A surrogate model based on Walsh decomposition for pseudo-boolean functions

Sébastien Verel, Bilel Derbel, Arnaud Liefooghe, Hernan Aguirre, Kiyoshi Tanaka

A Surrogate Model based on Walsh Decomposition for Pseudo-Boolean Functions

Sébastien Verel[1], Bilel Derbel[2,3], Arnaud Liefooghe[2,3], Hernán Aguirre[4], and Kiyoshi Tanaka[4]

[1] Univ. Littoral Côte d'Opale, LISIC, F-62100 Calais, France
[2] Univ. Lille, CNRS, Centrale Lille, UMR 9189 – CRIStAL, F-59000 Lille, France
[3] Inria Lille – Nord Europe, F-59650 Villeneuve d'Ascq, France
[4] Shinshu University, Faculty of Engineering, Nagano, Japan

Abstract. Extensive efforts so far have been devoted to the design of effective surrogate models aiming at reducing the computational cost for solving expensive black-box continuous optimization problems. There are, however, relatively few investigations on the development of methodologies for combinatorial domains. In this work, we rely on the mathematical foundations of discrete Walsh functions in order to derive a surrogate model for pseudo-boolean optimization functions. Specifically, we model such functions by means of Walsh expansion. By conducting a comprehensive set of experiments on nk-landscapes, we provide empirical evidence on the accuracy of the proposed model. In particular, we show that a Walsh-based surrogate model can outperform the recently-proposed discrete model based on Kriging.

1 Introduction

Context. Black-box optimization refers to the situation where no specific properties nor hypothesis are known about the problem to be solved. Nothing but the fitness/objective value associated to a given (candidate) solution can be used by the optimization process. This happens in different application fields in, e.g., engineering and multi-disciplinary design, where a complete mathematical modeling may not be available and/or where the problem formulation typically implies some numerical simulations [1]. Hence, solving a black-box optimization problem consists in exploring a number of carefully-generated solutions from the search space, based solely on the evaluation of their fitness value. When the cost of computing fitness values is time consuming, traditional black-box optimization techniques, such as evolutionary algorithms and metaheuristics can have a prohibitive computational cost. In this context, surrogate-assisted approaches, such as Kriging and the Efficient Global Optimization (EGO) approach [11], are methods of choice to 'predict' the quality of solutions without systematically computing their objective values.

Motivation. Surrogates models, also called meta-models, have in fact well-established foundations at the crossroad of optimization and machine learning [10]. Roughly speaking, a surrogate model can be viewed as an estimate of the function being optimized based solely of the points (learning data) sampled by the optimization process so far. This (cheap) estimate is then used to sample and

evaluate new points that are hopefully beneficial for the purpose of approaching the global optima, while significantly reducing the overall computational cost. When reviewing the abundant literature on surrogates, it is however a fact that, with the exception of few recent work [13, 1], most existing investigations are with respect to the continuous domain, that is problems where the decision variables are real-valued. When turning to the combinatorial setting, that is when the decision variables are discrete, we can safely claim that the adaptation of existing techniques is relatively scarce [1], and the development of dedicated surrogate models is in its very infancy beginning. It is worth noticing that expensive combinatorial optimization problems are a natural outcome for real-world applications from complex scheduling or neural networks, among others [10].

Contribution. In this paper, we are interested in pushing a step toward the establishment of novel surrogate models for combinatorial optimization. We focus on the class of pseudo-boolean functions for which the solution space is the set of binary strings. Our work is based on the application of Walsh functions [16], which form a complete orthogonal set of functions, and can be considered from a mathematical perspective as one of the discrete counterpart of the trigonometric functions used in Fourier analysis. As such, we propose to represent a pseudo-boolean function as a discrete decomposition on the basis of Walsh functions, which enables us to derive a new surrogate model for this class of optimization problems. Having the model established, we approximate its coefficients (hyper-parameters) using different optimization and machine learning techniques for linear regression, namely conjugate gradient (CG) and Least-Angle Regression (LARS). Using a comprehensive set of nk-landscapes [12], we first evaluate the accuracy of the so-obtained approximate model. We then conduct a comparative study with the recently-proposed Kriging surrogate model for combinatorial problems. Our experimental results allows us to show that the designed Walsh-based surrogate model is able to provide a highly accurate approximation of the considered instances, outperforming Kriging in a number of scenarios.

