\mathcal{L}(\boldsymbol{x}) = (y(\boldsymbol{x}) - y_K(\boldsymbol{x}))^2 \tag{12}$$

where $y(\boldsymbol{x})$ is the observed response that considers the true response with an additive noise. That is, $y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon$ where ϵ is a Gaussian process with a zero mean, which represents, e.g., the measurement error in physical experiments.

Then, the overall generalization error of the Kriging model can be obtained as

$$e = \int_{\boldsymbol{x} \in D} \mathbb{E}[\mathcal{L}(\boldsymbol{x})]\, d\boldsymbol{x} \tag{13}$$

where $\mathbb{E}[\mathcal{L}(\boldsymbol{x})]$ is the expectation of the prediction error, and can be decomposed as

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\boldsymbol{x})] &= (\mathbb{E}[y(\boldsymbol{x})] - \mathbb{E}[y_K(\boldsymbol{x})])^2 \\ &\quad + \mathbb{E}[(y_K(\boldsymbol{x}) - \mathbb{E}[y_K(\boldsymbol{x})])^2] \\ &\quad + \mathbb{E}[(y(\boldsymbol{x}) - \mathbb{E}[y(\boldsymbol{x})])^2] \end{aligned} \tag{14}$$

The first term on the right-hand side of this equation is the *bias* that represents the average difference between the predicted response and the observed response; the second term is the Kriging model's prediction *variance*; and the third term, which equals Var(ϵ), represents the intrinsic *noise* of the data, and thus is independent of the Kriging model. Due to the deterministic computer simulation codes considered here (Sacks et al., 1989), we ignore the noise term.

The bias-variance decomposition in Eq. (14) has been known in the field of machine learning for model selection over different datasets (Zhou et al., 2002). Here, we introduce it to explain why adaptive sampling that considers local exploitation is expected to outperform the pure variance-based sampling.

By ignoring the noise term, we reorganize Eq. (14) to obtain the *expected prediction error* (EPE) of a Kriging model at point $\boldsymbol{x}$ as

$$\mathrm{EPE}(\boldsymbol{x}) = \underbrace{\left(f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\right)^2}_{bias^2} + \underbrace{s^2(\boldsymbol{x})}_{variance} \tag{15}$$

It is found that in order to reduce the generalization error of a Kriging model in Eq. (13), we should reduce both the bias term and the variance term in Eq. (15). This insight gives a guideline for the selection of the next new point $\boldsymbol{x}_{new}$ in adaptive sampling. That is, $\boldsymbol{x}_{new}$ should be the one at which the Kriging model yields a large prediction error or a large prediction variance. In the context of adaptive sampling, the bias term, which favors points with large prediction errors for efficiently improving the Kriging predictions in regions of interest, acts as *local exploitation*; while the variance term, which prefers points with large prediction variances for improving the Kriging predictions in regions with large uncertainty, acts as *global exploration*. It is found that, due to the consideration of the additional bias term, the criterion in Eq. (15) is capable of gathering more informative points than the typical variance-based sampling criteria. Although adaptive sampling that considers local exploitation is flexible and powerful, in practice it only gains benefits with valid local exploitation and appropriate trade-off between local exploitation and global exploration.

It is found that the bias and variance at point $\boldsymbol{x}$ is required in order to select new points based on the EPE criterion (15). For $\boldsymbol{x} \in \mathbf{X}_D$, as an interpolation model, we have $\mathrm{EPE}(\boldsymbol{x}) = 0$. However, for $\boldsymbol{x} \in \mathbf{U}$, the bias term in Eq. (15), i.e., the prediction error $e_{\mathrm{true}} = |\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})|$, is undetermined because of the unknown true response $f(\boldsymbol{x})$. Therefore, we need to estimate the bias information in practice.

Alternatively, we employ the k-fold cross-validation (CV) approach to estimate the bias term. For the commonly used leave-one-out cross-validation approach ($k = 1$), it selects $\boldsymbol{x}_i$ as a validation point and uses the remaining $m-1$ points to refit a model $\hat{f}^{-i}$ for obtaining the CV error as

$$e^2_{\mathrm{CV}}(\boldsymbol{x}_i) = \left(f(\boldsymbol{x}_i) - \hat{f}^{-i}(\boldsymbol{x}_i)\right)^2 \tag{16}$$

For building $\hat{f}^{-i}$, we need to re-estimate the correlation parameters $\boldsymbol{\theta}$ by solving problem (8) after removing $\boldsymbol{x}_i$, which, however, is a time-consuming process, especially for cases with large sample sizes and high dimensions (Liu et al., 2017). To alleviate computational budget, the correlation parameters can be fixed as constants in the cross-validation process (Joseph et al., 2008; Queipo et al., 2005). In this context, the CV error can be calculated efficiently as (Sundararajan and Keerthi, 2001)

$$e^2_{CV}(\boldsymbol{x}_i) = \left(\frac{\left(\boldsymbol{R}^{-1}\right)_{i,:} \left(\mathrm{d} + \boldsymbol{H}_{:,i} \frac{\boldsymbol{d}_i}{1-\boldsymbol{H}_{ii}} \right)}{\left(\boldsymbol{R}^{-1}\right)_{ii}} \right)^2 \tag{17}$$

where the subscript i, : indicates the ith row of a matrix, :, i means the ith column, and ii indicates the ith element on the diagonal; $\boldsymbol{d}$ and $\boldsymbol{H}$ are respectively defined as

$$\boldsymbol{d} = \mathbf{y}_D - \boldsymbol{F}\hat{\boldsymbol{\beta}} \tag{18}$$

and

$$\boldsymbol{H} = \boldsymbol{F}(\boldsymbol{F}^T\boldsymbol{F})^{-1}\boldsymbol{F}^T \tag{19}$$

The fast Eq. (17) decreases the computational complexity of e^2_{CV} from $\mathcal{O}(m^4)$ to $\mathcal{O}(m^3)$, since it calculates the inverse of $\boldsymbol{R}$ only once.

Furthermore, in order to estimate the CV error at an arbitrary point $\boldsymbol{x}$, we consider the Voronoi partition of the domain D according to the observed sample set $\mathbf{X}_D$, and have the cell V_i as

$$V_i = \{\boldsymbol{x} \in D, |\boldsymbol{x} - \boldsymbol{x}_i| \le |\boldsymbol{x} - \boldsymbol{x}_j|, \forall j \ne i\}, \ 1 \le i, j \le m \tag{20}$$

Thereafter, the CV error $e^2_{\mathrm{CV}}(\boldsymbol{x})$ is simply estimated as

$$e^2_{\mathrm{CV}}(\boldsymbol{x}) = e^2_{\mathrm{CV}}(\boldsymbol{x}_i), \ \boldsymbol{x} \in V_i \tag{21}$$

Eq. (21) reveals that if two points fall into the same Voronoi cell, they have the same CV error.

Then, we replace the bias term in Eq. (15) with the cross-validation error as

$$\mathrm{EPE}_{\mathrm{CV}}(\boldsymbol{x}) = e^2_{\mathrm{CV}}(\boldsymbol{x}) + s^2(\boldsymbol{x}) \tag{22}$$

Now the $\mathrm{EPE}_{\mathrm{CV}}$ criterion consists of determined local exploitation and global exploration terms.

It is notable that the CV error term $e^2_{\mathrm{CV}}(\boldsymbol{x})$ is an estimation of the true error $e^2_{\mathrm{true}}(\boldsymbol{x})$. The quality of this estimation is found to affect the sampling performance. If the CV error under-/over-estimates the true error, which is a frequently encountered case in practice (Viana et al., 2009; Liu et al., 2016b), the $\mathrm{EPE}_{\mathrm{CV}}$ criterion will become biased due to the discrepancy between $e^2_{\mathrm{CV}}(\boldsymbol{x})$ and $e^2_{\mathrm{true}}(\boldsymbol{x})$. Therefore, the CV error term should be adjusted in order to make the sampling process unbiased. In other words, the $\mathrm{EPE}_{\mathrm{CV}}$ criterion needs a trade-off between local exploitation and global exploration, which has a great impact on the sampling performance.

To this end, we introduce a balance factor α to adjust the $\mathrm{EPE}_{\mathrm{CV}}$ criterion as

$$\mathrm{EPE}^{\alpha}_{\mathrm{CV}}(\boldsymbol{x}) = \alpha \times e^2_{\mathrm{CV}}(\boldsymbol{x}) + (1-\alpha) \times s^2(\boldsymbol{x}) \tag{23}$$

The factor α is derived from the idea of adjusting $e^2_{\mathrm{CV}}(\boldsymbol{x})$ so that it is as close to $e^2_{\mathrm{true}}(\boldsymbol{x})$ as possible. Meanwhile, this factor plays a role of balancing the local exploitation and global exploration. It is found that $\mathrm{EPE}^{\alpha}_{\mathrm{CV}}$ with a large α value prefers local exploitation, while a small α value prefers global exploration. Ideally, the adjust factor α should change with $\boldsymbol{x}$ because of the pointwise local and global terms. In practice, however, with little information about the estimation quality of e^2_{CV}, it is hard to present a pointwise version of α. Hence, here we use a constant α over the entire domain.

