Smoothing of Surrogate Models

Bernstein Alexander, Belyaev Mikhail, Burnaev Evgeny, Yanovich Yuri
Institute for Information Transmission Problems, Institute for System Analysis, Moscow Institute of Physics and Technology, DATADVANCE
a.bernstein@cpt-ran.ru, {mikhail.belyaev,evgeny.burnaev,yurij.yanovich}@datadvance.net

Abstract

In the present work methods for controlling smoothness of surrogate models is proposed and justified. Such control of smoothness is important in surrogate based optimization process.

1. Introduction

In [1, 2] a universal surface fitting algorithm has been developed. The algorithm allows to construct a surrogate model of a given full scale model, which could be given either as a black box function or as a data sample. In the former case the tool allows to construct a surrogate model with a specified accuracy, in the later case – with the best possible accuracy.

If the resulting surrogate model is to be used as either objective function or constraint of some optimization problem and if the gradient based method is to be used then a question of obtaining the gradient of this surrogate model do arises. Of course, it is desirable to have certain smoothness properties of both surrogate model and its gradient.

Another and much more important reason to require specific smoothness properties of surrogate models is the following. Most of engineer-level objective functions are intrinsically noisy. The often random noise is caused by various sources, e.g. it might be because of finite (perhaps, rather crude) precision of physical experiments or because of numerical modeling flaws (non-convergence of algorithms, their finite precision etc.). Leaving aside the question of noise origin (although it might be helpful to classify various sources) one could argue that in most cases experienced engineers could identify unreliable data points by analyzing the history and convergence of corresponding simulations supplemented with all his/her engineering wisdom. Therefore it is possible to rank all available data points with a kind of reliability estimate which might be very helpful in noise reduction technique. Indeed, the most promising approach to reduce both the influence of noise and the number of function evaluations seems to be the systematic exploitation of surrogate models. The prime observation is that any surrogate model one could construct in the vicinity of current guess does not reproduce the original data set exactly. Instead (depending on details of chosen model) it reproduces some regularized approximation to the original data reflecting the engineer's a priori knowledge of the model considered. Thus the art of surrogate modeling does not reduce to simply taking the most universal type of approximation algorithm, in order to be useful the surrogate model should provide the capabilities to control the smoothness properties of the model. For instance, if at engineer-level it is known that the model response (no matter how complex its parameter dependence might be) is continuous, the continuous surrogate model could greatly reduce the noise present in data. To the contrary, taking more "universal" surrogate model supporting discontinuities of various kinds makes no much sense with continuous but noisy objective function(s), physics is ought to be lost but noise will be mostly preserved by surrogate model.

All the existing methods for constructing approximations are unable to obtain the simultaneous solution of two interrelated problems – ensuring both the accuracy of the model and the smoothness of the model (except for trivial cases).

In the present work a new algorithm for construction of smoothed surrogate models has been developed. This algorithm ensures not only closeness of original function and its approximation but also ensures smoothness of gradient of the approximation.

Elaborated algorithm is intended for the automatic generation of smoothed surrogate models, which can be used to calculate approximate smooth value and approximate smooth gradient for the true values of the characteristics and their gradient correspondingly.

The article is structured as follows:

1. Section 2 outlines the main goals and particular details of mathematical modeling, surrogate modeling, their use in design optimization and gives the statement of the smooth surrogate modeling problem.

2. Section 3 contains the main theoretical approaches to filtering (smoothing) as well as description of a filtering (smoothing) algorithm based on these theoretical approaches and used for construction of smoothed surrogate models.

3. Section 4 gives example of application of elaborated algorithm to typical test function.

2. Problem statement

2.1. Mathematical modeling

Comparison of various technical solutions regarding an object's structure, parameters, operating principles and other aspects is an essential step in any design process. Rapid progress in mathematical modelling and computer engineering has made it possible to investigate plenty of design alternatives with different object configurations and parameters, predict an object's characteristics and find the best or most balanced solutions without any full-scale experiments. Therefore, mathematical modelling which involves series of computational experiments for the investigation of analytical models of the object and its environment has become one of the most common methods of analysis and optimization of an engineering object's structure (see [3, 5, 6]).

Let us denote by:

- X some p-dimensional vector describing object design (for example, X is a p-dimensional description of aircraft layout),

- $Y = f_M(X)$ some mathematical model used for calculation of q-dimensional vector of an object characteristics with description X. It is supposed that the value of the function $f_M(X)$ is noiseless. In the sequel without loss of generality let us consider the case $q = 1$.

Usually the model $Y = f_M(X)$ is used in some optimization process, for example, its minimum is searched. This minimum is a candidate for optimal design. For example, in order to find the optimal airfoil shape X for an aircraft wing, an engineer simulates the air flow around the wing for different shape variables (length, curvature, material, ...) and minimize the value of the drag coefficient $f_M(X)$, i.e. search for p-dimensional

$$X_0 = \arg\min_{X \in \mathbf{X}} f_M(X), \tag{1}$$

where $\mathbf{X}$ is a design space. Design space can also be defined as $\mathbf{X} = \{X : g(X) \le 0\}$ for some vector function $g(X)$. The most widely used method for solution of this problem consists of two phases (probably, repeated iteratively) and can be described as follows:

- Phase during which local minimum of $f_M(X)$ is obtained (LM-phase).

- Phase during which transition to a search for another local minimum is done in order to find global minimum (GM-phase).

During the LM-phase a sequence of points $X_1, X_2, X_3, \ldots$ is constructed such that $f_M(X_1) > f_M(X_2) > f_M(X_3) > \ldots$ and $\{X_i, i = 1, 2, \ldots\}$ converges to the local minimum X_0 (the construction process is stopped when the decrease of $f_M(X)$ is stopped). It is obvious that for some intermediate point X_k we should be able to find new value X_{k+1} with smaller value of the response function $f_M(X)$. Using theory of optimization we get that if $f'_M(X_k)$ is a non-zero value of the gradient of the function $f_M(X)$ for $X = X_k$, then there exists such number $T(X_k) > 0$ that

$$f_M(X_k + t f'_M(X_k)) < f_M(X_k) \tag{2}$$

for $0 < t < T(X_k)$. Thus during the LM-phase we should

- define multidimensional direction $e(X_k)$ of decrease of the function $f_M(X)$ (for example, using the gradient, i.e. $e(X_k) = f'_M(X_k)$),

- define $t \in (0, T(X_k))$ such that for $X_{k+1} = X_k + t f'_M(X_k)$ the value of $f_M(X_{k+1})$ is the smallest possible.