Outline. In Section 2, we provide an overview of the mathematical foundation of Walsh functions and related work. In Section 3, we describe the proposed Walsh-based model. In Section 4, we evaluate the accuracy of the model using nk-landscapes. In Section 5, we conclude the paper and discuss further research.

2 Walsh Functions: Background and Related Work

2.1 Walsh Functions Basics and Evolutionary Computation

Continuous Walsh decomposition. Walsh functions [16] constitute an enumerable set of functions $\varphi_k : [0, 1] \to \{-1, 1\}$ which composes a normal and orthogonal basis of the Hilbert space $L^2([0, 1])$. Like other basis of functions such as trigonometric functions of the Fourrier basis, and although the Walsh functions are not continuous since their values is either -1 and 1, they can be used to decompose any function of the Hilbert space; see [16] for the mathematical conditions. More formally, for any integer $k \in \mathbb{N}$ with the binary representation $k = \sum_{j=0}^{\infty} k_j 2^j$

and $k_j \in \{0,1\}$, the Walsh function φ_k is defined for any (real-valued) $x \in [0,1]$ with a natural binary representation $x = \sum_{j=1}^{\infty} x_j 2^{-j}$ and $x_j \in \{0,1\}$, by $\varphi_k(x) = (-1)^{\sum_{j=0}^{\infty} k_j x_{j+1}}$. Over the interval $[0,1]$, the graphical representation of the Walsh functions can be viewed in a similar way than the cosine functions. Indeed, for all $x \in [0,1]$, $\varphi_0(x) = 1$ is the constant function, the values taken by φ_1 change from 1 to -1 at $x = 0.5$, and so on. The *orthogonality* of Walsh functions means that for any positive integers j and $k \in \mathbb{N}$, $\int_0^1 \varphi_j(x)\varphi_k(x)dx = \delta_{jk}$ where δ_{jk} is the Kronecker delta. Thus, for any function f from $L^2([0,1])$, and for any $x \in [0,1]$, we have that $f(x) = \sum_{k=0}^{\infty} w_k \varphi_k(x)$, where $w_k \in \mathbb{R}$ are the *coefficients* given by the projection of f on φ_k: $w_k = \int_0^1 f(t)\varphi_k(t)dt$. The *order* of a Walsh function φ_k, denoted by $o(\varphi_k)$, is defined by the number of binary digit equals to 1 in the binary representation of k. For example, the function of order 0 is φ_0, the functions of order 1 are φ_{2^p} for all integers $p \geqslant 0$, the functions of order 2 are $\varphi_{2^p+2^{p'}}$ for all pairs of integers $p \neq p' \geqslant 0$, and so on. While the previous discussion is with respect to a continuous function (in x), similar considerations can be discussed for the discrete case, especially for pseudo-boolean functions.

Discrete Walsh decomposition and EAs. Tightly related to evolutionary algorithms (EA), discrete Walsh functions were considered by A. D. Bethke [2], a PhD student of J. Holland in the late seventies. This was further extended by D. Goldberg [8], S. Forrest and M. Mitchell [6] to offer a relevant theoretical framework on the properties of fitness functions related to the schemata theorem, and on deceptive functions in EAs. In this context, the Walsh functions are defined for any pseudo-boolean function as follows. For any integer $k \in [0, 2^n - 1]$ with the binary representation $k = \sum_{j=0}^{\infty} k_j 2^j$ and $k_j \in \{0,1\}$, the Walsh function $\varphi_k : \{0,1\}^n \to \{-1,1\}$ is defined for any binary string $x = (x_1, \ldots, x_j, \ldots, x_n) \in \{0,1\}^n$ as: $\varphi_k(x) = (-1)^{\sum_{j=0}^{n-1} k_j x_j}$. The so-defined (finite) set of discrete functions is a normal orthogonal basis for the space of pseudo-boolean functions. For any integer j, and $k \in [0, 2^n - 1]$, $\frac{1}{2^n} \sum_{x \in \{0,1\}^n} \varphi_j(x)\varphi_k(x) = \delta_{jk}$[5]. Therefore, any pseudo-boolean function $f : \{0,1\}^n \to \mathbb{R}$ can be written as a unique finite weighted sum of Walsh functions.

$$\forall x \in \{0,1\}^n,\ f(x) = \sum_{k=0}^{2^n-1} w_k.\varphi_k(x) \quad (1) \qquad \text{s.t. } w_k = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} f(x).\varphi_k(x) \quad (2)$$

