Open Archive Toulouse Archive Ouverte (OATAO)

OATAO is an open access repository that collects the work of some Toulouse researchers and makes it freely available over the web where possible.

This is an author's version published in: https://oatao.univ-toulouse.fr/24160

Official URL: https://doi.org/10.1016/j.advengsoft.2019.03.005

To cite this version:


Any correspondence concerning this service should be sent to the repository administrator:
tech-oatao@listes-diff.inp-toulouse.fr
A Python surrogate modeling framework with derivatives

Mohamed Amine Bouhlel⁎,a, John T. Hwangb, Nathalie Bartolic, Rémi Lafagec, Joseph Morlierd, Joaquim R.R.A. Martins⁢

⁎Corresponding author.
E-mail addresses: mbouhlel@umich.edu (M.A. Bouhlel), jhwang@eng.ucsd.edu (J.T. Hwang), nathalie.bartoli@onera.fr (N. Bartoli), remi.lafage@onera.fr (R. Lafage), joseph.morlier@isae-supaero.fr (J. Morlier), jrram@umich.edu (J.R.R.A. Martins).

1https://github.com/SMTorg/SMT.

ARTICLE INFO

Keywords: Surrogate modeling Gradient-enhanced surrogate modeling Derivatives

ABSTRACT

The surrogate modeling toolbox (SMT) is an open-source Python package that contains a collection of surrogate modeling methods, sampling techniques, and benchmarking functions. This package provides a library of surrogate models that is simple to use and facilitates the implementation of additional methods. SMT is different from existing surrogate modeling libraries because of its emphasis on derivatives, including training derivatives used for gradient-enhanced modeling, prediction derivatives, and derivatives with respect to training data. It also includes unique surrogate models: kriging by partial least-squares reduction, which scales well with the number of inputs; and energy-minimizing spline interpolation, which scales well with the number of training points. The efficiency and effectiveness of SMT are demonstrated through a series of examples. SMT is documented using custom tools for embedding automatically tested code and dynamically generated plots to produce high-quality user guides with minimal effort from contributors. SMT is maintained in a public version control repository.

1. Motivation and significance

In the last few decades, numerical models for engineering have become more complex and accurate, but the computational time for running these models has not necessarily decreased. This makes it difficult to complete engineering tasks that rely on these models, such as design space exploration and optimization. Surrogate modeling is often used in the context of design optimization because of the repeated model evaluations that are required. Derivatives play an important role in optimization, because problems with a large number of optimization variables require gradient-based algorithms for efficient scalability. Therefore, situations frequently arise where there are requirements for surrogate models associated with the computation or use of derivatives.

There are three types of derivatives in surrogate modeling: prediction derivatives, training derivatives, and output derivatives. Prediction derivatives are the derivatives of the surrogate model outputs with respect to the inputs, and they are the derivatives required when using a surrogate model in gradient-based optimization. Training derivatives are derivatives of the training outputs with respect to the training inputs that provide additional training data that increase the accuracy of the surrogate model. Output derivatives are derivatives of the prediction outputs with respect to the training outputs, which are required if a surrogate model is reconstructed within a gradient-based optimization process. We describe these derivatives in more detail in Section 3.3.

Various packages that build surrogate models have been developed using different programming languages, such as Scikit-learn in Python [1], SUMO in MATLAB [2], and GPML in MATLAB and Octave [3]. However, these packages do not handle the derivatives.

In this paper, we introduce a new Python package called the surrogate modeling toolbox (SMT). SMT is different from existing surrogate modeling libraries because of its emphasis on derivatives, including all three types of derivatives described above. SMT also includes newly developed surrogate models that handle derivatives and
do not exist elsewhere: partial least-squares (PLS)-based surrogate models [4–6], which are suitable for high-dimensional problems, and the regularized minimal-energy tensor-product spline (RMTS), which is suitable for low-dimensional problems with up to hundreds of thousands of sampling points [7]. To use SMT, the user should first provide a set of training points. This could be done either by using the sampling techniques and benchmarking functions implemented in SMT or by directly importing the data. Then, the user can build a surrogate model based on the training points and make predictions for the function values and derivatives.

The main goal of this work is to provide a simple Python toolbox that contains a set of surrogate modeling functions and supports different kinds of derivatives that are useful for many applications in engineering. SMT includes sampling techniques, which are necessary for the construction of a surrogate model. Various sample functions are also included to facilitate the benchmarking of different techniques and for reproducing results. SMT is suitable for both novice and advanced users and is supported by detailed documentation available online with examples of each implemented surrogate modeling approach. It is hosted publicly in a version-controlled repository, and is easily imported and used within Python scripts. It is released under the New BSD License and runs on Linux, macOS, and Windows operating systems. Regression tests are run automatically on each operating system whenever a change is committed to the repository.

The remainder of this paper is organized as follows. First, we present the architecture and the main implementation features of SMT in Section 2 and then we describe the methods implemented in SMT in Section 3. Section 4 gives an example of SMT usage and presents a set of benchmarking functions implemented within SMT. We apply SMT to two engineering problems in Section 5 and present the conclusions in Section 6.