Even if some additional functional constraints $g(X) \le 0$ should be taken into account during the optimization then the general scheme given above is not changed, only some additional restrictions are imposed on selected multidimensional direction $e(X_k)$.

Thus in order to accomplish LM-phase we should be able to calculate the gradient $L(f_M)(X) = f'_M(X)$ for the function $f_M(X)$. In case there is no explicit formulae for the gradient $L(f_M)$ numerical approximation of the functional $L(f_M)$ is evaluated using another functional $L_s(f_M)$ depending on the finite number $n(s)$ of the values of $f_M(X)$, such that

$$L_s(f_M) \to L(f_M) \tag{3}$$

for $s \to 0$. In particular, for the case of $L(f_M)(X) = f'_M(X)$ the following functional is used as $L_s(f_M)$

$$L_s(f_M) = (f_M(X + s) - f_M(X))/s. \tag{4}$$

At the same time when using (4) it is taken into account that all calculations (computation of the function $f_M(X)$ value, division operation, etc.) are done with some finite accuracy (for example, due to finite digit capacity of the computer). In particular, this fact defines the choice of the parameter s.

2.2. Surrogate Modeling

Traditionally, mathematical model $Y = f_M(X)$ is based on "process physics" and describes physical processes and phenomena occurring in the course of an object's operation by complex partial differential equations with boundary conditions, for which in most cases nothing is known either about the theorems of existence or uniqueness of the solution or the dependence of the solution from the parameters or boundary conditions. These equations are solved using complicated numerical methods that require significant computing resources and a lot of effort for preparing input data and computational meshes. Thus model based on process physics has a limited scope of application, especially at the early (conceptual) design phase where a lot of various design alternatives have to be considered and making the wrong choice can have far-reaching consequences.

One way of alleviating this burden is by constructing approximation models, known as surrogate models or metamodels, that mimic the behavior of the simulation model as closely as possible while being computationally cheaper to evaluate (see [3, 5, 6]). Surrogate models are constructed using a data-driven, bottom-up approach. The exact, inner working of the simulation code is not assumed to be known (or even understood), solely the input-output behavior is important. A model is constructed based on modeling the response of the simulator using a limited number of intelligently chosen data points of expensive experiments and/or simulations, i.e. using some training sample

$$S_{train} = \{X_i, Y_i = f_M(X_i), i = 1, \ldots, N\} \quad (5)$$

such function (surrogate model) $f_{SM}(X)$ is constructed that

$$f_{SM}(X) \approx f_M(X). \quad (6)$$

The surrogate model $f_{SM}(X)$ is cheap to evaluate, so it is used in optimization process instead of the initial function $f_M(X)$ to predict designs with promising performance. The remaining budget of expensive experiments/simulations are run for these candidate designs to check them. The process usually takes the form of the following search/update procedure:

1. Initial sample S_{train} is constructed.
2. Initial surrogate model $f_{SM}(X)$ is constructed.
3. Constructed surrogate model $f_{SM}(X)$ is used in optimization process, for example, its minimum is searched. Obtained solution of the optimization problem is a candidate for optimal design.
4. Experiments/simulations are done at new location(s) found during the previous step and added to the existing sample S_{train}.
5. Steps 2 to 4 are iterated until out of time or "good enough" design is found.

2.3. Surrogate model for the gradient

In order to obtain accurate candidates (for optimal design) as a solution of the corresponding optimization problem (see Step 3 in the previous section) we should provide

- the maximal possible accuracy $f_{SM}(X) \approx f_M(X)$ necessary for replacement of the model $f_M(X)$ by the surrogate model $f_{SM}(X)$,
- the surrogate model $(f'_M(X))_{SM}$ for the gradient $f'_M(X)$ necessary for accomplishment of LM-phase (see Step 3 in the previous section), such that $(f'_M(X))_{SM} \approx f'_M(X)$.

Suppose that we can obtain both the value of the model $f_M(X)$ and the value of its gradient $f'_M(X)$, i.e. for each input vector X we can calculate the vector $(f_M(X), f'_M(X))$. Thus, using the extended training sample

$$S^*_{train} = \{X_i, (Y_i = f_M(X_i), Y'_i = f'_M(X_i)), i = 1, \ldots, N\}, \quad (7)$$

we can construct the surrogate model $f^*_{SM}(X)$ with vector output $(f_{SM}(X), (f'_M(X))_{SM})$, where the first component $f_{SM}(X)$ approximates the model $f_M(X)$ and the second component $(f'_M(X))_{SM}$ approximates the gradient $f'_M(X)$ of the model (this means that we use the Sobolev norm to measure the error of approximation). However if the value of the gradient $f'_M(X)$ is unknown then we simply do not have data for construction of such extended surrogate model $f^*_{SM}(X)$ (there is no data about the object for which we want to construct a surrogate model).

Numerical estimation of the gradient (calculation of finite differences) for construction of extended sample (7) is rather problematic. Finite differences (4) can have satisfactory quality only for small values of all components of the vector s. Such requirement is almost impossible to fulfill in multidimensional case due to high sparseness of the input space. Even if we can generate new data then numerical estimation of the gradient is very computationally intensive. Moreover, numerically estimated gradient is rather noisy since it highly depends on the parameters of the used numerical method.

Thus in general we are not able to construct a surrogate model $(f'_M(X))_{SM}$ for the gradient $f'_M(X)$ and the only source of information about the function $f_M(X)$ and its gradient $f'_M(X)$ is the available surrogate model $f_{SM}(X)$.