Schemata theory. The average of fitness values over a schemata of order p can be computed with a subset of Walsh functions of lower orders [8]. In fact, let us recall that a schemata is a hypercube of the binary space. Usually, a schemata is written with the alphabet $\{*, 0, 1\}$ where 0, and 1 give the fixed position bits of the hyperplane. The order of the schemata is the number of 0/1 in the string. For instance, the schemata $h = *01***0$ is a schemata of length 7, and of order 3. The average fitness of a schemata h is then $f(h) = \sum_{k \subset h} w_k.\varphi_k(x)$ where $k \subset h$ means that the 1 in the binary representation of k corresponds to 0 or 1 of the schemata. Hence, it is possible to design deceptive functions to challenge EAs [8].

[5] Indeed, the matrix $(\varphi_k(x_j))_{jk}$ of dimension $2^n \times 2^n$ is a Hadamard matrix.

Walsh decomposition in combinatorial optimization. Besides their initial theoretical interest, there was recently a renewed interest to Walsh decomposition, which remains the subject of active research in the optimization community [9]. In particular, in the so-called *grey-box* optimization setting [3], standard problems such as nk-landscapes, or max-SAT are regarded as a decomposition of Walsh functions. Within such a perspective, the fitness value, the fitness distribution, or the best solution at a given Hamming distance is fast to compute, hence enabling the design of effective and efficient optimization techniques. Additionally, the Walsh decomposition can be used to detect accurate crossover points, and to identify independent sub-space problems that lead to the solving of very large combinatorial optimization problems with an impressively reduced cost [4].

2.2 Surrogate Models for Combinatorial Optimization

Surrogate models. A standard surrogate-assisted optimization framework consists in an iterative process, where at each iteration: (step i) a model is built on the basis of the solutions (learning data) evaluated so far at previous iterations, (step ii) compute best (believed, predicted) solution(s) on the basis of the so-constructed (cheap) model, (step iii) evaluate the so-chosen solution(s) using the real (expensive) black-box function f. Each of these three steps comes with different challenges and different techniques and tools to address and implement. In our work, we are interested in the application of Walsh functions for designing a surrogate model dedicated to pseudo-boolean functions. We thereby focus very specifically on the very first step of the aforementioned framework, that is, the definition and the building of a highly accurate model that can eventually be used as a substitute of the real function. It is worth noticing that this is of crucial importance towards the design of effective and efficient surrogate-assisted optimization techniques. As mentioned in the introduction, this is especially motivated by the fact that little is known about the design and the application of effective surrogates for combinatorial optimization, which is to contrast with the real breakthrough made in the field of (global) continuous black-box optimization by both the machine learning and the optimization communities [10].

Discrete surrogates. As summarized in [17], when looking at the specialized literature, we can find a number of approaches ranging from the most naive ones, ignoring the discrete nature of variables and simply following a standard machine learning framework where the data is regarded as a vector, to more specialized approaches where either the model is inherently discrete or a similarity / distance 'measure' between discrete solutions is used to leverage existing continuous models. The work presented in this paper falls in this last category, encompassing a number of noticeable techniques [1]. To cite a few, in [13], it is shown how to leverage existing distance-based surrogates, by considering more general (not necessarily continuous) metric spaces. This idea is then illustrated using Radial Basis Function Networks (RBFN). Later in [18, 19], a seemingly similar principle is adopted in order to derive a Kriging (Gaussian Process) like surrogate model. Kriging has the interesting feature of providing a measure of uncertainty when determining

predictions. This can be used to calculate the Expected Improvement (EI) of a solution, which is then be used as the main criteria to balance exploitation and exploration when sampling candidate solutions in the so-called Efficient Global Optimization (EGO) approach [11]. Such an EGO approach [19] is shown to outperform Kriging and RBFN [13] on a number of nk-landscapes [12] considered as difficult adversarial pseudo-boolean benchmark functions. In our work, we also validate empirically our model using nk-landscapes as a case study, while comparing to the Kriging approach considered as a baseline competitor. Notice that using the proposed Walsh model to sample promising points (as performed in EGO) is left for future work since our main goal is to investigate the accuracy of the Walsh model in correctly rendering the original expensive function.