Let q denote the iteration number, in the $(q-1)$th iteration we obtain a new point $\boldsymbol{x}_{m+q-1}$ and its true response $f(\boldsymbol{x}_{m+q-1})$ via expensive simulator. Intuitively, the newest information at this point can act a feedback to help us determine the α value for the qth iteration. To this end, we present an adaptive balance strategy that dynamically adjusts the factor α by comparing the true prediction error and the CV error at $\boldsymbol{x}_{m+q-1}$.

The balance factor α is calculated as

$$\alpha = \begin{cases} 0.5, & q = 1 \\ 0.99 \times \min\left[0.5 \times \dfrac{e^2_{\mathrm{true}}(\boldsymbol{x}_{m+q-1})}{e^2_{\mathrm{CV}}(\boldsymbol{x}_{m+q-1})}, 1\right], & q > 1 \end{cases} \tag{24}$$

It is observed that in the first iteration ($q = 1$) α is initialized as 0.5, that is, the $\mathrm{EPE}^{\alpha}_{\mathrm{CV}}$ criterion treats local exploitation and global exploration equally. For $q > 1$, suppose that in the $(q-1)$th iteration we obtain a new point $\boldsymbol{x}_{m+q-1}$, and evaluate it through f and the Kriging model $\hat{f}_{q-1}$ constructed in this iteration to obtain the true error $e^2_{\mathrm{true}}(\boldsymbol{x}_{m+q-1})$ and the CV error $e^2_{\mathrm{CV}}(\boldsymbol{x}_{m+q-1})$, respectively. The balance factor α is adjusted as shown in Fig. 2.

- If $e^2_{\mathrm{CV}}(\boldsymbol{x}_{m+q-1}) > e^2_{\mathrm{true}}(\boldsymbol{x}_{m+q-1})$, it means the CV error overestimates the true error. Therefore, in the qth iteration the balance factor α is decreased from 0.5 to $0.5 \times \frac{e^2_{\mathrm{true}}(\boldsymbol{x}_{m+q-1})}{e^2_{\mathrm{CV}}(\boldsymbol{x}_{m+q-1})}$. In this context, $\mathrm{EPE}^{\alpha}_{\mathrm{CV}}$ favors global exploration.
- If $e^2_{\mathrm{CV}}(\boldsymbol{x}_{m+q-1}) = e^2_{\mathrm{true}}(\boldsymbol{x}_{m+q-1})$, it indicates that the CV error is the same as the true error at $\boldsymbol{x}_{m+q-1}$. Therefore, we have a balance factor of $\alpha = 0.5$, which results in equal treatment of local exploitation and global exploration.
- If $e^2_{\mathrm{CV}}(\boldsymbol{x}_{m+q-1}) < e^2_{\mathrm{true}}(\boldsymbol{x}_{m+q-1})$, it means the CV error underestimates the true error. We hence increase the α value

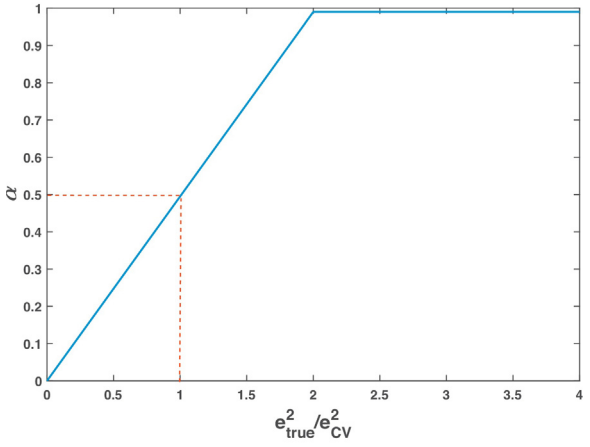

Fig. 2. The balance factor α when $q > 1$.

as $\min\left[0.5 \times \frac{e^2_{\text{true}}(\boldsymbol{x}_{m+q-1})}{e^2_{\text{CV}}(\boldsymbol{x}_{m+q-1})}, 1\right]$. As a result, $\text{EPE}^{\alpha}_{\text{CV}}$ favors local exploitation. Note that in this case, we use a min[., .] operator to prevent the α value from exceeding one when $0.5 \times \frac{e^2_{\text{true}}(\boldsymbol{x}_{m+q-1})}{e^2_{\text{CV}}(\boldsymbol{x}_{m+q-1})} > 1$.

- Finally, it is observed that Eq. (24) employs a coefficient of 0.99 to prevent the α value from being one. If $\alpha = 1$, then $\text{EPE}^{\alpha}_{\text{CV}}$ pays all attention on the local exploitation term $e^2_{\text{CV}}(\boldsymbol{x})$. As stated before, Eq. (21) implies that all the points in V_i produce the same CV error. In this context, maximizing the $\text{EPE}^{\alpha}_{\text{CV}}$ criterion with $\alpha = 1$ would result in an infinite number of new points. According to the above discussions, we have $\alpha \in [0, 1)$.

Note that the proposed adaptive balance strategy adjusts the balance factor α via the error information at a single point $\boldsymbol{x}_{m+q-1}$. We choose $\boldsymbol{x}_{m+q-1}$ because it is the only new data we can obtain in current stage. Meanwhile, this single point is believed to be representative since it is regarded as the point yielding the maximal expected prediction error in the previous iteration. Besides, Romero et al. (2015) used the present-stage error criterion that employs the error at current single point to estimate the true model error during the sequential Kriging refinement process. Test results revealed that this single point error is a good estimator of the true model error. Hence, the error information at $\boldsymbol{x}_{m+q-1}$ may estimate the true model error information well.

Finally, the proposed adaptive sampling approach selects the new point in the qth iteration by maximizing the $\text{EPE}^{\alpha}_{\text{CV}}$ criterion (MEPE) as

$$\boldsymbol{x}_{m+q} = \underset{\boldsymbol{x} \in \mathbf{U}}{argmax}\, \text{EPE}^{\alpha}_{\text{CV}}(\boldsymbol{x}) \tag{25}$$

Compared to typical variance-based sampling approaches that consider only the variance information to select new points, e.g., $\boldsymbol{x}_{new} = \text{argmax}s^2(\boldsymbol{x})$ in the MMSE sampling approach (Jin et al., 2002), the proposed MEPE sampling approach, however, considers not only the prediction variance but also the bias information to gather more informative new points.

Finally, the proposed MEPE adaptive sampling approach is elaborated in Algorithm 1. In line 1 the initial points should fill the whole domain evenly to ensure that everywhere has equal chance to be probed. As a result, the initial Kriging model can better represent the underlying function for guiding the subsequent adaptive sampling process. Some existing space-filling sampling approaches, e.g., the optimal Latin hypercube design (LHD) approaches (Kenny et al., 2000; Viana et al., 2010; Crombecq et al., 2011; Liu et al., 2015), could be used here. As for the stopping criterion in line 4, we often use the limited computational resource, e.g., the allowed maximal number of points (simulations).

Algorithm 1. The MEPE adaptive sampling approach

Input: An expensive black-box function f defined in the unit space $D \in [0, 1]^n$; an iteration number $q = 1$; and an initial balance factor $\alpha = 0.5$
Output: The final Kriging model $\hat{f}(\boldsymbol{x})$
1: Generate m initial points $\boldsymbol{X}_D = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m]^T$ by a space-filling sampling approach
2: Call f to evaluate their responses $\boldsymbol{y}_D = [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_m)]^T$
3: Construct the initial Kriging model as $\hat{f}(\boldsymbol{x}) \longleftarrow (\boldsymbol{X}_D, \boldsymbol{y}_D)$
4: **while** the stopping criterion is not met **do**
5: **Obtain a new point**:
6: Use Eq. (24) to update the balance factor α
7: Calculate the CV error $e^2_{\text{CV}}(\boldsymbol{x}_i)$ at each observed point using Eq. (17)
8: Use the local exploitation term $e^2_{\text{CV}}(\boldsymbol{x})$ estimated by Eq. (21), the global exploration term $s^2(\boldsymbol{x})$ estimated by Eq. (10), and the balance factor α to form the $\text{EPE}^{\alpha}_{\text{CV}}(\boldsymbol{x})$ criterion in Eq. (23)
9: Obtain the new point $\boldsymbol{x}_{m+q}$ by solving the auxiliary problem (25)
10: Call f to obtain the response of the new point as $f(\boldsymbol{x}_{m+q})$
11: **Update information**:
12: $\boldsymbol{X}_D = \boldsymbol{X}_D \cup \boldsymbol{x}_{m+q}$; $\boldsymbol{y}_D = \boldsymbol{y}_D \cup f(\boldsymbol{x}_{m+q})$; $q = q + 1$
13: Refit the Kriging model as $\hat{f}(\boldsymbol{x}) \longleftarrow (\boldsymbol{X}_D, \boldsymbol{y}_D)$
14: **return** The final Kriging model $\hat{f}(\boldsymbol{x})$

3.2. Discussions about MEPE

It is observed that there are two key parts in the proposed MEPE sampling approach. The first part is to use the estimated CV error for local exploitation, and the second part is to use the adaptive balance strategy for offering a dynamic trade-off between local exploitation and global exploration. Here, we attempt to validate the quality of the two parts on a 1D test case.