Let us consider more thoroughly how $f_{SM}(X)$ is related to $f_M(X)$. The surrogate model $f_{SM}(X)$ is considered to be good when there is no dependency between residuals $f_{SM}(X) - f_M(X)$ for close input points (otherwise subsequent improvement of the accuracy of the surrogate model $f_{SM}(X)$ is possible), i.e. in the limit the residuals $f_{SM}(X) - f_M(X)$ behaves like white noise. Thus

- the surrogate model $f_{SM}(X)$ can be represented at least approximately as

$$f_{SM}(X) = f_M(X) + \varepsilon(X), \quad (8)$$

where $\varepsilon(X)$ are residuals of the surrogate model;

- we consider $f_M(X)$ to be rather smooth function without noise since the function $f_M(X)$ represents some real physical process;
- the error $\varepsilon(X)$ of the surrogate model is considered to be small compared to the value of $f_M(X)$ (otherwise the surrogate model $f_{SM}(X)$ is considered to have low accuracy and should not be used in optimization process);
- $\varepsilon(X)$ is modeled (at least approximately) by some random process with zero mean, being uncorrelated or having some dependence structure.

Rather often engineer-level models are intrinsically noisy. Usually random noise is caused by various sources, e.g. it might be because of finite (perhaps, rather crude) precision of physical experiments or because of numerical modeling flaws (non-convergence of algorithms, their finite precision etc.). In such case the sample $S_{train} = \{X_i, Y_i, i = 1, \ldots, N\}$ will contain such random noise, i.e. $Y_i = f_M(X_i) + \delta_i, i = 1, \ldots, N$, for some random noise process δ. In such case in (8) we consider the function $f_M(X)$ to be an ideal unknown function without noise, representing true physical process. Thus the source of the error term $\varepsilon(X)$ can be not only the discrepancy between the surrogate model $f_{SM}(X)$ and the initial model $f_M(X)$ (appeared due to the finite size of the sample S_{train}), but also some intrinsic noise presented in the sample S_{train}.

Model (8) means that we estimate the value of the function $f_M(X)$ with a random error, i.e. for the same input X we can obtain measurements with different values of the error $\varepsilon(X)$. Thus the function $f_M(X)$ can be considered as a trend and the error $\varepsilon(X)$ can be considered as a diffusion. In general the error term $\varepsilon(X)$ cannot be reduced to zero by the increase of calculation accuracy (for example by means of increase of sample size or by increase of computer digit capacity).

Let us now consider more thoroughly how the gradient $f'_{SM}(X)$ of $f_{SM}(X)$ is related to the gradient $f'_M(X)$ of $f_M(X)$. Given the model (8) the problem of the gradient $f'_M(X)$ estimation is reduced to the estimation of the functional $L(f_M)(X) = f'_M(X)$ using "noisy" data $f_{SM}(X)$. Due to the noise term $\varepsilon(X)$ in (8) convergence of the estimate to the true value should be considered in the mean square sense in contrast to the usual deterministic convergence in (3). Thus we should construct the functional $L_s(f_M)$ depending on the finite number $n(s)$ of the noisy values of $f_{SM}(X)$, such that

$$\mathbb{E}(L_s(f_{SM}) - L(f_M))^2 \to 0 \quad (9)$$

for $s \to 0$, where $\mathbb{E}$ denotes mathematical expectation with respect to the probability law of $\varepsilon(X)$ and distribution of the input vector X (for example, if the input domain is bounded then we can consider uniform distribution as such distribution of input vector X). Usual procedure based on finite differences (4) cannot be used for estimation of the gradient, since it is designed for noise-free case. Actually, for $L_s(f_{SM}) = (f_{SM}(X+s) - f_{SM}(X))/s$ it can be easily proved that for $s \to 0$

$$\mathbb{E}(L_s(f_{SM}) - L(f_M))^2 = (s^2 \cdot f''_M(X^*))/2 + (\mathbf{Var}(\varepsilon(X+s) - \varepsilon(X)))/s^2, \quad (10)$$

where **Var** denotes variance with respect to the probability law of $\varepsilon(X)$ and distribution of the input vector X, X^* is a some point belonging to the segment with endpoints X and $X+s$. For the most of standard situations (measurements are independent or the noise $\varepsilon(X)$ is white etc.) even if the variance $\mathbf{Var}(\varepsilon(X+s) - \varepsilon(X))$ tends to zero, it's convergence rate is not faster than O(s). Thus calculation of the gradient using finite differences applied to $f_{SM}(X)$ is prohibitive for the case of noisy function, since $\mathbb{E}(L_s(f_{SM}) - L(f_M))^2 \to \infty$ for $s \to 0$.

2.4. Optimization of the surrogate model

Due to the representation (8) the typical examples of the surrogate model behavior can be described as follows:

(1) significantly varied "saw-toothed" regions corresponding to a big number of close local minima with similar values of the response function $f_{SM}(X)$ (at the same time teeth of the saw can be sufficiently blunt to ensure the existence of the gradient).

(2) significantly varied "flat" segments corresponding to a big number of close "continuous" local minima with similar values of the response function $f_{SM}(X)$ (at the same time flat segments can be smoothly joined in order to ensure the existence of the gradient). This is precisely the behavior of some local methods (for example, k-nearest neighbor method from modeFrontier [7]).

General optimization algorithms used for accomplishment of the LM-phase (see Step 3 in the section 2.2) are usually designed for rather complex but though rather "good" functions. By "good" functions here we mean functions with either one minimum or several minima located sufficiently distant from each other. Moreover, often these minima correspond to significantly different output values of the response function. Thus typical all-purpose optimization algorithms are not very well tailored for optimization of surrogate models, since for example

- In (1) (see above) even for precise value of the gradient $e(X) = f'_{SM}(X)$ the value of $T(X)$ (see (2)) will be very small, resulting in small optimization steps. Thus effectiveness of LM-phase will be decreased and GM-phase (see section 2.1) will become incredibly difficult.

- In (2) the gradient will be zero and optimization algorithm will not work.

Thus the gradient $f'_{SM}(X)$ has "bad" behavior and can significantly differ from the gradient $f'_M(X)$ (see also section 2.3 for details), so optimization of the surrogate model $f_{SM}(X)$ "as is" using the gradient $f'_{SM}(X)$ and typical all-purpose optimization algorithms is complicated and can lead to inaccurate candidates for optimal design.