3 Surrogate Model based on Walsh Functions

The Walsh-based surrogate model. Given any pseudo-boolean function $f : \{0,1\}^n \rightarrow \mathbb{R}$, we have a closed form decomposition of f using the set of Walsh functions φ_k, as given in Eq. (1). It is worth noticing that the functions φ_k are problem independent, and hence uniquely defined irrespective to f. The values of the coefficients w_k in Eq. (1) depend however on function f, as given in Eq. (2). They are here assumed to be unknown and black-box. Moreover, although the number of coefficients is in general exponential in n, there might exist a significantly large number of zero coefficients, namely, 2^n. For instance, for nk-landscapes, the number of non-zero coefficients is bounded by $n \cdot 2^{k+1}$ [9] and the maximum order is equals to $k+1$. Hence, our idea is to consider an approximation of f using solely the Walsh functions of a constant order $d << n$ and using an estimate $\widehat{w}_k$ of the (unknown) coefficient w_k. More formally, we shall assume that the pseudo-boolean function f can be approximated by the following model constituting the core of our proposed surrogate model:

$$\forall x \in \{0,1\}^n,\ \hat{f}(x) = \sum_{k\ :\ o(\varphi_k) \leqslant d} \widehat{w}_k . \varphi_k(x) \tag{3}$$

Obviously, the previous equation is to recall standard (finite) Taylor series for continuous function expansion. The larger the order d, the better the approximation; and the better the quality of the estimate coefficients $\widehat{w}_k$, the more accurate the expansion. In the following, we shall focus on how to provide a good estimate of the Walsh coefficients, assuming that d is fixed to some constant. For clarity, the choice of the setting of the order d is discussed later on.

Model approximation. Given the black-box nature of the true pseudo-boolean function f, one idea would be to consider a sample of solutions for which we know the true f-values. Let us assume given such a set denoted $\mathcal{S}$. For now, we do not make any further assumption on $\mathcal{S}$. Then, the question we are trying to solve is: find an estimate $\widehat{w}_k$ of the coefficients w_k using the data set $\{(x, f(x) \mid x \in \mathcal{S}\}$. One answer to this question could be to simply use the mean as estimator by setting $\widehat{w}_k = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} f(x)\varphi_k(x)$. By a routine verification, we can show that

the bias of the estimate is $\widehat{w}_k - w_k = \sum_{j \neq k} w_j \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \varphi_j(x)\varphi_k(x)$, which is to be interpreted as the degree of 'non-orthogonality' of the Walsh functions on $\mathcal{S}$. While being informative, such an estimate might be misleading, since it might be challenging to design a sample data set $\mathcal{S}$ verifying such properties w.r.t. Walsh functions. In the following, we discuss two techniques to estimate the Walsh coefficients required by the proposed surrogate.

Mean squared error estimation using conjugate gradient (CG). The Walsh decomposition of a pseudo-boolean function is a *linear model* where the predictors are the Walsh functions' values. As a consequence, classical methods for non-sparse and sparse approximation can be used to estimate the coefficients of the regression. Our first technique is based on a standard approach which consists in minimizing the mean squared error of the surrogate (linear) model with respect to data set $\mathcal{S}$. More formally,

$$mse(\widehat{w}) = \sum_{x \in \mathcal{S}} \left(\sum_{k \,:\, o(\varphi_k) \leqslant d} \widehat{w}_k . \varphi_k(x) - f(x) \right)^2 \tag{4}$$

We then find the coefficients $\widehat{w}^*$ minimizing Eq. (4), that is: $\widehat{w}^* = \text{argmin}_w mse(w)$. To solve this equation, we propose to use a non-sparse method, namely the conjugate gradient (CG) approach [14].

Least-angle regression (LARS) coefficients estimate. When the number of predictors in a linear regression is large, sparse techniques can be used to minimize the number of non-zero coefficients. Among others, lasso is one classical technique for sparse approximation [15]. In this work, we propose to use the least-angle regression (LARS) algorithm [5] to fit the model. The LARS algorithm is in the same family of regularization/sparse methods and follows a forward stepwise selection regression mechanism. It has the major advantage of being computationally fast and effective when fitting high-dimensional data of relatively small size. Hence, it is a method of choice in our context since the number of Walsh functions of order d might be greater than the number of samples in the training data set $\mathcal{S}$.