3.2.1. The quality of local exploitation

It is found that Lam (2008) proposed an *expected improvement for global fit* (EIGF) criterion derived from the well-known expected improvement (EI) criterion (Jones et al., 1998). The EIGF criterion measures the expected improvement of $\boldsymbol{x}$ over the nearest observed point $\boldsymbol{x}_*$ as

$$\text{EIGF}(\boldsymbol{x}) = \left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x}_*)\right)^2 + s^2(\boldsymbol{x}) \tag{26}$$

Though derived from different views, the $\text{EPE}^{\alpha}_{\text{CV}}$ criterion and the EIGF criterion have a similar formula. Regardless of the adaptive balance factor, the only difference between Eqs. (22) and (26) is that EIGF uses the nearest response variation $\left(\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x}_*)\right)^2$ to do local exploitation. The nearest response variation can be regarded as an representation of the local first-order derivative to some degree. That is, EIGF aims to sample more points in regions with drastic response changes. Compared to the CV error $e^2_{\text{CV}}(\boldsymbol{x})$ used in $\text{EPE}^{\alpha}_{\text{CV}}$, this nearest response variation, however, is not a direct estimation of the local prediction error. As a result, the EIGF criterion keeps highlighting the local exploitation in regions where the target function produces high first-order gradients, even if the metamodel has fit well in those regions. Hence, the local exploitation term in EIGF may be inappropriate in some circumstances.

To validate the quality of local exploitation, we employ a one-dimensional case as

$$f(x) = 3(1-x)^2 e^{(-x^2-1)} - 10\left(\frac{x}{5} - x^3\right)e^{-x^2}, \; x \in [-4, 1] \tag{27}$$

for the purpose of comparison. As shown in Fig. 3, the 1D case has a flat region in the range $[-4, -3]$, while it yields multimodal responses in the remaining region.

In the numerical experiments, we use a version of MEPE, denoted as MEPE_{CV}, with the balance factor α fixed at 0.5 for a fair comparison. MEPE_{CV} and EIGF start with five initial points evenly

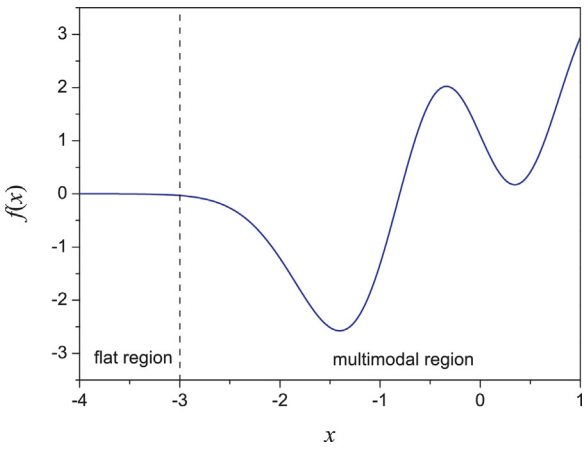

Fig. 3. Plot of the 1D test case.

Table 1
The Kriging modeling results obtained by $MEPE_{CV}$ and EIGF for the 1D test case.

Criterion	$MEPE_{CV}$	EIGF
RMSE	**1.7494E−6**	3.3274E−3

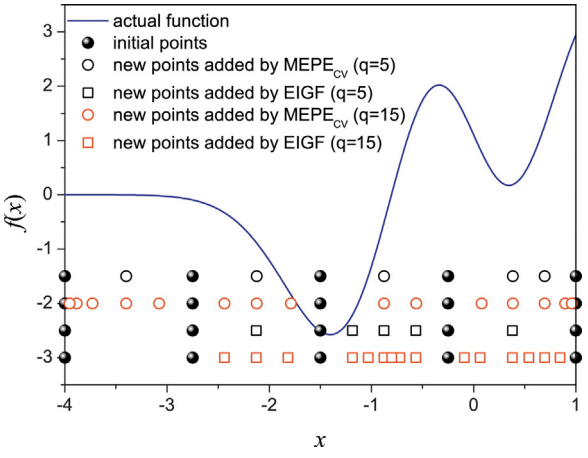

Fig. 4. The sample distributions of $MEPE_{CV}$ and EIGF with the iteration number q respectively being 5 and 15.

distributed along the axis, and then sequentially add 15 new points. For $MEPE_{CV}$ and EIGF, we use 1000 evenly distributed candidate points to form a pool for solving the auxiliary optimization problem like (25) in order to obtain new points. In each iteration, after involving a new point, the Kriging model built by the Matlab toolbox DACE (Lophaven et al., 2002) with the Gaussian correlation function (*@corrgauss*) and the first-order polynomial regression (*@regpoly1*) is updated. We use the root mean square error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{t}\sum_{i=1}^{t}\left(f(\boldsymbol{x}_i) - \hat{f}(\boldsymbol{x}_i)\right)^2} \tag{28}$$

with $t = 1000$ random test points to assess the prediction quality of the Kriging model for this 1D case.

Table 1 shows the Kriging modeling results obtained by $MEPE_{CV}$ and EIGF for the 1D test case, with the best result marked in bold. It is observed that $MEPE_{CV}$ performs much better than EIGF. To account for the remarkable performance of $MEPE_{CV}$, Fig. 4 illustrates the sample distributions of $MEPE_{CV}$ and EIGF with the iteration number q respectively being 5 and 15. It is observed that in the early stage ($q = 5$), both $MEPE_{CV}$ and EIGF identify the multimodal responses in the range [−3, 1]. Hence, $MEPE_{CV}$ samples four out of the five new points in this range, while EIGF puts all the five new points there. When the sampling process keeps proceeding, $MEPE_{CV}$ uses the CV error to identify that the prediction errors in the multimodal region are relatively small, thus it begins to explore the flat region. As a result, the entire domain has been fully explored and the final Kriging model yields sufficiently accurate predictions. On the contrary, because of the response variation term, EIGF always identifies that the multimodal region is informative and keeps exploiting it by putting all the 15 new points there. As a result, the final Kriging model produces poor predictions.

Table 2
The Kriging modeling results obtained by MEPE and $MEPE_{CV}$ for the 1D test case.

Criterion	MEPE	$MEPE_{CV}$
RMSE	**4.7965E−7**	1.7494E−6

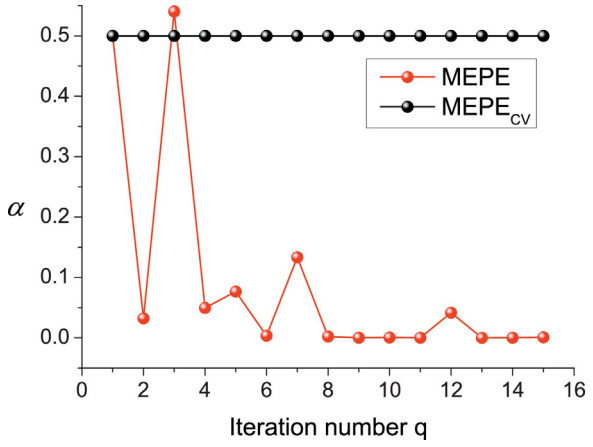

Fig. 5. The α values obtained during the MEPE and $MEPE_{CV}$ sampling processes.

3.2.2. The impact of adaptive balance strategy

Compared to $MEPE_{CV}$, the complete MEPE approach employs an additional adaptive balance strategy in order to improve the sampling performance. To show the merits of this adaptive balance strategy, we apply MEPE and $MEPE_{CV}$ to the 1D test case for generating 15 new points. Table 2 shows the Kriging modeling results obtained by MEPE and $MEPE_{CV}$ for the 1D test case, with the best result marked in bold. It is found that MEPE provides more accurate predictions than $MEPE_{CV}$.