2.5. Statement of the smooth high dimensional approximation problem

As it follows from the previous sections it is natural to consider the surrogate model $f_{SM}(X)$ to be analogue of the noisy model (8) producing noisy data $f_{SM}(X)$ with trend $f_M(X)$ and diffusion $\varepsilon(X)$.

The main problem of noisy data analysis is the so-called filtering (de noising) problem consisting in estimation of various characteristics of the trend $f_M(X)$ (output values for particular input values, behavior of $f_M(X)$ for changing input vector X, described by the gradient $f'_M(X)$, etc.) using noisy data (8) (see for details [8, 4]).

Thus relying only on "noisy" data $f_{SM}(X)$ we can consider only the problem of trend $f_M(X)$ estimation (including trend estimation for the gradient $f'_M(X)$). At the same time exactly the initial data is smoothed rather than values of non-robust functions of the initial data (for example, numerical estimates of the gradient are non-robust functions of the initial data, i.e. they can produce huge estimation errors especially in multi-dimensional case).

Therefore in order to construct the object for surrogate modeling (of the gradient in the considered case) it is necessary to use filtering (for the initial model $f_{SM}(X)$) and state the problem of surrogate model construction for the trend (smoothed function).

Thus the statement of the smooth high dimensional approximation problem is the following:

1. Surrogate model $f_{SM}(X)$ is intended for the most possible precise approximation of the initial model $f_M(X)$ and the more accurate the surrogate model $f_{SM}(X)$ reproduces the initial model $f_M(X)$ the better it is considered to be.

2. Problems of surrogate model construction for the gradient $f'_M(X)$, optimization of the initial model $f_M(X)$ etc. should be considered if not only pointwise prediction of the values of the initial model $f_M(X)$ should be obtained but also its behavior for changing input vector X (characteristics of the variation of $f_M(X)$ along different directions of the input vector X, minimum point of $f_M(X)$ etc.) should be investigated.

3. Since the values of the initial function $f_M(X)$ are known only for some finite set of input vectors X then well-posed problem statement of the gradient estimation $f'_M(X)$ can be set only for the smoothed surrogate model. In particular, the problem of the gradient estimation consists in estimation of the gradient of the trend $f_M(X)$ (see (8)) obtained by its filtration from the noisy data $f_{SM}(X)$.

4. The function $f_{SM}(X)$ should be smoothed (the trend should be estimated) in order to facilitate optimization. The gradient of the trend should be used in order to accomplish LM-phase of the optimization process and obtain accurate candidates for optimal design (see the section 2.2).

5. Let us denote by

   - $f_{SM,s}(X)$ the estimate of the trend $f_M(X)$, obtained by filtering of $f_{SM}(X)$, where s denotes some smoothing parameter ($f_{SM,s}(X)$ is a smoothed/de-noised surrogate model). Smoothing parameter s can be interpreted, for example, as the size of the neighborhood of the given input point X, over which the output of $f_{SM}(X)$ is averaged in order to obtain the value $f_{SM,s}(X)$.
   - $L(f_{SM,s})(X) = f'_{SM,s}(X)$ and $L(f_{SM})(X) = f'_{SM}(X)$ the gradients of $f_{SM,s}(X)$ and $f_{SM}(X)$ correspondingly. Suppose that the model (8) holds true, the function $f_M(X)$ is smooth enough (at least $f_M(X)$ has finite second derivatives), then the estimate $f_{SM,s}(X)$ should have the following asymptotical properties:

   $$\mathbb{E}(f_{SM,s} - f_M)^2 \to 0, \tag{11}$$
   $$\mathbb{E}(L(f_{SM,s}) - L(f_M))^2 \to 0, \tag{12}$$

   for $s \to 0$.

Remark. It is natural to obtain accurate candidates (for optimal design) as a solution of the corresponding optimization problem (see Step 3 in the section 2.2) using filtered (smoothed) surrogate model:

1. Filtered (smoothed) surrogate model $f_{SM,s}(X)$ and its gradient $f'_{SM,s}(X)$ are used in optimization process. Since the function $f_{SM,s}(X)$ has "good" behavior (opposed to the initial surrogate model $f_{SM}(X)$,

see section 2.4) then reasonable local minima can be found.

2. Using the initial surrogate model $f_{SM}(X)$ local optimization is done in the vicinity of local minima obtained during the previous step. This step is necessary in order to refine candidate designs.

3. Smoothing of Surrogate Model

According to the results of the previous section Smoothing of Surrogate Model is reduced to filtering (de-noising) of the trend $f_M(X)$ and its gradient $f'_M(X)$ from noisy data $f_{SM}(X)$, generated by the model (8). In the present section we are going to present theory and algorithms used for solution of this filtering problem.

3.1. Filtering: theoretical considerations

In order to estimate the trend $f_M(X)$ the following operator is used

$$f_{SM,s}(X) = \int K_s(Z-X) f_{SM}(Z) dF_s(Z), \qquad (13)$$

where $K_s(X)$ is a scaled kernel function, F_s is a measure concentrated in a finite number of available input points. Usually scaled kernel function $K_s(X)$ is constructed as a scaling of specific kernel function $K(X) = K(x_1,...,x_p)$, $X = (x_1,...,x_p)$ (formula (28) below gives an example of such kernel function $K(X)$) according to the formula

$$K_s(X) = \frac{1}{\prod_{k=1}^{p} h_k(s)} K\left(x_1/h_1(s), \ldots, x_p/h_p(s)\right), \qquad (14)$$

where $h_k(s) = s \cdot s_k$, $k = 1,\ldots,p$ define kernel width and s_k is a standard deviation of the k-th coordinate of the input vector X. Usually in practice s_k is estimated using the sample S_{train} (5).

It is assumed that for $s \to 0$ the measure F_s converges (in the weak sense after possible normalization) to some measure F determined by the design of experiment (for example, F is a uniform distribution).

Example. If rectangular window is used as a kernel (with width tending to zero when $s \to 0$) and F_s is a counting measure then the results of filtering take the form $f_{SM,s}(X) = \sum_i K_s(X_i - X) f_{SM}(X_i)$, where $\{X_i, i = 1,2,\ldots,N\}$ is a some set of inputs.