2. Software architecture, documentation, and automatic testing

SMT is composed of three main modules (sampling methods, problems, and surrogate models) that implement a set of sampling techniques, benchmarking functions, and surrogate modeling techniques, respectively. Each module contains a common interface inherited by the corresponding methods, and each method implements the functions required by the interface, as shown in Fig. 1.

SMT’s documentation is written using reStructuredText and is generated using the Sphinx package for documentation in Python, along with custom extensions. The documentation pages include embedded code snippets that demonstrate the usage. These code snippets are extracted dynamically from actual tests in the source code, ensuring that the code snippets are always up to date. The print output and plots from the code snippets are also generated dynamically by custom Sphinx extensions and embedded in the documentation page. This leads to high-quality documentation with low effort, requiring only the addition of a custom directive and the path to locating the test code. Similarly, another custom directive embeds a table of options, default values, valid types, valid values, and descriptions in the documentation for the surrogate model, sampling method, or benchmarking problem. The documentation also uses existing Sphinx directives to embed descriptions of the user-callable methods in the classes.

In addition to the user documentation, we also provide developer documentation that explains how to contribute code to SMT. The developer documentation includes a different list of application programming interface (API) methods for the SurrogateModel, SamplingMethod, and Problem classes, which are classes that must be implemented to create a new surrogate modeling method, sampling technique, or benchmarking problem, respectively.

When a developer issues a pull request, the request is merged once the automatically triggered tests run successfully and at least one reviewer approves it. The repository on GitHub is linked to two continuous integration testing services, Travis CI (for Linux and macOS) and AppVeyor (for Windows), which trigger test suite execution whenever code is committed and prevent changes from being merged if any of the tests fail.

3. Surrogate modeling methods

To build a surrogate model, two main steps are necessary. First, we generate a set of training points from the input space where the quantity of interest is computed. This step can be done by using one of the sampling techniques implemented in SMT (Section 3.1), or by uploading an existing training dataset. Second, we train the desired surrogate model on those points and make a prediction of the output values and derivatives (Section 3.2).

3.1. Sampling methods

SMT contains a library of sampling methods used to generate sets of points in the input space, either for training or for prediction.

**Random sampling** This class creates random samples from a uniform distribution over the design space.

**Latin hypercube sampling** This is a statistical method for generating a quasi-random sampling distribution. It is among the most popular sampling techniques in computer experiments thanks to its simplicity and projection properties in high-dimensional problems. Five construction criteria are available in SMT: four criteria defined in the pyDOE package and the enhanced stochastic evolutionary criterion [8].

**Full-factorial sampling** This is a common sampling method where all input variables are set at two levels each. These levels are usually denoted by +1 and −1 for high and low levels, respectively. Full-factorial sampling computes all possible combinations of these levels across all input variables.

Figure 2 shows the implementation of the Random class, which inherits from the SamplingMethod class.

3.2. Surrogate models

Table 1 lists the surrogate modeling methods currently available in SMT and summarizes the advantages and disadvantages of each method. The methods include both well-established methods and methods recently developed by the authors that use derivative information. Among the well-established methods, we implement kriging [9], radial basis functions (RBF) [10], inverse distance weighting (IDW) [11], least squares (LS) [12, ch. 3], and quadratic polynomials (QP) [12, ch. 3]. In Fig. 3, we show the implementation of the RBF class, which inherits from the SurrogateModel class.

The methods that we recently developed are kriging combined with partial least squares (KPLS) [4], KPLSK for the construction of a standard kriging model in high-dimensional problems [5], gradient-enhanced KPLS (GE-KPLS) [6], and RMTS [7].

Surrogate models provide a vector of prediction outputs \( \mathbf{y} \) for a given vector of prediction inputs \( \mathbf{x} \in \mathbb{R}^n \), where \( n \) is the number of dimensions, and can be expressed as

\[
\mathbf{y} = f(\mathbf{x}, \mathbf{x}_i, \mathbf{y}_i).
\]
where $x_i \in \mathbb{R}^{n_x}$ and $y_i$ are the vectors of training inputs and outputs, respectively, which are used to build the surrogate model a priori, and $x$ is the unknown point to predict with the surrogate model. We now describe the surrogate modeling methods available in SMT.

### 3.2.1. LS and QP

The LS method [12] fits a linear model with coefficients $\beta = (\beta_0, \beta_1, ..., \beta_{n_x})$ to minimize the residual sum of the squares between the observed responses in the dataset, and the responses predicted by the linear approximation, i.e.,

$$\min_{\beta} ||X\beta - y||^2_2,$$

where $X = (1, x^{(1)}_1, ..., x^{(n_x)}_1) \vdots (1, x^{(1)}_n, ..., x^{(n_x)}_n)$. The vectors in $X$ have $(n_t \times n_x + 1)$ dimensions and $n_t$ is the number of training points. The square polynomial model is given by Hastie et al. [12]

$$y = X\beta + \epsilon,$$

where $\epsilon$ is a vector of random errors and

$$X = \begin{bmatrix} 1 & x^{(1)}_1 & x^{(1)}_1^2 & \cdots & x^{(1)}_1 x^{(n_x)}_1 & x^{(1)}_2 & \cdots & x^{(1)}_n x^{(n_x)}_n & x^{(1)}_1^2 & \cdots & x^{(1)}_n^2 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 1 & x^{(n_x)}_1 & x^{(n_x)}_1^2 & \cdots & x^{(n_x)}_1 x^{(n_x)}_n & x^{(n_x)}_2 & \cdots & x^{(n_x)}_n x^{(n_x)}_n & x^{(n_x)}_1^2 & \cdots & x^{(n_x)}_n^2 \end{bmatrix}. \tag{4}$$