The results of MEPE and $MEPE_{CV}$ are directly caused by the value of balance factor α. Fig. 5 shows the α values obtained during the MEPE and $MEPE_{CV}$ sampling processes. It is found that MEPE adjusts the α value dynamically. In the early stage, MEPE often assigns a relatively large α value in order to exploit the multimodal region; thereafter, MEPE prefers global exploration by assigning small α values. On the contrary, $MEPE_{CV}$ fixes the α value at 0.5 through the whole process. Hence, it cannot flexibly adapt to the characteristics of the target function.

Finally, Fig. 6 offers the convergence curves of MEPE, $MEPE_{CV}$ and EIGF for the 1D test case. It is found that because of the local-exploitation biased mechanism, EIGF has a good performance in the early stage ($q \leq 11$), but it quickly fails to further improve the quality of Kriging predictions. By using the valid local exploitation and the adaptive balance strategy, the proposed MEPE approach performs better through the whole process, and finally refines the Kriging model to higher accuracy.

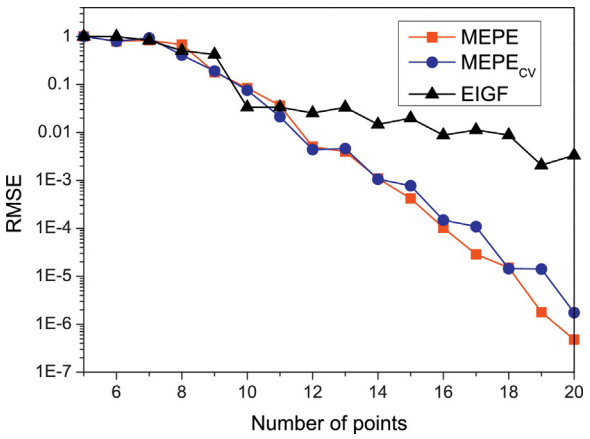

Fig. 6. The convergence curves of MEPE, MEPE$_{CV}$ and EIGF for the 1D test case.

4. Additional numerical experiments

4.1. Test cases

To comprehensively assess the performance of the proposed MEPE adaptive sampling approach for Kriging metamodeling, we employ four benchmark cases and four engineering cases from 2D to 10D. The mathematical formulas of four benchmark cases are listed below.

Peak (PK):

$$f(\boldsymbol{x}) = 3(1-x_1)^2 e^{(-x_1^2-(x_2+1)^2)} - 10\left(\frac{x_1}{5} - x_1^3 - x_2^5\right) e^{(-x_1^2-x_2^2)} - \frac{1}{3} e^{(-(x_1+1)^2-x_2^2)}, \quad \boldsymbol{x} \in [-4,4]^2 \tag{29}$$

Goldstein & Price (GP):

$$f(\boldsymbol{x}) = \left[1+(x_1+x_2+1)^2(19-14x_1+13x_1^2-14x_2+6x_1x_2+3x_2^2)\right] \times \left[30+(2x_1-3x_2)\times(18-32x_1+12x_1^2-48x_2-36x_1x_2+27x_2^2)\right], \quad \boldsymbol{x} \in [-2,2]^2 \tag{30}$$

Shubert (SH):

$$f(\boldsymbol{x}) = \left(\sum_{i=1}^{5} i\cos((i+1)x_1+i)\right) \times \left(\sum_{i=1}^{5} i\cos((i+1)x_2+i)\right), \quad \boldsymbol{x} \in [1,3]^2 \tag{31}$$

Hart6 (H6):

$$f(x) = -\sum_{i=1}^{4} c_i \exp\left[-\sum_{j=1}^{6} a_{ij}(x_j - p_{ij})^2\right], \quad \boldsymbol{x} \in [0,1]^6$$

where

$$\left[a_{ij}\right]_{j=1,\ldots,6} = \begin{bmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{bmatrix}$$

$$c_i = [1 \quad 1.2 \quad 3 \quad 3.2]^T \tag{32}$$

$$\left[p_{ij}\right]_{j=1,\ldots,6} = \begin{bmatrix} 0.1312 & 0.1696 & 0.5596 & 0.0124 & 0.8283 & 0.5886 \\ 0.2329 & 0.4139 & 0.8307 & 0.3736 & 0.1004 & 0.9991 \\ 0.2348 & 0.1451 & 0.3522 & 0.2883 & 0.3047 & 0.6550 \\ 0.4047 & 0.8828 & 0.8732 & 0.5743 & 0.1091 & 0.0381 \end{bmatrix}$$

It is found that the PK case yields most of the nonlinear responses around the origin. This case is expressed as the combination of three bivariate Gaussian density functions with different centers and some polynomial terms. The GP case has most of the nonlinear responses around the boundaries, and the boundary responses have several orders of magnitude larger than the interior responses. The SH case has many peaks and valleys over the entire domain. Finally, compared to PK, the 6D H6 case, which is also a combination of some Gaussian density functions, has higher dimensionality and more local optima, i.e., more complex responses.

We also assess the performance of the proposed adaptive sampling approach on four real-world problems. The first engineering case is a 2D Mach-Cruise (MaCru) design problem for an aircraft. The two design variables are the Mach number in the range [0, 0.8] and the cruise altitude (m) in the range [0,10,000], and the output is the fuel flow (kg/s) calculated by a simulation software.

The second 5D engineering case (Yang and Xue, 2015) is the design of a direct methanol fuel cell system that is used to generate electricity via reaction with oxygen in the air. This kind of power system is potential for mobile devices such as laptops and cell phones, due to the low pollution and high fuel efficiency. In this system, it has an anode with flow channel for methanol, polymer electrolyte membrane, and cathode with flow channel for air. The fuel cells, which are connected in the form of a stack, are connected with pumps, fuel tank, control system, etc. The output voltage V of a fuel cell system is usually influenced by some operating parameters, e.g., the current density, temperature, methanol concentration, flow rate of methanol, and flow rate of air. Based on extensive experiments (Yang et al., 2011), a semi-empirical model is developed for approximating the voltage V as

$$\begin{aligned} V = {} & 1.21 - 3.7534\times10^{-5}T - 3.1534\times10^{-4}T\ln C_{ME} \\ & + 6.62\times10^{-5}T\ln F_{AIR} - 0.7499 - 6.9897e^{\left(\frac{916.91}{T} - 4.6392\right)I} \\ & - [1.2658\times10^{5}I^3 + 46196I^2 - 4281I - 0.4029T \\ & - 18.8094C_{ME}^2 + 18.8094C_{ME} + 10.496] \\ & \times [\ln I - 3.9056 + 2.9582\times10^{-4} \\ & \times \left(\ln C_{ME} + \ln\left(1 - \frac{1}{5.3466\times10^{7}e^{-5182.4/T}C_{ME}^2}\mathrm{I}\right)\right)] \\ & \times [-1.2687\times10^{5}I^3 - 46221I^2 + 4283.6I + 0.4033T \\ & + 18.818C_{ME}^2 - 18.818C_{ME} - 10.572] \\ & \times \left[\ln \mathrm{I} - 3.8959 - 8.2402\times10^{-4}\ln F_{AIR}\right] + 31.583I^2\ln F_{ME} \end{aligned} \tag{33}$$

Table 3
The input variables and their ranges for the fuel cell case.

Variable	Range	Description
I	[0.0003,0.08]	Current density (A/cm^2)
T	[298,343]	Temperature (K)
C_{ME}	[0.25,2]	Methanol concentration (M)
F_{ME}	[3.5,5.5]	Flow rate of methanol (ccm)
F_{AIR}	[81.2,140.8]	Flow rate of air (ccm)

Table 3 describes the five input variables and their ranges.

The third engineering case is a 8D abalone dataset from UCI repository (Lichman, 2013). The abalone dataset, which predicts the age of abalone from physical measurements, contains 4177 points with known output responses. This dataset has eight inputs as the sex, length, diameter, height, whole weight, shucked weight, viscera weight and shell weight, and has a single output as rings. Note that since the sex is a nominal indicator, we delete it in the dataset.

Table 4
The input variables and their ranges for the wing weight case.

Variable	Range	Description
S_w	[150,200]	Wing area (ft^2)
W_{fw}	[220,300]	Weight of fuel in the ring (lb)
A	[6,10]	Aspect ratio
Λ	[−10,10]	Quarter-chord sweep (deg)
q	[16,45]	Dynamic pressure at cruise (lb/ft^2)
λ	[0.5,1]	Taper ratio
tc	[0.08,0.18]	Aerofoil thickness to chord ratio
N_Z	[2.5,6]	Ultimate load factor
W_{dg}	[1700,2500]	Flight design gross weight
W_p	[0.025,0.08]	Paint weight (lb/ft^2)

Table 5
The sampling configurations for the eight test cases.