It is obvious that

$$f_{SM,s}(X) = \int K_s(Z-X) f_M(Z) dF_s(Z) + \int K_s(Z-X)\varepsilon(Z) dF_s(Z). \qquad (15)$$

The first term in (15) defines the bias

$$b_s(X) = \int K_s(Z-X) f_M(Z) dF(Z) - f_M(X). \qquad (16)$$

Sometimes for the given class of functions $f_M(X)$ such kernels can be constructed that the bias (16) is exactly zero (class of functions with reproducing kernel). Small wonder that such kernels depend on unknown function $f_M(X)$ being estimated. In general case

$$\begin{aligned} b_s(X) &= \int K_s(Z) f_M(X+Z) dF(Z) - f_M(X) \\ &= f_M(X) \cdot \left(\int K_s(Z) dF(Z) - 1\right) + \\ & f'_M(X) \cdot \int K_s(Z) Z dF(Z) + \\ & (f''_M(X)/2) \cdot \left(\int K_s(Z) Z^2 dF(Z)\right) + \ldots. \end{aligned} \qquad (17)$$

From (17) the following natural requirements to kernel functions follows:

$$\int K_s(Z) dF(Z) = 1, \qquad (18)$$

$$\int K_s(Z) Z^j dF(Z) = 0, j = 1,2,\ldots,m, \qquad (19)$$

where $m \geq 1$ is a given number. Moreover, the kernel should provide convergence to zero (in some probabilistic sense) of the random term $\int K_s(Z-X)\varepsilon(Z) dF_s(Z)$ for $s \to 0$. In such case the function $f_{SM,s}(X)$ provides consistent estimate of the trend $f_M(X)$ and fulfills the requirement (11).

As it follows from the results of section 2.3 (see (10)) we cannot use the functional (4) applied to the observed function $f_{SM}(X)$ in order to estimate the trend. Also we cannot use the functional (4) applied to the filtered function $f_{SM,s}(X)$ (15) in order to estimate the trend. The thing is that in practice we consider only finite values of s and even the filtered function $f_{SM,s}(X)$ will contain noise $\int K_s(Z-X)\varepsilon(Z) dF_s(Z)$. According to the results of section 2.3 (see (10)) this noise will result in inconsistent estimate of the gradient.

However based on the structure (15) of the filtered function we can estimate the gradient of the filtered function $f_{SM,s}(X)$ (13). In fact this is the main purpose of the filtration rather than point-wise estimation of trend values, since the differentiability of the function $f_{SM,s}(X)$ (13) depends on the smoothness of the kernel K_s. Differentiating, we obtain that

$$\begin{aligned} & f'_{SM,s}(X) = \left(\int K_s(Z-X) f_M(Z) dF_s(Z) + \right. \\ & \left. \int K_s(X-Z)\varepsilon(Z) dF_s(Z)\right)' = \int K_s(Z) f'_M(X+Z) dF_s(Z) + \\ & \int K'_s(X-Z)\varepsilon(Z) dF_s(Z) = f'_M(X) \int K_s(Z) dF_s(Z) + \\ & f''_M(X) \int K_s(Z) Z dF_s(Z) + (f'''_M(X)/2) \int K_s(Z) Z^2 dF_s(Z) \\ & + \ldots + \int K'_s(X-Z)\varepsilon(Z) dF_s(Z). \end{aligned} \qquad (20)$$

Since the measure F_s weakly converges to some measure F for $s \to 0$ and requirements (18), (19) to the kernel K_s are fulfilled then

$$f'_{SM,s}(X) \approx f'_M(X) + \int K'_s(X-Z)\varepsilon(Z)dF_s(Z). \qquad (21)$$

Usually such kernel $K_s(X)$ is considered that not only $\int K_s(X-Z)\varepsilon(Z)dF_s(Z)$ converges to zero for $s \to 0$ (necessary for point-wise convergence of the estimate to the real trend) but also $\int K_s^{(j)}(X-Z)\varepsilon(Z)dF_s(Z) \to 0$, $j = 1,\ldots,m$ (in probabilistic sense) for $s \to 0$. Such additional requirements ensure that functions

$$f_{SM,s(j)}(X) = \int K_s^{(j)}(X-Z) f_{SM}(Z)dF_s(Z) \qquad (22)$$

for $j = 1,\ldots,m$ provide consistent estimates of derivatives $f_M^{(j)}(X)$, $j = 1,2,\ldots,m$ and fulfill the requirement (12). Thus the functions $f_{SM,s}(j)(X)$, $j = 1,2,\ldots,m$ can be considered as the surrogate models for the derivatives of the trend $f_M(X)$.

If the model (8) does not contain noise $\varepsilon(X)$ (the surrogate model $f_{SM}(X)$ is absolutely exact, i.e. $f_{SM}(X) = f_M(X)$) then the error of the filtration is determined only by the bias (16) which can be easily controlled. Filtration applied to already smooth function will not significantly distort such smooth function. Thus when estimating gradient it is reasonable to use preliminary smoothing in any case, since among other things such smoothing reduces errors of numerical calculations.

Therefore in order to estimate the trend $f_M(X)$ and its derivatives up to the m-th order the functions

$$f_{SM,s(j)}(X) = \int K_s^{(j)}(X-Z) f_{SM}(Z)dF_s(Z) \qquad (23)$$

for $j = 1,\ldots,m$ should be used and smoothing of the surrogate model is realized by the functions (23) for $m = 1$.

Remark. It can be seen that for all j in (23) weighted sums of the values of the same model are considered, namely

- the values of the initial model $f_M(X)$ (in case the surrogate model $f_{SM}(X)$ is constructed using local methods), or
- the values of the constructed surrogate model $f_{SM}(X)$ in other cases.

The measure F_s is concentrated on these values in all cases.

Optimal width s of the scaled kernel $K_s(X)$ is obtained as a solution of the following optimization problem

$$s_{opt} = \arg\min_{s>0} \mathbb{E}(f_{SM,s(0)} - f_M)^2, \qquad (24)$$

where $f_{SM,s(0)}(X)$ is defined by the formula (23) and $f_M(X)$ is an initial unknown model (trend), see (8).