The vector of estimated polynomial regression coefficients using ordinary LS estimation is

$$\beta = X^T X^{-1} X^T y. \tag{5}$$

### 3.2.2. IDW

The IDW model is an interpolating method where the unknown points are calculated with a weighted average of the sampling points uniformly over the design space.

$$y = \begin{cases} \sum_{i=1}^{n_t} \frac{\beta_i x^{(i)}_1}{\sum_{i=1}^{n_t} \beta_i x^{(i)}_1 x^{(i)}_1}, & \text{if } x \neq x^{(i)}_1 \quad \forall \ i \\ y_1^{(i)}, & \text{if } x = x^{(i)}_1 \text{ for some } i \end{cases} \tag{6}$$
where \(\mathbf{x} \in \mathbb{R}^n\) is the prediction input vector, \(y \in \mathbb{R}\) is the prediction output, \(\mathbf{x}^{(i)} \in \mathbb{R}^n\) is the input vector for the \(i\)th training point, and \(y^{(i)} \in \mathbb{R}\) is the output value for the \(i\)th training point. The weighting function \(\beta\) is defined as
\[
\beta(x^{(i)}, x^{(j)}) = ||x^{(i)} - x^{(j)}||_2^p, 
\]
where \(p\) is a positive real number called the power parameter, which must be strictly greater than one for the derivatives to be continuous.

### 3.2.3. RBF

The RBF surrogate model \([10]\) represents the interpolating function as a linear combination of basis functions, one for each training point. RBF are named as such because the basis functions depend only on the distance from the prediction point to the training point for the basis function. The coefficients of the basis functions are computed during the training stage. RBF are frequently augmented to global polynomials to capture the general trends. The prediction equation for RBF is given by
\[
y = p(x)w_p + \sum_{i=1}^{m} \phi(x, x^{(i)}) w_i, 
\]
where \(\mathbf{x} \in \mathbb{R}^n\) is the prediction input vector, \(y \in \mathbb{R}\) is the prediction output, \(\mathbf{x}^{(i)} \in \mathbb{R}^n\) is the input vector for the \(i\)th training point, \(p(x) \in \mathbb{R}^p\) is the vector mapping the polynomial coefficients to the prediction output, \(\phi(x, x^{(i)}) \in \mathbb{R}^m\) is the vector mapping the RBF coefficients to the prediction output, \(w_p \in \mathbb{R}^m\) is the vector of polynomial coefficients, and \(w_i \in \mathbb{R}^m\) is the vector of RBF coefficients. The coefficients, \(w_p\) and \(w_i\), are computed by solving the following linear augmented Gram system:
\[
\begin{bmatrix}
\phi(x^{(1)}, x^{(1)}) & \cdots & \phi(x^{(1)}, x^{(m)}) & p(x^{(1)})^T \\
\vdots & \ddots & \vdots & \vdots \\
\phi(x^{(m)}, x^{(1)}) & \cdots & \phi(x^{(m)}, x^{(m)}) & p(x^{(m)})^T \\
\end{bmatrix}
\begin{bmatrix}
w_1 \\
\vdots \\
w_m \\
w_p \\
\end{bmatrix}
= 
\begin{bmatrix}
y^{(1)} \\
\vdots \\
y^{(m)} \\
\end{bmatrix}, 
\]
where \(\phi(x^{(i)}, x^{(j)}) = \exp\left(-\frac{||x^{(i)} - x^{(j)}||^2}{d_0}\right)\) is the RBF kernel, and \(d_0\) is a scaling parameter.

### 3.2.4. Kriging-based models

Kriging \([9]\) is an interpolating model that is a linear combination of a known function \(f(x)\) added to a realization of a stochastic process, \(Z(x)\), to obtain
\[
\hat{y} = \sum_{i=1}^{k} \beta_i f_i(x) + Z(x), 
\]
where \(\beta_i\) is the \(i\)th linear regression coefficients and \(k\) is the number of linear regression coefficients to be determined. Here, \(Z(x)\) has a mean of zero and a spatial covariance function given by
where $\sigma^2$ is the process variance and $R$ is the correlation. Two types of correlation functions are available in SMT: the exponential (Ornstein–Uhlenbeck process) and the Gaussian correlation function.