Case	n	Initial no.	Sequential no.	Candidate no.	Test no.
PK	2	20	130	10,000	10,000
GP	2	20	130	10,000	10,000
SH	2	20	130	10,000	10,000
H6	6	60	340	15,000	15,000
MaCru	2	20	130	10,000	10,000
CellVol	5	50	250	15,000	15,000
Abalone	7	70	430	2500	1177
Wing	10	100	800	20,000	20,000

The final engineering case is a conceptual level estimate of the weight of a Cessna C172 Skyhawk aircraft wing (Raymer, 2006), which is formulated as

$$f(\boldsymbol{x}) = 0.036 S_w^{0.758} W_{fw}^{0.0035} \left(\frac{A}{\cos^2\Lambda}\right) q^{0.006} \lambda^{0.04} \left(\frac{100tc}{\cos\Lambda}\right) (N_Z W_{dg})^{0.49} + S_w W_p \quad (34)$$

This 10D case is derived from fitting a curve to existing aircraft data. Table 4 describes the ten input variables and their ranges for the wing weight case.

4.2. Test scheme

In the numerical experiments, we compare MEPE with three Kriging-based sampling approaches including the EIGF sampling approach (Lam, 2008), the AME sampling approach (Liu et al., 2016a), and the MMSE sampling approach (Jin et al., 2002).

Table 5 shows the sampling configurations for the eight test cases. For all the test cases except MaCru and Abalone that have no analytical expressions, the four sampling approaches start with a set of initial points generated by the MCSR space-filling sequential sampling approach (Liu et al., 2015), which is capable to efficiently generate good semi-optimal LHD points. The initial sample size is set as $10n$ (n is the number of inputs), since Loeppky et al. (2009) provided reasons and evidence supporting that this is an appropriate initial sample size when we use Kriging to approximate the target function. It is notable that for each test case, MCSR is repeated ten times to generate ten initial sets. With the ten different initial sets, we can assess the performance of a sampling approach comprehensively. Finally, for the Abalone dataset, the sampling approaches also start with ten initial sets, each of which has $10n$ points that are randomly selected from the given 4177 points; for the MaCru dataset, the $10n$ initial sets are generated by MCSR, and then they are evaluated through a design software to obtain the associated responses.

Note that each of the employed four sampling approaches selects a new point in each iteration by solving an auxiliary optimization problem like (25). This kind of optimization problem usually has a sophisticated surface with many local optima, and is hard to be solved by current global optimization algorithms, e.g., the genetic algorithm (GA) (Crombecq et al., 2011). Thus, for the sake of simplicity, we maintain a pool **U** containing numerous candidate points, from which the sampling approach selects the one with the highest score as $\boldsymbol{x}_{new}$ in each iteration. The numbers of sequential points, candidate points and test points for each case are shown in Table 5. For the 2D PK, GP, SH and the MaCru dataset, the 10,000 candidates are grid points; for H6, CellVol and Wing, the candidate points are generated by the Matlab routine *lhsdesign* and the *maximin* criterion with a maximum of 1000 iterations. Besides, the number of random test points to assess the RMSE value of the Kriging model keeps the same as that of the candidate points. For the Abalone dataset, the candidate and test points are randomly selected from the dataset.

Table 6
The average Kriging modeling results obtained by different sampling approaches for the eight test cases.

Case	MEPE	EIGF	AME	MMSE
PK	**3.5283E−2** (**1.0660E−2**)	3.9373E−2 (1.4830E−2)	4.3690E−2 (2.9819E−2)	1.2316E−1 (1.9501E−2)
GP	**9.3308E+0** (**7.8620E+0**)	6.1733E+3 (1.9560E+3)	2.2480E+1 (6.2488E+0)	1.3509E+1 (1.0570E+1)
SH	2.0887E−3 (3.9440E−4)	1.8493E−2 (5.4914E−3)	6.0671E−3 (1.7826E−3)	**1.9107E−3** (**2.6789E−4**)
H6	**1.3053E−1** (**2.1749E−2**)	1.4638E−1 (3.4316E−2)	1.5006E−1 (1.9825E−2)	2.2343E−1 (2.8999E−2)
MaCru	**5.2991E−9** (**2.7656E−10**)	2.0122E−7 (1.4648E−7)	2.4860E−8 (9.7092E−9)	**5.2870E−9** (**3.6609E−10**)
CellVol	**7.6308E−3** (**6.4154E−4**)	2.0910E−2 (6.8130E−3)	8.2911E−3 (1.1621E−3)	8.5635E−3 (6.9469E−4)
Abalone	1.7686E+0 (1.3296E−1)	**1.7146E+0** (**9.0160E−2**)	1.7525E+0 (1.3168E−1)	1.8819E+0 (1.7698E−1)
Wing	**5.5436E+2** (**1.7321E+1**)	5.6960E+2 (4.4379E+0)	5.6756E+2 (4.7139E+1)	5.7858E+2 (2.9088E+0)

4.3. Results and discussions

Table 6 shows the average Kriging modeling results (over ten trials) obtained by different sampling approaches for the eight test cases, with the best results marked in bold. The values in parentheses refer to standard deviations. It is observed that the proposed MEPE adaptive sampling approach stands out in six out of the eight cases. Among the remaining three sampling approaches, AME outperforms EIGF in five cases, and is comparable to MMSE.

In detail, for the first four benchmark cases with different characteristics, Fig. 7 shows the average convergence curves of different sampling approaches. For the PK case which contains a nonlinear center region and a flat boundary region, MEPE, EIGF and AME that consider local exploitation distinctly outperform the space-filling MMSE sampling approach. Among the first three adaptive sampling approaches, due to the local-exploitation biased mechanism, the converge of EIGF becomes unstable in the later stage. Since AME uses a fixed balance pattern which contains a sequence of weights from global to local, it outperforms EIGF but is worse than MEPE. While the proposed MEPE is capable of refining the Kriging model more efficiently due to the effective local exploitation and the dynamic adaptive balance strategy. Consequently, MEPE has the fastest convergence speed during most of the sampling process.

For the GP case with its main characteristics lying around the boundaries, it is observed that MEPE still provides the best Kriging predictions, but MMSE shows a competing performance during most of the sampling process. This is because that MMSE favors sampling more points around the boundaries rather than the interior regions, especially at the beginning of the sampling process. For MMSE, with the increase of dimensionality, the number of boundary points grows rapidly such that the ratio of which to the total

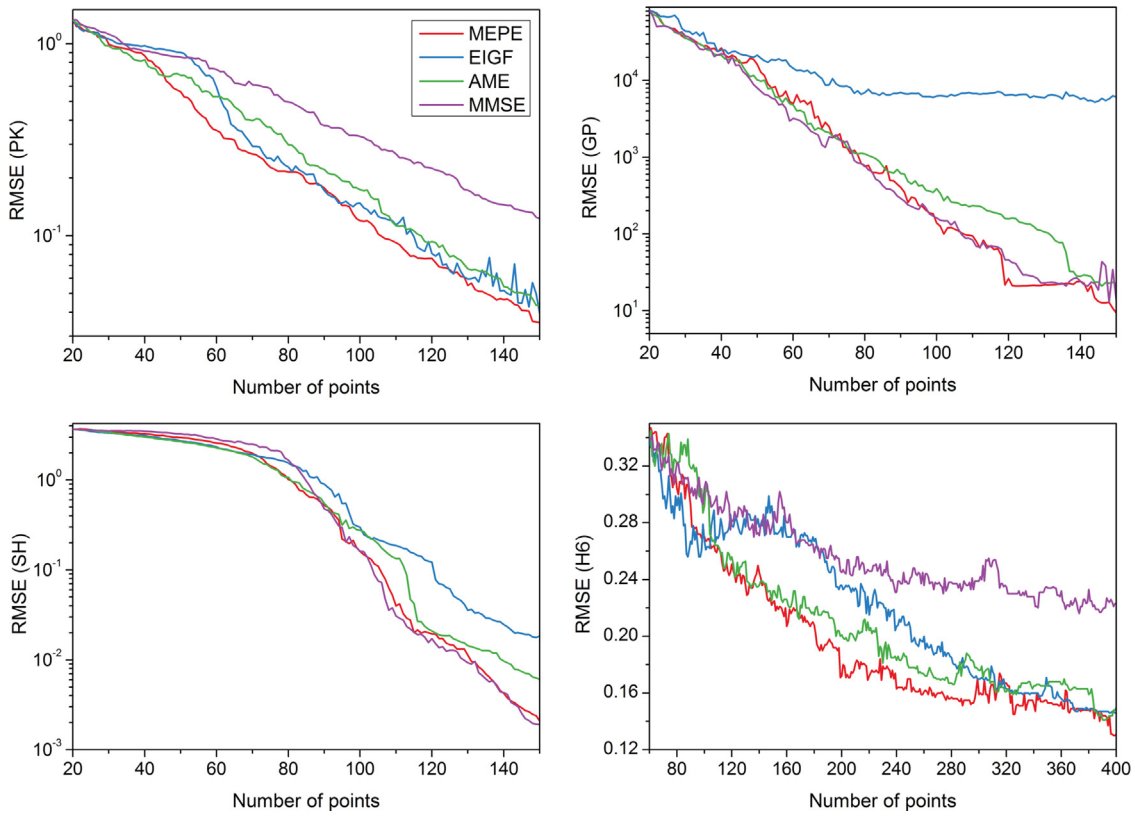

Fig. 7. The average convergence curves of different sampling approaches for the four benchmark cases.