Example. If F_s is a counting measure concentrated on the points $\{X_i, i = 1,2,\ldots,N\} \in S_{train}$ and the distribution of the input vector X is uniform in some bounded domain then it can be proved that asymptotically for $s \to 0$, $Ns^p \to \infty$ (see [8, 4])

$$s_{opt} \approx (C_1 \cdot p/4C_2 \cdot N)^{1/(p+4)}, \qquad (25)$$

where

$$C_1 = \|K\|_2^2 \int \sigma^2(X)dX / \prod_{j=1}^{p} s_j, \qquad (26)$$

$$C_2 = (\mu_2(K))^2 \int (\mathrm{Tr}(X))^2 dX/4, \qquad (27)$$

$\sigma^2(X)$ is a variance of the noise term $\varepsilon(X)$ in (8), $\mathrm{Tr}(X) = \mathrm{tr}\left(\sigma^{\mathrm{T}} \cdot H_p(X) \cdot \sigma\right)$, $\sigma = \mathrm{diag}(s_1,\ldots,s_p)$, $H_p(X) = \left(\frac{\partial^2 f_M(X)}{\partial x_i \partial x_j}\right)_{i,j=1}^{p}$, $\mathrm{tr}(\cdot)$ is a trace of the given matrix, $\|K\|_2^2 = \int K^2(X)dX$, $\mu_2(X)$ is a such number that $\int uu^{\mathrm{T}} K(u)du = \mu_2(K)I_p$ with $u = \begin{pmatrix} u_1 \\ \ldots \\ u_p \end{pmatrix}$, and identity matrix $I_p \in R^{p\times p}$. For such optimal parameter s_{opt} the mean integrated square error is equal to $\mathbb{E}\left(f_{SM,s_{opt}(0)} - f_M\right)^2 = N^{-4/(p+4)} \cdot (C_1)^{4/(p+4)} \cdot (C_2)^{p/(p+4)} \cdot \{(p/4)^{-p/(p+4)} + (p/4)^{4/(p+4)}\}$. Thus when we increase the sample size for $s = s_{opt}(N,p)$ the error of approximation decreases as $\sim N^{-4/(p+4)}$. So, if we want the error of filtering to be E then the size of the sample should be equal to $N_{opt} \approx (C(p)/E)^{p/4+1}$, where $C(p) = (C_1)^{4/(p+4)} \cdot (C_2)^{p/(p+4)} \cdot \{(p/4)^{-p/(p+4)} + (p/4)^{4/(p+4)}\}$. Therefore the optimal size N_{opt} of the sample depends on the integrated variance of the noise and integrated trace of the squared scaled Hessian of the unknown function $f_M(X)$.

3.2. Filtering: algorithmic considerations

According to the results of the previous section in order to smooth surrogate model the following things should be specified:

1. Kernel function $K(X)$ for use in (23).
2. Algorithm for selection of the optimal (in the sense of the criterion (11)) width s of the kernel $K_s(X)$.
3. Algorithm for efficient evaluation of the integral in (23). Usually input vector X has dimension greater than one, so the algorithm should be scalable and weakly dependent on the dimension of input vector X.

Let us consider these issues more thoroughly. It is proposed to use so-called Epanechnikov kernel function as

a kernel function $K(X)$

$$K(X) = \left(1 - \sum_{k=1}^{p} x_k^2\right) \mathrm{I}\left(\sum_{k=1}^{p} x_k^2 \le 1\right) \Gamma(p/2+2)/\pi^{p/2}, \tag{28}$$

where $X = (x_1, \ldots, x_p)$, $\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt$ is a gamma-function, $\mathrm{I}(A)$ is an indicator function of the event A. This kernel was chosen because of its compact support and optimality in a minimum variance sense ([8, 4]). Used scaled version $K_s(X)$ of the Epanechnikov kernel has the form

$$K_s(X) = \frac{\Gamma(p/2+2)}{\left(\pi^{p/2} \prod_k^p h_k(s)\right)} \left(1 - \sum_{k=1}^{p} x_k^2 / h_k^2(s)\right) \times \mathrm{I}\left(\sum_{k=1}^{p} x_k^2 / h_k^2(s) \le 1\right) \tag{29}$$

Optimal width s of the kernel $K_s(X)$ is selected on basis of formulas (25), (26) and (27). For application of these formulas initial estimates of the unknown function $f_M(X)$ and its derivatives are necessary. Such initial estimates are obtained by the same formulas (23) with kernel width estimated by the formula of Bowman and Azzalini ([8, 4]). Efficient numerical algorithms were developed for estimation of C_1 and C_2 in (26), (27).

Usually surrogate model $f_{SM}(X)$ has the form (see, for example, [3, 1])

$$f_{SM}(X) = \sum_{i=1}^{d} V_i \cdot \psi_i(X), \tag{30}$$

where $\{\psi_i(X)\}_{i=1}^{d}$ are adaptive basis functions (squared anisotropic Gaussian functions, generalized sigmoid-type functions, multivariate adaptive regression spines etc.) tuned to particular training sample S_{train} (5) by the learning algorithm implemented for example in [1, 2]. For such structure of the surrogate model the formula (23) takes the form

$$f_{SM,s(j)} = \sum_{i=1}^{d} V_i \cdot K_{i,j}(s,X),\ j = 0,1,\ldots,m, \tag{31}$$

where

$$K_{i,j}(s,X) = \int K_s^{(j)}(X-Z)\psi_i(X)dF_s(Z), \quad i = 1,\ldots,d,\ j = 0,1,\ldots,m. \tag{32}$$

In most cases adaptive basis functions has specific structure. For example, multivariate generalized sigmoid-type function can be represented as $\psi_i(X) = \sigma(\gamma_i^{\mathrm{T}} \cdot X)$, where $\gamma_i^{\mathrm{T}} = (\gamma_1, \ldots, \gamma_p)$ is a some direction, and $\sigma(\cdot)$ is a one-dimensional generalized sigmoid-type function. Due to this fact multivariate integrals in (32) can be reduced to one- or two-dimensional integrals. Efficient numerical routines can be developed for their calculation based on conventional approaches to numerical integration.