KPLS [4] is a kriging model that uses the PLS method. KPLS is faster than kriging because fewer hyperparameters need to be estimated to achieve high accuracy. This model is suitable for high-dimensional problems owing to the kernel constructed through the PLS method. The PLS method is a well-known tool for high-dimensional problems that searches the direction that maximizes the variance between the input and output variables. This is done by a projection in a smaller space spanned by the principal components. The PLS information is integrated into the kriging correlation matrix to scale the number of inputs by reducing the number of hyperparameters. The number of principal components, $h$, which corresponds to the number of hyperparameters in KPLS is much lower than $n_x$. For example, the PLS-Gaussian correlation function is

\begin{equation}
R(x^0, x^0) = \prod_{i=1}^{n_h} \exp(-\theta_i (x_i^0 - x_i^0)^2),
\end{equation}

where $\theta_i \in \mathbb{R}^+$. The number of hyperparameters $\theta$ is equal to the number of variables $n_h$. They are estimated by maximizing the likelihood function using the gradient-free optimization algorithm COBYLA [13]. These two correlation functions are called abs_exp (exponential) and squar_exp (Gaussian) in SMT. The deterministic term, the sum in the kriging prediction (11), is replaced by a constant, a linear model, or a quadratic model, all of which are available in SMT.

\begin{equation}
\text{cov}[Z(x^0), Z(x^0)] = \sigma^2 R(x^0, x^0),
\end{equation}
KPLS and estimate the hyperparameters expressed in the reduced space with \( h \) dimensions. The second step is to express the estimated hyperparameters in the original space with \( n_t \) dimensions, and then use this as a starting point to optimize the likelihood function of a standard kriging model. The idea here is to guess a “good” initial hyperparameter and apply a gradient-based optimization using a classic kriging kernel. This guess is provided by the KPLS construction: the solutions \((\beta_k^0, ..., \beta_k^h)\) and the PLS-coefficients \((w^{(1)}_k, ..., w^{(h)}_k)\) for \( k = 1, ..., h \). Using the change of variables

\[
\eta_i = \frac{h}{\sum_{l=1}^{h} \beta_l^k w^{(k)}_l x^{(1)}_i},
\]

for \( i = 1, ..., n_t \), we express the initial hyperparameters point in the original space. We demonstrate this using a KPLS-Gaussian correlation function \( R_{\text{KPLS}} \) in the following example [5]:

\[
R_{\text{KPLS}}(x^{(i)}_1, x^{(j)}_1) = \prod_{k=1}^{n_t} \prod_{l=1}^{h} \exp\left(-\beta_l^k w^{(k)}_l (x^{(i)}_l - x^{(j)}_l)^2\right)
= \exp\left(\sum_{l=1}^{h} \sum_{l=1}^{n_t} - \beta_l^k w^{(k)}_l (x^{(i)}_l - x^{(j)}_l)^2\right)
= \exp\left(\sum_{l=1}^{h} \eta_l(x^{(i)}_l - x^{(j)}_l)^2\right)
= \prod_{l=1}^{n_t} \exp\left(-\eta_l(x^{(i)}_l - x^{(j)}_l)^2\right),
\]

where the change of variables is performed on the third line, and the last line is a standard Gaussian correlation function. The hyperparameters point (14) provides a starting point for a gradient-based optimization applied on a standard kriging method.

3.2.5. Gradient-enhanced KPLS models

GE-KPLS [6] is a gradient-enhanced kriging (GEK) model with the PLS approach. GEK is an extension of kriging that exploits gradient information. GEK is usually more accurate than kriging; however, it is not computationally efficient when the number of inputs, the number of sampling points, or both, are high. This is primarily owing to the size of the corresponding correlation matrix that increases in proportion to both the number of inputs and the number of sampling points.

To address these issues, GE-KPLS exploits the gradient information with a slight increase in the size of the correlation matrix and reduces the number of hyperparameters. The key idea of GE-KPLS is to generate a set of approximating points around each sampling point using a first-order Taylor approximation. Then, the PLS method is applied several times, each time on a different number of sampling points with the associated sampling points. Each PLS provides a set of coefficients that contribute to each variable nearby the associated sampling point to the output. Finally, an average of all PLS coefficients is computed to estimate the global influence to the output. Denoting these coefficients by \((w^{(k)}_1, ..., w^{(h)}_h)\), the GE-KPLS Gaussian correlation function is given by

\[
R_{\text{GE-KPLS}}(x^{(i)}_1, x^{(j)}_1) = c^{n_t} \prod_{l=1}^{h} \prod_{k=1}^{h} \exp\left(-\beta_l^k (w^{(k)}_l x^{(i)}_l - w^{(k)}_l x^{(j)}_l)^2\right).
\]

This approach reduces the number of hyperparameters (reduced dimension) from \( n_t \), to \( h \), where \( h \ll n_t \).

As mentioned previously, PLS is applied several times with respect to each sampling point, which provides the influence of each input variable around that point. The idea here is to add only \( m \) approximating points \((m \in [1, n_t])\) around each sampling point. Only the \( m \) highest coefficients given by the first principal component are considered, which usually capture the most useful information. Bouhel and Martins [6] provide more details on this approach.

3.2.6. RMTS Method

The RMTS model is a type of surrogate model for low-dimensional problems with large datasets that has fast prediction capability [7]. The underlying mathematical functions are tensor-product splines, which limits RMTS to up to four-dimensional problems, or five-dimensional problems in certain cases. On the other hand, tensor-product splines enable a fast prediction time that does not increase with the number of training points. Unlike other methods, such as kriging and RBF, RMTS is not susceptible to numerical issues when there is a large number of training points or when points are too close together. The prediction equation for RMTS is given by

\[
y = F(x)w.
\]

where \( x \in R^n \) is the prediction input vector, \( y \in R \) is the prediction output, \( w \in R^w \) is the vector of spline coefficients, and \( F(x) \in R^w \) is the vector mapping the spline coefficients to the prediction output.