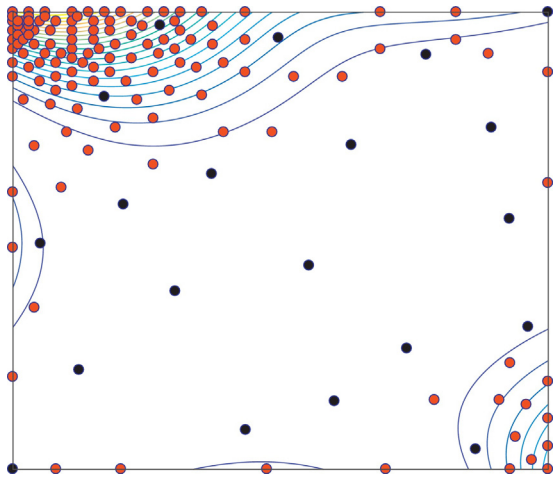

Fig. 8. The sample distribution selected from a run of EIGF for the GP case.

sample size is $(1-(1-2/N)^n)$ given N^n grid points (Beck and Guillas, 2016). This boundary property helps MMSE yield good results for GP; however, it makes MMSE perform worse for PK and H6 which have main characteristics in the interior regions.

In addition, it is observed that EIGF provides the worst performance for the GP case. The poor performance of EIGF is caused by the same reason for PK. But it becomes more serious for GP because the boundary responses of GP have several orders of magnitude larger than the interior responses. Consequently, this extremely large response variation makes EIGF fall into exploitation around the boundaries. Fig. 8 shows the sample distribution selected from a run of EIGF for GP where the blue circles are initial points and the red circles are sequentially added new points. It is observed that EIGF ignores the interior region and puts all the new points around the boundaries.

Next, let us move to the SH case which represents a frequently encountered practical scenario where the complex problem shows the main features over the entire domain, not just in several subregions. For this kind of function, the sampling approaches that tend to fill the domain evenly are expected to be a good choice. The results of SH show that the space-filling MMSE approach has the best performance. Compared to MMSE, it is found that the proposed MEPE adaptive sampling approach, however, shows a very competing performance, e.g., it outperforms MMSE in four out of the ten runs. While AME shows a little bit worse performance for SH because of the fixed balance strategy. As for EIGF, it performs good in the early stage, but quickly becomes worse due to the inappropriate local exploitation term.

Finally, for the H6 case which is similar to PK but has higher dimensionality and more complex responses, the test results are similar to that of PK. The first three adaptive sampling approaches outperform MMSE, and MEPE is the best. In addition, it is observed that compared to PK, the average convergence curves of the four sampling approaches for H6 become more fluctuant because of the increase of dimensionality. This fluctuation phenomenon is again obviously observed in some other high-dimensional cases below, e.g., the Wing case.

Fig. 9 shows the average convergence curves of different sampling approaches for the four engineering cases. For the MaCru case with the characteristics over the entire domain, similar to the results for SH, it is found that MMSE provides the fastest convergence speed in most of the sampling process. While MEPE shows a comparable performance, and it catches up and yields almost the same performance with that of MMSE in the final stage. The AME sampling approach shows a competing performance in the early stage, but thereafter it becomes incapable of efficiently improv-

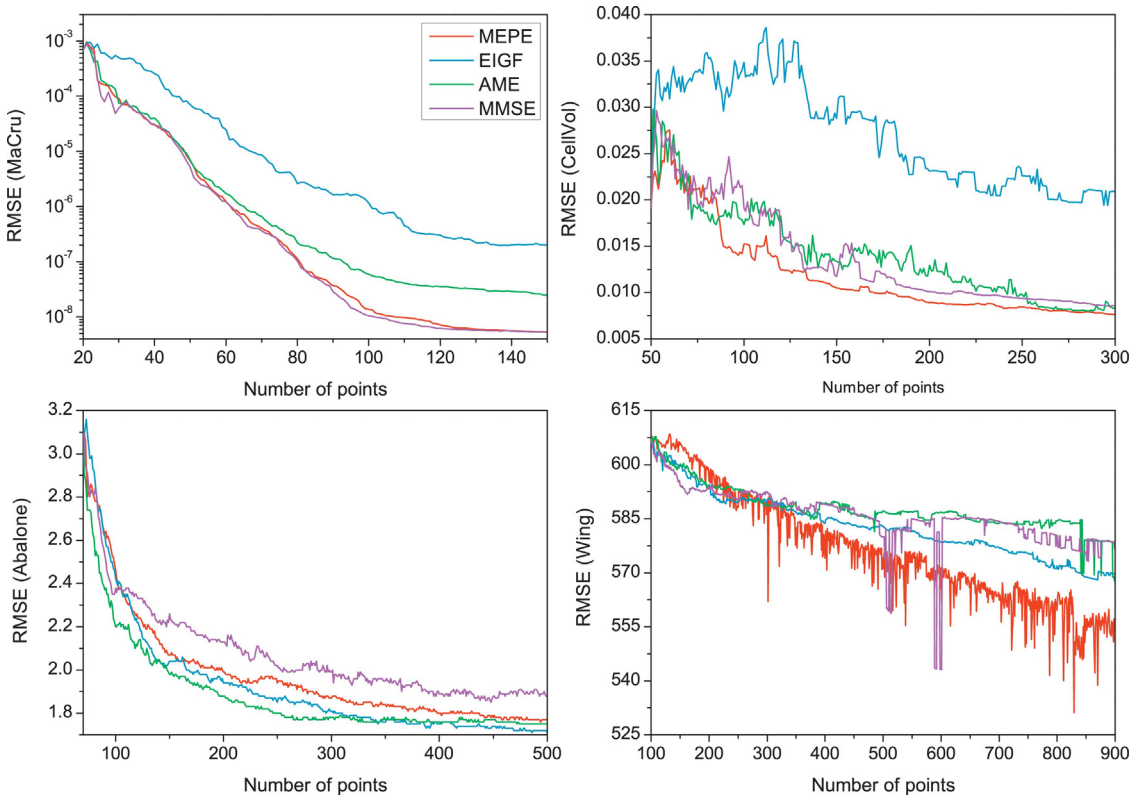

Fig. 9. The average convergence curves of different sampling approaches for the four engineering cases.

ing the Kriging predictions as much as MMSE and MEPE. The EIGF sampling approach, however, shows the worst results through the whole sampling process.

For the CellVol case, EIGF again provides the worst performance. For this case, MEPE converges faster than AME and MMSE through the whole sampling process. Besides, MMSE outperforms AME during most of the sampling process, but AME stands out and has a slightly better performance in the last stage.

Though having been shown that EIGF adopts an inappropriate local exploitation term, this local-exploitation biased mechanism, however, may help EIGF perform well in some scenarios, e.g., the Abalone case that needs local exploitation. On the contrary, for the MMSE sampling approach that does not consider local exploitation, it provides the worst performance for the Abalone case.

Finally, for the 10D Wing case, it is found that with the adding of new points, MEPE converges faster than other approaches. Besides, it is observed that all the convergence curves show a global downward trend but have drastic fluctuations. For example, in some iterations of MEPE and MMSE, the RMSE value suddenly decreases to a much small value and then quickly raises back to the normal level. This is caused by the fact that the metamodel is hard to fit well in high dimensions, and the required number of points usually grows exponentially with dimensions, known as "the curse of dimensionality".

To summarize, due to the cross-validation based local exploitation and the adaptive balance strategy, the proposed MEPE adaptive sampling approach is very promising for problems with diverse characteristics from low to high dimensions.

4.4. More studies on the adaptive balance strategy

The adaptive balance strategy in MEPE utilizes the actual prediction error at a single point obtained in the previous iteration to sequentially adjust the α value, thereby dynamically balancing the local exploitation and global exploration. The test results for the 1D case have revealed that the adaptive balance strategy helps MEPE achieve a better sampling performance. Since the adaptive balance strategy is the key component of the proposed MEPE sampling approach, more numerical experiments are required to comprehensively validate the effectiveness of this strategy.