Let us describe how the already constructed surrogate model can be smoothed. Algorithm for construction of smoothed surrogate models takes as input data set of $N \times (p+q)$ matrix consisting of N rows where each row is $(p+q)$-dimensional vector. Each vector is composed of p-dimensional numerical (digital) description of some object X and q-dimensional value of the characteristics $Y = f_M(X)$ for input X. As output algorithm returns the functions (23) for m = 1. These functions are constructed as follows:

1. Using the training sample S_{train} (5) and an algorithm for construction of approximation, for example, the one described in [1, 2], the surrogate model $f_{SM}(X)$ is constructed.

2. Using the surrogate model $f_{SM}(X)$ and the training sample S_{train} (5) the optimal width s_{opt} of the kernel $K_s(X)$ (25) is estimated as described above.

3. The following parameters are passed to the filtering (smoothing) algorithm:

   a. Parameters of the surrogate model $f_{SM}(X)$.

   b. Parameters of the optimal kernel width, namely the vector

   $$H_{opt} = [h_1(s_{opt}), \ldots, h_p(s_{opt})].$$

and the smoothed version of the constructed surrogate model is obtained as described above.

Let us denote by $\alpha \in [0,1)$ the parameter used for controlling the smoothness of the surrogate model. The approximate smooth value $f_{SM,s}(X|\alpha)$ and approximate smooth gradient $f'_{SM,s}(X|\alpha)$ for the given values of p-dimensional numerical (digital) description of some object X, and given value of the parameter $\alpha \in [0,1)$ can be calculated as follows:

1. The vector $H = H(\alpha)$ is calculated by the formula

   $$H(\alpha) = \alpha/(1-\alpha) \cdot H_{opt}, \alpha \in [0,1). \tag{33}$$

   In the sequel this vector is used as a width of the scaled kernel $K_s(X)$ (29). It can be seen that $H(\alpha) = H_{opt}$ for $\alpha = 1/2$, i.e. the smoothing will be optimal. In case $\alpha \in [0,1/2)$ we will undersmooth the surrogate model $f_{SM}(X)$ and its gradient and in case $\alpha \in (1/2,1)$ we will oversmooth the approximation $f_{SM}(X)$ and its gradient. In the limit $\alpha \to 0$ the scaled kernel $K_s(X)$ (29) with such width $H(\alpha)$ will converge to the delta-function, i.e. in (23) we will get equality $f_{SM,s(j)}(X) = f_{SM}(X), j = 0,1,\ldots,m$.

2. The values of $f_{SM,s(0)}(X)$ and $f_{SM,s(1)}(X)$ are calculated according to the formulas (23), (30), (31)

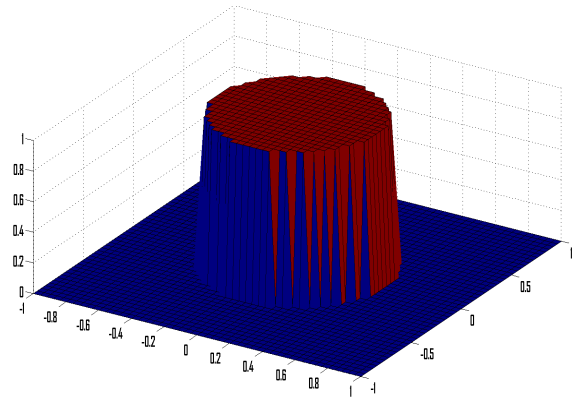

Figure 1. Initial function $f_M(X)$.

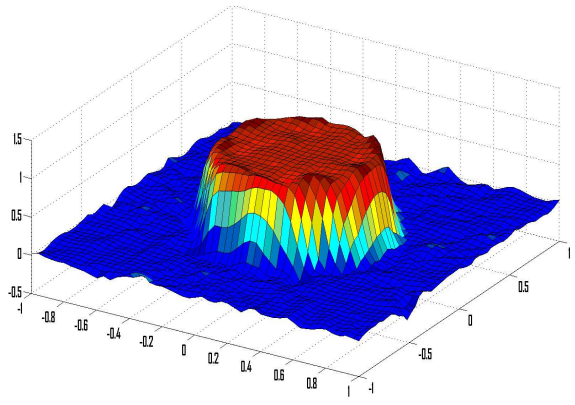

Figure 2. Surrogate model $f_{SM}(X)$.

and (32) using the kernel width equal to $H(\alpha)$. Obtained values are outputted as approximate smooth value $f_{SM,s}(X|\alpha)$ and approximate smooth gradient $f'_{SM,s}(X|\alpha)$ correspondingly.

4. Example of Application

In the present subsection an example of surrogate model smoothing is given. Plots of the following functions are constructed:

1. Initial function $f_M(X)$ (figure 1);
2. Surrogate model $f_{SM}(X)$ (figure 2);
3. Smoothed surrogate model $f_{SM,s}(X|\alpha)$, obtained with $\alpha = 0.8$ (figure 2);
4. Derivative $\frac{\partial f_{SM}(X)}{\partial x_1}$, $X = (x_1, x_2)$ (figure 4);
5. Smoothed derivative $\frac{\partial f_{SM,s}(X|\alpha)}{\partial x_1}$, $X = (x_1, x_2)$, obtained with $\alpha = 0.8$ (figure 5);
6. Derivative $\frac{\partial f_{SM}(X)}{\partial x_2}$, $X = (x_1, x_2)$ (figure 6);
7. Smoothed derivative $\frac{\partial f_{SM,s}(X|\alpha)}{\partial x_2}$, $X = (x_1, x_2)$, obtained with $\alpha = 0.8$ (figure 7);

It can be seen from these figures that proposed approach allows constructing smoothed surrogate models which facilitates surrogate based optimization.

5. Conclusions

In order to perform surrogate based optimization methods for controlling smoothness of surrogate models should be elaborated. In the present work a new algorithm for construction of smoothed surrogate models has been developed. This algorithm ensures not only

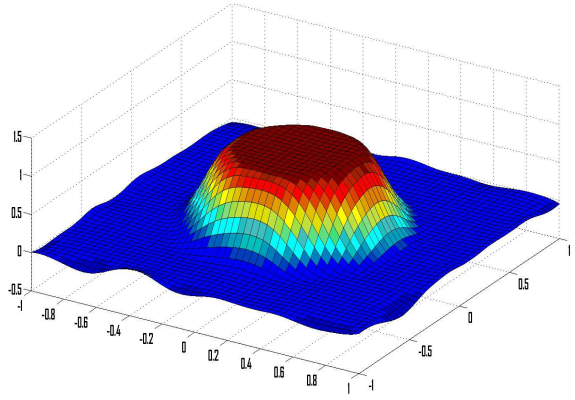

Figure 3. Smoothed surrogate model $f_{SM,s}(X|\alpha)$, obtained with $\alpha = 0.8$.

closeness of original function and the corresponding surrogate model but also ensures smoothness of gradient of the surrogate model which is important for surrogate based optimization.