RMTS computes the coefficients of the splines, \( w \), by solving an energy minimization problem subject to the conditions that the splines pass through the training points. This is formulated as an unconstrained optimization problem where the objective function consists of a term with the second derivatives of the splines, a term that represents the approximation error for the training points, and another term for regularization. Thus, this optimization problem can be written as

\[
\min_{w} \frac{1}{2} w^T H w + \frac{1}{2} \beta w^T w + \frac{1}{2} \frac{\alpha}{\lambda} \sum_{i=1}^{n} [F(x^{(i)}_l)w - x^{(i)}_l]^2,
\]

where \( x^{(i)}_l \in R^n \) is the input vector for the \( i \)th training point, \( x^{(i)}_l \in R^l \) is the output value for the \( i \)th training point, \( H \in R^{nw \times nw} \) is the matrix of second derivatives, \( F(x^{(i)}_l) \in R^{w} \) is the vector mapping the spline coefficients to the \( i \)th training output, and \( \alpha \) and \( \beta \) are the regularization coefficients.

In problems with a large number of training points relative to the number of spline coefficients, the energy minimization term is not necessary and can be set to zero by setting the \( \text{reg\_cons} \) option to zero. In problems with a small dataset, the energy minimization is necessary. When the true function has high curvature, the energy minimization can be counterproductive. This can be addressed by increasing the quadratic approximation term to one of higher order and using Newton’s method to solve the resulting nonlinear system. The nonlinear formulation is given by

\[
\min_{w} \frac{1}{2} w^T H w + \frac{1}{2} \beta w^T w + \frac{1}{2} \frac{\alpha}{\lambda} \sum_{i=1}^{n} [F(x^{(i)}_l)w - x^{(i)}_l]^p,
\]

where \( p \) is the order set by the \text{approx\_order} option. The number of Newton iterations is specified in the \text{nonlinear\_maxiter} option.

RMTS is implemented in SMT with two choices of splines: B-splines and cubic Hermite splines. RMTB uses B-splines with a uniform knot vector in each dimension. The number of B-spline control points and the B-spline order in each dimension are options that trade off efficiency and precision of the interpolant. For the cubic Hermite splines, RMTC divides the domain into tensor-product cubic elements. For adjacent elements, the values and derivatives are continuous. The number of elements in each dimension is an option that trades off efficiency and precision. B-splines are usually the better choice when training time is the most important factor, whereas cubic Hermite splines are the better choice when the accuracy of the interpolant is most important [7].

3.3. Derivatives

There are three types of derivatives in SMT.

Prediction derivatives \((dy/dx)\) are derivatives of predicted outputs with respect to the inputs at which the model is evaluated. These are computed together with the prediction outputs when the surrogate model is evaluated [14]. These are required for gradient-based optimization algorithms based on surrogate models [15].

Training derivatives \((dy/dx)\) are derivatives of the training outputs with respect to the corresponding inputs. These are provided by the user and are used to improve the model accuracy in GE-KPLS. When the adjoint method is used to compute training derivatives, a high-quality surrogate model can be constructed with a low relative
cost, because the adjoint method computes these derivatives at a cost independent of the number of inputs. Output derivatives (dy/dy) are derivatives of predicted outputs with respect to training outputs, which is a measure of how the prediction changes with a change in training outputs, accounting for the retraining of the surrogate model. These post-training derivatives are used when the surrogate model is trained within an optimization iteration. This feature is not commonly available in other frameworks; however, it is required when the training of the surrogate model is embedded in a gradient-based optimization. In this case, derivatives of the prediction outputs with respect to the training outputs must be computed and combined with derivatives from other parts of the model using, for example, the chain rule.

Given its focus on derivatives, SMT is synergistic with the OpenMDAO framework [16], which is a software framework for gradient-based multidisciplinary analysis and optimization [17–19]. An SMT surrogate model can be a component that is part of a larger model developed in OpenMDAO and can provide the derivatives that OpenMDAO requires from its components to compute the coupled derivatives of the multidisciplinary model.

3.4. Additional surrogate modeling methods

To extend surrogate modeling to higher-level methods, we implemented a component within SMT named extensions. These methods require additional and sometimes different steps than the usual surrogate modeling. For example, multi-fidelity methods combine data generated from different sources, such as coarse and fine mesh solutions. Another example is the mixture of experts (MoE) class of models and expectation maximization [20].

Variable-fidelity modeling (VFM) samples points using both low-fidelity function evaluations that are cheap to evaluate and more costly high-fidelity evaluations. It then uses the differences between the high- and low-fidelity evaluations to construct a bridge function that corrects the low-fidelity model [21]. SMT implements additive and multiplicative bridge functions.

Multi-fidelity kriging (MFK) uses a correlation factor \( \rho(x) \) (constant, linear, or quadratic) and a discrepancy function \( \delta(x) \). The high-fidelity model is given by

\[
y_{\text{high}}(x) = \rho(x)y_{\text{low}}(x) + \delta(x).
\]

SMT follows the formulation proposed by Le Gratiet [22], which is recursive and can be easily extended to multiple levels of fidelity.