To this end, we apply the MEPE_{CV} sampling approach with the balance factor α fixed as 0.5 to the above eight cases. The other running parameters in the testing keep the same as that for MEPE. Fig. 10 shows the boxplots of the Kriging modeling results obtained by MEPE and MEPE_{CV} for the eight test cases. It is observed that MEPE performs better than MEPE_{CV} in seven out of the eight cases, especially for GP, SH and CellVol.

Fig. 11 shows the α values obtained in two runs of MEPE for PK and SH, respectively. It is observed that there are fluctuations in α over the iterations, which lead to frequent switches between local exploitation and global exploration. Here, MEPE almost conducts a pure global exploration like MMSE when $\alpha < 0.01$. It is found that for SH that has characteristics over the entire domain, MEPE conducts a pure global exploration in 46 out of the 130 iterations; while for PK that contains a significantly nonlinear subregion, MEPE only conducts a pure global exploration in 23 out of the 130 iterations.

4.5. Running time

To validate the sampling efficiency, Table 7 offers the average running times (over 10 trials) of different sampling approaches for the eight test cases. Not that all the codes are run in the Matlab environment with an Intel 3.70 GHz processor. It is found that MMSE spends the least running time, while AME is the most time-consuming approach.

The running time of the proposed MEPE sampling approach comes at two aspects. The first is the cross-validation process which needs to refit the Kriging model m times in order to calculate the

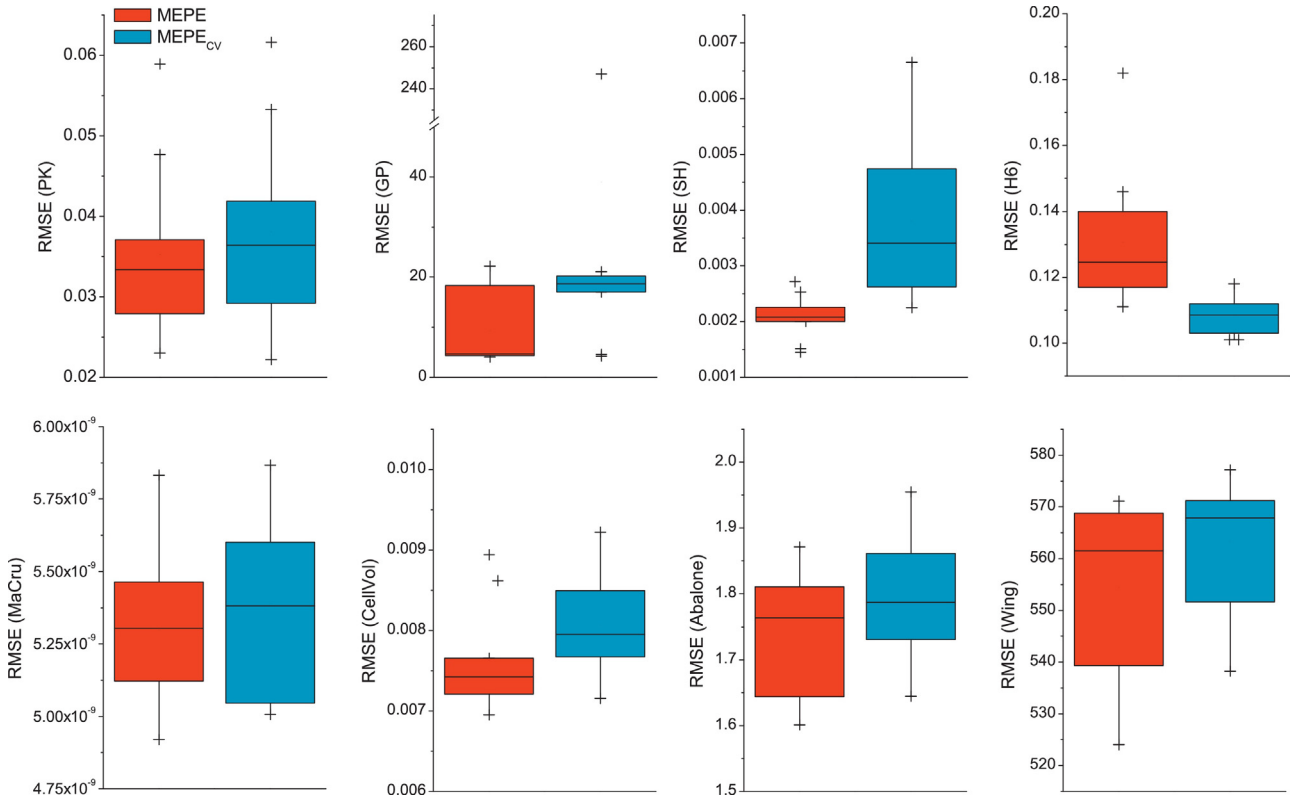

Fig. 10. Boxplots of the Kriging modeling results obtained by MEPE and MEPE$_{CV}$ for the eight test cases.

Table 7
The average running times (s) of different sampling approaches for the eight test cases.

Case	MEPE	EIGF	AME	MMSE
PK	61.4 (1.0)	36.8 (0.5)	93.2 (1.4)	21.9 (0.3)
GP	63.5 (0.9)	39.2 (0.8)	94.5 (1.0)	22.4 (0.3)
SH	60.8 (4.5)	35.4 (0.8)	92.8 (0.7)	22.4 (0.4)
H6	1090.5 (29.1)	638.2 (5.4)	1623.1 (34.5)	525.8 (5.6)
MaCru	64.3 (1.0)	48.5 (0.6)	113.6 (1.7)	31.2 (0.3)
CellVol	368.4 (4.1)	260.2 (2.2)	591.7 (7.4)	218.5 (3.2)
Abalone	1328.0 (46.6)	408.9 (29.2)	2497.7 (193.8)	408.1 (27.6)
Wing	22,531.6 (552.8)	12,268.6 (127.3)	45,611.0 (2105.4)	8125.8 (63.6)

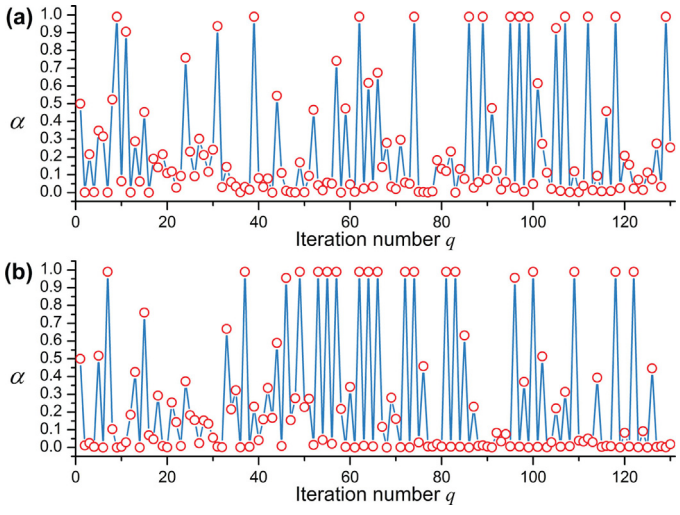

Fig. 11. The α values obtained in two runs of MEPE for (a) PK and (b) SH, respectively.

CV errors at m observed points. The second is to solve the auxiliary optimization problem (25) in order to obtain a new point. In this part, it should estimate the prediction variances of the candidate points, and compute the distances between the candidate points and the observed points.

The other three sampling approaches more or less contain the similar parts. MMSE only needs to solve an auxiliary optimization problem by estimating the prediction variances of the candidate points. EIGF needs to solve an auxiliary optimization problem by computing the prediction variances and distances of the candidate points. Thus, the two approaches run faster than MEPE. While AME not only contains the cross-validation process and the auxiliary optimization process, but also conducts many matrix operations. Thus, it turns out to be the most time-consuming approach.

The above discussions offer an insight into the computational complexity of each sampling approach. It is worth noting that these sampling approaches are designed for simulation-based problems, wherein a simulation may take hours or even days. Therefore, compared to the time spent on simulation, the running times of these sampling approaches in practice are negligible.

5. Concluding remarks

This article presents a novel adaptive sampling approach for Kriging metamodeling for simulation-based engineering problems. It selects new points by maximizing an expected prediction error criterion that considers both the bias and variance information. Moreover, this approach employs a novel adaptive trade-off strategy to dynamically balance the local exploitation and global exploration via the error information from the previous iteration.