References

[1] E.V. Burnaev, M.G. Belyaev, P.V. Prihodko, *Approximation of multidimensional dependency based on an expansion in parametric functions from the dictionary*, Proceedings of 9th International Conference Computer Data Analysis and Modeling: Complex Stochastic Data and Systems, September 7-11, Minsk, Belarus, 2010.

[2] E. Burnaev, M. Belyaev, P. Prihodko, *About hybrid algorithm for tuning of parameters in approximation based on linear expansions in parametric functions*, Proceedings of the 8th International conference "Intelligent Information Processing", Republic of Cyprus, Paphos, October 17-24, 2010.

[3] A.I.J. Forrester, A. Sobester, A.J. Keane, *Engineering Design via Surrogate Modelling. A Practical Guide*, Wiley, New-York, 2008.

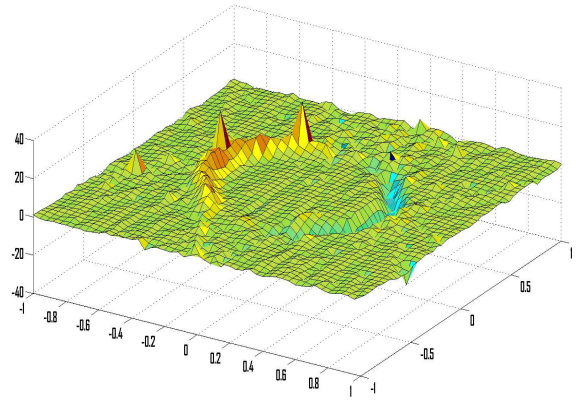

Figure 4. Derivative $\frac{\partial f_{SM}(X)}{\partial x_1}$, $X = (x_1, x_2)$.

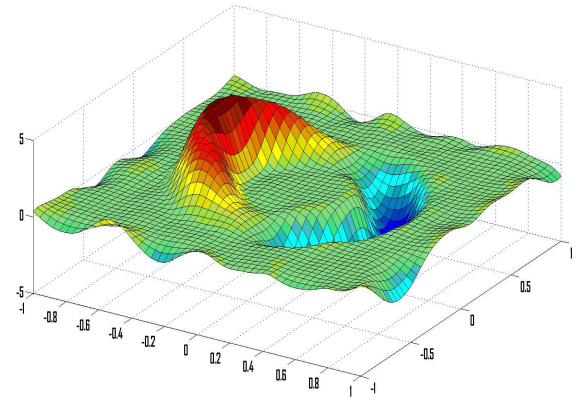

Figure 5. Smoothed derivative $\frac{\partial f_{SM,s}(X|\alpha)}{\partial x_1}$, $X = (x_1, x_2)$, obtained with $\alpha = 0.8$.

[4] W. Hardle, M. Muller, S. Sperlich, A. Werwatz, *Nonparametric and Semiparametric Models*, Berlin, Springer, 2004.

[5] A.P. Kuleshov, A.V. Bernstein, *Cognitive technologies in adaptive models of complex plants*, Proceedings of the 13th IFAC Symposium on Information Control Problems in Manufacturing (INCOM'09), June 3-5, Moscow, Russia, 2009.

[6] A.P. Kuleshov, A.V. Bernstein, E.V. Burnaev, *Adaptive models of complex systems based on data handling*, Proceedings of the 3rd International Conference on Inductive Modelling (ICIM'2010), May 16-22, Kyiv, Ukraine, 2010.

[7] Multidisciplinary and multi-objective optimization and design environment *modeFRONTIER*, available at `www.modefrontier.com`.

[8] L. Wasserman, *All of Nonparametric Statistics*, Springer Texts in Statistics, Berlin, 2007.

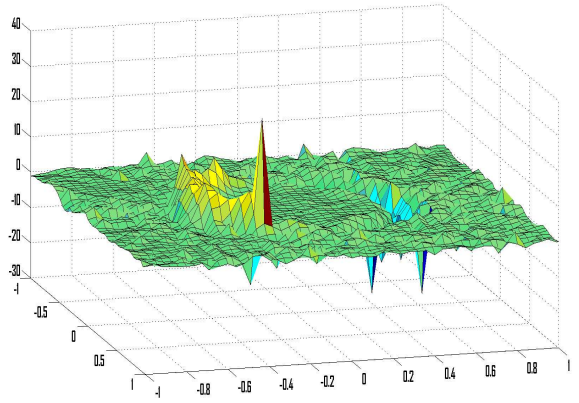

Figure 6. Derivative $\frac{\partial f_{SM}(X)}{\partial x_2}$, $X = (x_1, x_2)$.

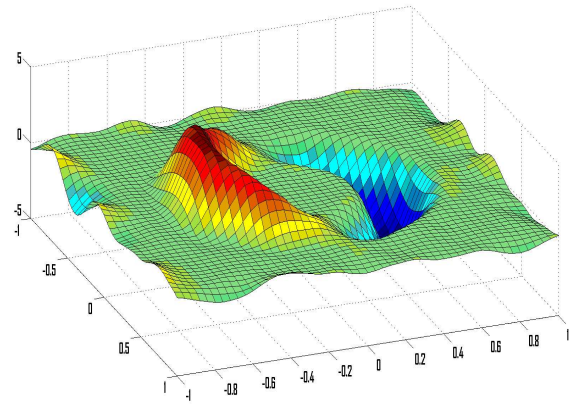

Figure 7. Smoothed derivative $\frac{\partial f_{SM,s}(X|\alpha)}{\partial x_2}$, $X = (x_1, x_2)$, obtained with $\alpha = 0.8$.