4. Example and benchmark functions

SMT contains a library of analytical and engineering problems that can be used for instructional or benchmarking purposes. The analytical functions currently implemented in SMT are the sphere, Branin, \( L_p \) norm, Rosenbrock, and tensor-product functions, all of which are detailed in Appendix A. The engineering functions currently implemented in SMT are the cantilever beam, robot arm, torsion vibration, water flow, welded beam, and wing weight problems (described in Appendix B).

Figure 4 shows the implementation of the Sphere class, which inherits from the Problem class. This function is a simple example that serves as a good first test for newly developed surrogate modeling and surrogate-based optimization methods. This function is a continuous and convex function where the global minimum is at the origin.

Figure 5 shows an example of the use of RMTC within SMT for the robot arm function [23]. This function is scalable; however, the number of dimensions must be an even number (we use two dimensions in this example). This function gives the position of a robot arm, which is made by multiple segments, in a two-dimensional space where the shoulder of the arm is fixed at the origin. This function is highly nonlinear, and the use of training derivatives samples is particularly beneficial in this case [6].

5. Applications

In this section, we describe two applications that highlight the unique features of SMT. We also provide an overview of the main previous applications realized using SMT. The first application is the computation and validation of the prediction derivatives for an airfoil analysis tool, where we apply GE-KPLS. The second application describes a practical use for output derivatives, where we use RMTS to compute the outputs derivatives for a surrogate model that is dynamically trained with an optimization iteration.

5.1. Airfoil analysis and shape optimization tool

Li et al. [14] used the GE-KPLS implementation in SMT with an MoE technique to develop a data-driven approach to airfoil and shape optimization based on computational fluid dynamics (CFD) simulations. They used a database of 1100 existing airfoils\(^8\) and enriched the database with 100,000 more generated airfoils. The data was used to create a surrogate model of force coefficients (lift, drag, and moment) with respect to flight speed (Mach number), the angle of attack, and airfoil shape variables. The shape design variables consisted of shape modes obtained by singular-value decomposition of the existing airfoil shapes. The approach proved successful, enabling shape optimization in 2 sec using a personal computer with an error of less than 0.25% for subsonic flight conditions. This is thousands of times faster than optimization using direct calls to CFD.

To validate the prediction derivatives provided by SMT, we built a surrogate model of the airfoil drag coefficient similar to the one above using the following four inputs: first thickness mode \( t_1 \), first camber mode \( c_1 \), Mach number \( M \), and angle of attack \( \alpha \). We compute the partial derivatives of the drag coefficient \( C_D \) with respect to these four variables. The range of the first thickness and camber modes are such that we cover 55% of the airfoil database. The range of \( M \) is \([0.3,0.4]\) and the range of \( \alpha \) is \([0,4]\). To train the surrogate model, we use 20 training points with their respective training derivatives. We use the GE-KPLS model, where we provide training derivatives using an adjoint approach [24,25]. To validate the prediction derivatives of the constructed surrogate model, we generate 100 new points different from the training points and compute the relative error on those points using

\[
\varepsilon = \frac{\left| \frac{\partial C_D}{\partial x} \right| - |\partial C_D/\partial x|}{|\partial C_D/\partial x|},
\]

where \( \partial C_D/\partial x \) is the vector of 100 surrogate model prediction derivatives with respect to one of the four variables \( x = t, c, M, \alpha \), \( \partial C_D/\partial x \) is the vector of corresponding reference derivative values with respect to \( x \), and \(|\cdot|_2\) is the \( L_2 \) norm. The resulting relative errors corresponding to each variable (listed in Table 2) show that the surrogate model provides an accurate estimate of the derivatives with a relative expected precision.

\(^7\) https://github.com/mdolab/adflow

\(^8\) http://webfoil.engin.umich.edu
Using the constructed surrogate, we compare two gradient-based optimization algorithms:

- **Optimization 1** uses the surrogate model derivatives provided by SMT;
- **Optimization 2** uses finite-difference derivatives.

The optimization problem is to minimize the drag coefficient ($C_d$) subject to a lift constraint ($C_l = 0.5$). Optimization 1 achieved a slightly lower drag (0.03%) than Optimization 2 using 21 times fewer evaluations (Fig. 6).

### 5.2. Aircraft design optimization considering dynamics

The design optimization of commercial aircraft using CFD is typically done by modeling the aircraft performance at a small number of representative flight conditions [26,27]. This simplification is made because it would be prohibitively expensive to simulate the full, discretized mission profile using CFD at all points. However, full-mission simulation is sometimes necessary, such as when the flight is short or when considering morphing aircraft designs [28,29]. This requires a surrogate model of the aerodynamic performance as a function of a small number of parameters such as flight speed and altitude. In a design optimization context, the surrogate model must be retrained each optimization iteration because the aircraft design changes from iteration to iteration as the shape is being optimized.