We compared the proposed MEPE adaptive sampling approach to some existing adaptive sampling approaches on four benchmark cases and four engineering cases. The numerical results reveal that MEPE usually converges faster and builds Kriging models with more

accurate predictions using the same number of points. Thus, it is a very promising sampling approach to construct Kriging models for problems with diverse characteristics.

Acknowledgments

This work was conducted within the Rolls-Royce@NTU Corporate Lab with support from the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme.It is also partially supported by the Data Science and Artificial Intelligence Research Center (DSAIR) and the School of Computer Science and Engineering at Nanyang Technological University.

References

Beck, J., Guillas, S., 2016. Sequential design with mutual information for computer experiments (mice): emulation of a tsunami model. SIAM/ASA J. Uncertain. Quantif. 4, 739–766.
Busby, D., 2009. Hierarchical adaptive experimental design for Gaussian process emulators. Reliab. Eng. Syst. Saf. 94, 1183–1193.
Busby, D., Farmer, C.L., Iske, A., 2007. Hierarchical nonlinear approximation for experimental design and statistical data fitting. SIAM J. Sci. Comput. 29, 49–69.
Christen, J.A., Sansó, B., 2011. Advances in the sequential design of computer experiments based on active learning. Commun. Stat.: Theory Methods 40, 4467–4483.
Cohn, D.A., 1996. Neural network exploration using optimal experiment design. Neural Netw. 9, 1071–1083.
Crombecq, K., Laermans, E., Dhaene, T., 2011. Efficient space-filling and non-collapsing sequential design strategies for simulation-based modeling. Eur. J. Oper. Res. 214, 683–696.
Deschrijver, D., Crombecq, K., Nguyen, H.M., Dhaene, T., 2011. Adaptive sampling algorithm for macromodeling of parameterized-parameter responses. IEEE Trans. Microw. Theory Tech. 59, 39–45.
Eason, J., Cremaschi, S., 2014. Adaptive sequential sampling for surrogate model generation with artificial neural networks. Comput. Chem. Eng. 68, 220–232.
Farhang-Mehr, A., Azarm, S., 2005. Bayesian meta-modelling of engineering design simulations: a sequential approach with adaptation to irregularities in the response behaviour. Int. J. Numer. Methods Eng. 62, 2104–2126.
Garud, S.S., Karimi, I., Kraft, M., 2017. Smart sampling algorithm for surrogate model development. Comput. Chem. Eng. 96, 103–114.
Jin, R., Chen, W., Sudjianto, A., 2002. On sequential sampling for global metamodeling in engineering design. In: ASME 2002 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers, pp. 539–548.
Jones, D.R., Schonlau, M., Welch, W.J., 1998. Efficient global optimization of expensive black-box functions. J. Glob. Optim. 13, 455–492.
Joseph, V.R., Hung, Y., Sudjianto, A., 2008. Blind kriging: a new method for developing metamodels. J. Mech. Des. 130, 031102.
Journel, A.G., Huijbregts, C.J., 1978. Mining Geostatistics. Academic Press.
Kennedy, M.C., O'Hagan, A., 2000. Predicting the output from a complex computer code when fast approximations are available. Biometrika 87, 1–13.
Kenny, Q.Y., Li, W., Sudjianto, A., 2000. Algorithmic construction of optimal symmetric latin hypercube designs. J. Stat. Plan. Inference 90, 145–159.
Kleijnen, J.P., 2009. Kriging metamodeling in simulation: a review. Eur. J. Oper. Res. 192, 707–716.
Lam, C.Q., 2008. Sequential adaptive designs in computer experiments for response surface model fit (Ph.D. thesis). The Ohio State University, Columbus, OH.
Lichman, M., 2013. UCI Machine Learning Repository.
Lin, Y., Mistree, F., Allen, J.K., Tsui, K.L., Chen, V.C., 2004. A sequential exploratory experimental design method: development of appropriate empirical models in design. In: ASME 2004 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. American Society of Mechanical Engineers, pp. 1021–1035.
Liu, H., Xu, S., Chen, X., Wang, X., Ma, Q., 2017. Constrained global optimization via a direct-type constraint-handling technique and an adaptive metamodeling strategy. Struct. Multidiscip. Optim. 55, 155–177.
Liu, H., Xu, S., Ma, Y., Chen, X., Wang, X., 2016a. An adaptive Bayesian sequential sampling approach for global metamodeling. J. Mech. Des. 138, 011404.
Liu, H., Xu, S., Wang, X., 2015. Sequential sampling designs based on space reduction. Eng. Optim. 47, 867–884.
Liu, H., Xu, S., Wang, X., Meng, J., Yang, S., 2016b. Optimal weighted pointwise ensemble of radial basis functions with different basis functions. AIAA J. 54, 3117–3133.
Loeppky, J.L., Sacks, J., Welch, W.J., 2009. Choosing the sample size of a computer experiment: a practical guide. Technometrics 51, 366–376.
Lophaven, S.N., Nielsen, H.B., Søndergaard, J., 2002. Aspects of the matlab toolbox DACE, Technical Report. Informatics and Mathematical Modelling, Technical University of Denmark, DTU.
Mackman, T., Allen, C., 2010. Investigation of an adaptive sampling method for data interpolation using radial basis functions. Int. J. Numer. Methods Eng. 83, 915–938.
Martin, J.D., 2009. Computational improvements to estimating kriging metamodel parameters. J. Mech. Des. 131, 084501.
Morris, M.D., Mitchell, T.J., Ylvisaker, D., 1993. Bayesian design and analysis of computer experiments: use of derivatives in surface prediction. Technometrics 35, 243–255.
Queipo, N.V., Haftka, R.T., Shyy, W., Goel, T., Vaidyanathan, R., Tucker, P.K., 2005. Surrogate-based analysis and optimization. Prog. Aerosp. Sci. 41, 1–28.
Raymer, D.P., 2006. Aircraft Design: A Conceptual Approach and RDS-Student, Software for Aircraft Design, Sizing, and Performance Set (AIAA Education). American Institute of Aeronautics and Astronautics, Washington, DC.
Romero, D.A., Marin, V.E., Amon, C.H., 2015. Error metrics and the sequential refinement of kriging metamodels. J. Mech. Des. 137, 011402.
Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P., 1989. Design and analysis of computer experiments. Stat. Sci., 409–423.
Settles, B., 2010. Active learning literature survey, Technical Report. University of Wisconsin, Madison.
Shewry, M.C., Wynn, H.P., 1987. Maximum entropy sampling. J. Appl. Stat. 14, 165–170.
Singh, P., Deschrijver, D., Dhaene, T., 2013. A balanced sequential design strategy for global surrogate modeling. In: 2013 Winter Simulations Conference (WSC). IEEE, pp. 2172–2179.
Sundararajan, S., Keerthi, S.S., 2001. Predictive approaches for choosing hyperparameters in Gaussian processes. Neural Comput. 13, 1103–1118.
Turner, C.J., Crawford, R.H., Campbell, M.I., 2007. Multidimensional sequential sampling for NURBs-based metamodel development. Eng. Comput. 23, 155–174.
Viana, F.A., Haftka, R.T., Steffen Jr., V., 2009. Multiple surrogates: how cross-validation errors can help us to obtain the best predictor. Struct. Multidiscip. Optim. 39, 439–457.
Viana, F.A., Simpson, T.W., Balabanov, V., Toropov, V., 2014. Special section on multidisciplinary design optimization: metamodeling in multidisciplinary design optimization: how far have we really come? AIAA J. 52, 670–690.
Viana, F.A., Venter, G., Balabanov, V., 2010. An algorithm for fast optimal latin hypercube design of experiments. Int. J. Numer. Methods Eng. 82, 135–156.
Wang, Z., Ierapetritou, M., 2017. A novel feasibility analysis method for black-box processes using a radial basis function adaptive sampling approach. AIChE J. 63, 532–550.
Xiong, Y., Chen, W., Apley, D., Ding, X., 2007. A non-stationary covariance-based kriging method for metamodelling in engineering design. Int. J. Numer. Methods Eng. 71, 733–756.
Xu, S., Liu, H., Wang, X., Jiang, X., 2014. A robust error-pursuing sequential sampling approach for global metamodeling based on Voronoi diagram and cross validation. J. Mech. Des. 136, 071009.
Yang, Q., Kianimanesh, A., Freiheit, T., Park, S., Xue, D., 2011. A semi-empirical model considering the influence of operating parameters on performance for a direct methanol fuel cell. J. Power Sources 196, 10640–10651.
Yang, Q., Xue, D., 2015. Comparative study on influencing factors in adaptive metamodeling. Eng. Comput. 31, 561–577.
Zhou, Z.H., Wu, J., Tang, W., 2002. Ensembling neural networks: many could be better than all. Artif. Intell. 137, 239–263.