This is the situation in an aircraft allocation-mission-design optimization problem recently solved using RMTS [30]. In this work, the optimization problem maximized airline profit by optimizing the twist, span, sweep, and shape of the wing of a next-generation commercial aircraft. The profit was computed by simulating the fuel efficiency and flight time for the aircraft on a set of routes operated by the hypothetical airline. To do this, a surrogate model was generated for the wing lift and drag coefficients ($C_L$ and $C_D$) as a function of the angle of attack ($\alpha$), Mach number ($M$), and altitude ($h$). In each optimization iteration, the training $C_L$ and $C_D$ values were computed at a series of points in $M-\alpha-h$ space using CFD. The training outputs were used to retrain the RMTS surrogate model, and the mission simulations for all the airline routes were performed using inexpensive evaluations of the trained surrogate model, as shown in Fig. 7. However, because the overall optimization problem was solved using a gradient-based algorithm, we required derivatives of the prediction outputs ($C_L$ and $C_D$ values at the fixed $M-\alpha-h$ points at which we evaluated the surrogate) with respect to the training outputs ($C_L$ and $C_D$ values at the fixed $M-\alpha-h$ points where the training points were located).

Figure 8 shows the surrogate model with the altitude axis eliminated by projecting onto the other two axes, where $C_L$ replaces $\alpha$ as one of the inputs. The horizontal and vertical axes represent the inputs. The red points are the training points and the black points are the prediction outputs.
Therefore, the output derivatives are the derivatives of the outputs at the black points with respect to the training outputs at the red points.

5.3. Other applications

In addition to the two applications we just described, SMT has been used to solve other engineering problems. We summarize these applications in Table 3. The number of input variables ranges from 2 to 99. This demonstrates SMT’s ability to solve different engineering problems of various complexities.

6. Conclusions

SMT is unique compared with existing surrogate modeling libraries in that it is designed from the ground up to handle derivative information effectively. The derivative information includes training derivatives used for gradient-enhanced modeling, prediction derivatives used for surrogate-model-based optimization, and derivatives with respect to the training data used when the optimization loop includes reconstructing the surrogate model. However, SMT does not need to involve derivatives and provides a simple general interface to various surrogate modeling techniques. Together with the provided problems, SMT is a convenient framework to develop new methods for sampling and surrogate modeling, and then benchmarking them against existing methods.

SMT arose from our own research needs, namely the development of the GE-KPLS and RMTS surrogate models, which require the handling of derivatives. GE-KPLS is a method we recently developed that scales much better with the number of inputs than the other methods, thanks in part to the use of the derivative information. RMTS is another method that we recently developed that is best suited to low-dimensional problems and handles both structured and unstructured training points.

SMT is distributed through an open-source license and is freely available online\(^9\). We provide documentation that caters to both users and potential developers.\(^10\) SMT allows users collaborating on the same

---

**Table 2**

Relative error of the prediction derivative of \(C_d\) with respect to \(c_1\), \(c_2\), \(M\), and \(\alpha\).

<table>
<thead>
<tr>
<th>(\frac{\partial C_d}{\partial c_1})</th>
<th>(\frac{\partial C_d}{\partial c_2})</th>
<th>(\frac{\partial C_d}{\partial M})</th>
<th>(\frac{\partial C_d}{\partial \alpha})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.010</td>
<td>0.007</td>
<td>0.009</td>
<td>0.002</td>
</tr>
</tbody>
</table>

\(^9\)https://github.com/SMTorg/SMT
\(^10\)http://smt.readthedocs.io/en/latest/

---

**Fig. 5.** Example of use of RMTC within SMT on the robot arm function [23].
project to have a common surrogate modeling tool that facilitates the exchange of methods and reproducibility of results.

SMT is a recently developed tool and, so far, it has primarily been used in aerospace engineering applications. However, this tool is useful to anyone needing to use or develop surrogate modeling techniques, regardless of the application. In particular, given the advantages of the newer surrogate modeling techniques (GE-KPLS and RMTS), using SMT is particularly attractive.

Acknowledgments

This work was partially supported by the Air Force Office of Scientific Research (AFOSR) MURI on “Managing multiple information sources of multi-physics systems,” Program Officer Jean–Luc Cambier, Award Number FA9550-15-1-0038. This work is part of the activities of ONERA–ISAE–ENAC joint research group and was partially supported by an ONERA internal project on multi-fidelity methods (MUFIN). The authors would like to thank Rémi Vauclin and Mostafa Meliani for the implementation of MFK, and Remy Priem for the implementation of MoE.
Appendix A. Analytical functions implemented in SMT

In the following descriptions, \( n \) is the number of dimensions.

**Sphere.** The sphere function is quadratic, continuous, convex, and unimodal. It is given by

\[
\sum_{i=1}^{n} x_i^2, \quad -10 \leq x_i \leq 10, \quad \text{for } i = 1, \ldots, n.
\]

**Branin** [35]. The Branin function is commonly used in optimization and has three global minima. It is given by

\[
f(x) = \left( a x_2 - \frac{5.1}{4\pi^2} x_1^2 + \frac{5}{\pi} x_1 - 6 \right)^2 + 10 \left( 1 - \frac{1}{8\pi} \right) \cos(x_1) + 10.
\]
where \(x = (x_1, x_2)\) with \(-5 \leq x_1 \leq 10\) and \(0 \leq x_2 \leq 15\). Rosenbrock [36]. The Rosenbrock function is a continuous, nonlinear, and non-convex function and usually used in optimization. The minimum of this function is situated in a parabolic-shaped flat valley and is given by

\[
\sum_{i=1}^{n_x} ((x_{i+1} - x_i)^2 + (x_i - 1)^2), \quad -2 \leq x_i \leq 2, \quad \text{for } i = 1, ..., n_x.
\]

Tensor-product. The tensor-product function approximates a step function, which causes oscillations with some surrogate models [7], and is given by

\[
\prod_{i=1}^{n_x} \cos(a_i \pi x_i), \quad -1 \leq x_i \leq 1, \quad \text{for } i = 1, ..., n_x,
\]

or

\[
\prod_{i=1}^{n_x} \exp(x_i), \quad -1 \leq x_i \leq 1, \quad \text{for } i = 1, ..., n_x,
\]

or

\[
\prod_{i=1}^{n_x} \tanh(x_i), \quad -1 \leq x_i \leq 1, \quad \text{for } i = 1, ..., n_x,
\]

or

\[
\prod_{i=1}^{n_x} \exp(-2x_i^2), \quad -1 \leq x_i \leq 1, \quad \text{for } i = 1, ..., n_x.
\]

Appendix B. Engineering problems implemented in SMT

**Cantilever beam** [37]

This function models a simple uniform cantilever beam with vertical and horizontal loads

\[
\frac{50}{600} \sum_{j=1}^{12} \left[ \frac{12}{b_j h_j} \left( \left( \sum_{i=1}^{4} l_i \right) - \left( \sum_{j=1}^{12} l_j \right) \right) \right],
\]

where \(b_j \in [0.01, 0.05], \quad h_j \in [0.3, 0.65], \quad l_i \in [0.5, 1]\).

**Robot arm** [23]

This function gives the position of a robot arm

\[
\left( \sum_{j=1}^{4} L_i \cos \left[ \sum_{i=1}^{4} \theta_j \right] \right)^2 + \left( \sum_{i=1}^{4} L_i \sin \left[ \sum_{i=1}^{4} \theta_j \right] \right)^2,
\]

where \(L_i \in [0, 1]\) for \(i = 1, ..., 4\) and \(\theta_j \in [0, 2\pi]\) for \(j = 1, ..., 4\).

**Torsion vibration** [38]

This function gives the low natural frequency of a torsion problem

\[
\frac{1}{2a} \sqrt{-b - \sqrt{b^2 - 4ac}},
\]

where \(K_1 = \frac{\pi^2 d_1}{4d_3}, \quad M_1 = \frac{\pi^4 d_3}{8d_1}, \quad J_1 = 0.5M_1 d_3, \quad a = 1, \quad b = \left( \frac{K_1 + K_3}{2} \right)^2 + \left( \frac{K_1 + K_3}{2} \right), \quad c = \frac{K_1 K_2 + K_1 K_3 + K_2 K_3}{2d_3}, \quad \text{for } d_1 \in [1.8, 2.2], \quad L_1 \in [9, 11], \quad G_1 \in [105300000, 128700000], \quad d_2 \in [1.638, 2.002], \quad L_2 \in [10.8, 13.2], \quad G_2 \in [5580000, 6820000], \quad d_3 \in [2.025, 2.475], \quad L_3 \in [7.2, 8.8], \quad G_3 \in [3510000, 4290000], \quad D_1 \in [10.8, 13.2], \quad t_1 \in [2.7, 3.3], \quad p_1 \in [0.252, 0.308], \quad D_2 \in [12.6, 15.4], \quad t_2 \in [3.6, 4.4], \quad \text{and } p_2 \in [0.09, 0.11].

**Water flow** [39]

This function characterizes the flow of water through a borehole that is drilled from the ground surface through two aquifers

\[
\ln \left( \frac{T_r}{T_0} \right) + \frac{2L_T}{\ln \left( \frac{T_r}{T_0} \right) K_w + \frac{L_T}{T_r}},
\]

where \(0.05 \leq w_r \leq 0.15, \quad 100 \leq r \leq 50000, \quad 63070 \leq T_r \leq 115600, \quad 990 \leq H_r \leq 1110, \quad 63.1 \leq T_l \leq 116, \quad 700 \leq H_l \leq 820, \quad 1120 \leq L \leq 1680, \quad \text{and } 9855 \leq K_w \leq 12045.\)
The stress of a welded beam problem is given by
\[
\sqrt{\frac{1}{2.5(2+(h+r)^3)}} \cdot \left( 6000(1.4+0.5S) \cdot 0.08 + \frac{0.25}{h} \right),
\]
where \( r = \frac{6000}{276} \), \( t = \frac{6000(1.4+S)}{1700+2200} \), for \( h \in [0.125, 1] \), and \( l \in [5, 10] \).

Wing weight [35]

The estimate of the weight of a light aircraft wing is given by
\[
0.036S^2 \frac{88}{\cos^2 A} \left( \frac{A}{\cos^2 A} \right)^{0.0064} 1000c_{10}^2 \frac{A}{\cos^2 A} \left( N, W_0 \right)^{0.49} + S_0 W_p,
\]
where \( 150 \leq S_0 \leq 200 \), \( 220 \leq W_p \leq 300 \), \( 6 \leq A \leq 10 \), \( -10 \leq \alpha \leq 10 \), \( 16 \leq q \leq 45 \), \( 0.5 \leq t_c \leq 1 \), \( 0.08 \leq \xi \leq 0.18 \), \( 2.5 \leq N_t \leq 6 \), \( 1700 \leq W_p \leq 25000 \), and \( 0.025 \leq W_p \leq 0.08 \).

References