Order setting. Finally, we need to specify a value for the maximum order d to be set in the estimate. Intuitively, the larger the order, the larger the number of Walsh coefficient to be estimated, and hence the better the approximation. However, the larger the number of coefficients, the more difficult and time consuming their estimation using the previously-described techniques. Let n_d be the number of Walsh coefficients of order d. Then, we have that: $n_0 = 1$ and $n_d = n_{d-1} + \binom{n}{d}$ for $d > 0$. This makes the choice of large d values problematic. Nonetheless, we argue that the number of non-zero coefficients is typically much less than n_d and a value of d of at most 3, for which $n_d = O(n^3)$,

Table 1. Number of Walsh coefficients

	10	15	20	25
0	1	1	1	1
1	11	16	21	26
2	46	121	211	326
3	176	576	1351	2626

should be sufficient for an accurate approximation of multi-modal (difficult-to-optimize) functions, as supported by our empirical results. Table 3 shows the values of n_d for different values of n (columns) and d (row).

4 Experimental Analysis

4.1 Experimental Setup and Methodology

Test functions. As in previous studies [13, 19], we consider nk-landscapes [12] as benchmark pseudo-boolean functions. For every binary string $x = (x_1, \ldots, x_n)$ of size n, $f(x)$ is defined as the average value of the *contributions* associated with each variable x_i. For every $i \in \{1, \ldots, n\}$, a component function $f_i \colon \{0,1\}^{k+1} \mapsto [0,1]$ assigns a real-valued contribution for every combination of x_i and its k *epistatic interactions* $\{x_{i_1}, \ldots, x_{i_k}\}$. In other words, the individual contribution of a variable x_i to $f(x)$ depends on its value and on the values of $k < n$ other variables $\{x_{i_1}, \ldots, x_{i_k}\}$. The function f is hence defined as follows: $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x_i, x_{i_1}, \ldots, x_{i_k})$. The k epistatic interactions w.r.t. a variable x_i are set uniformly at random among the $(n-1)$ variables other than x_i [12]. The f_i values are uniformly distributed in $[0,1]$. It is important to remark that by increasing the number of epistatic interactions k from 0 to $(n-1)$, problem instances can be gradually tuned from smooth to rugged, which make nk-landscapes an abstract adversarial optimization benchmark that can eventually cover a wide range of (real-world) pseudo-boolean function classes, as commented further in the following.

Experimental Setup. We consider a comprehensive set of nk-landscapes with $n \in \{10, 15, 20, 25\}$, and $k \in \{0, 1, 2\}$. Notice that for $k = 0$, the function is linear which makes it easy to optimize. For $k = 1$, the function is quadratic which, informally speaking, falls in the same class than the widely studied Unconstrained Binary Quadratic Problem. For $k = 2$, every variable is in interaction with two other ones which, informally speaking, recalls the max-3-SAT problem. For every parameter combination ($4 \times 3 = 12$), we generate 5 instances for which every competing model/algorithm is run for 5 independent runs. The reported results are over the $5 \times 5 = 25$ independent runs. All algorithms and experiments are implemented in R using standard machine learning and optimization packages.

Validation methodology. In our work, we focus on studying the accuracy of the Walsh expansion surrogate in providing a high fidelity approximation. Consequently, we follow the experimental procedure depicted in the template of Algorithm 1. First, we generate a (test) set Q of $N = 1000$ solutions generated uniformly at random (i.e., each bit is set to 0 or 1 with equal probability) which is used as input of our experimental procedure. For each instance, and for every iteration $t > 0$ of Algorithm 1, we generate uniformly at random a new solution x_t and evaluate its true fitness value $f(x_t)$. Next, we build a surrogate model $\widehat{f}_t$ using the (training) data set $S_t = S_{t-1} \cup \{x_t\}$. We then record an error measure (denoted ϵ_t) rendering the quality of the fit $(\widehat{f}_t, S_t)$ with respect to the (test) data set Q. This shall allow us to study the ability of the surrogate model to fit the real

Algorithm 1: Experimental procedure

Input: A test set Q

$S_0 \leftarrow \varnothing$;
for $t = 1, 2, \ldots, Max\_Budget$ **do**
  $x_t \leftarrow$ a solution generated uniformly at random;
  $S_t \leftarrow S_{t-1} \cup \{(x_t, f(x_t))\}$;
  $\widehat{f}_t \leftarrow$ build a surrogate model for f on the basis of (the training set) S_t;
  $\epsilon_t \leftarrow$ a measure of the quality of the accuracy of $\widehat{f}_t$ using the test set Q;
end

function as the size of the available sample data grows, that is, as the available budget in terms of (expensive) function evaluations is given. The maximum allowed budget is actually variable in the size of the considered nk-landscape.

As a baseline, we use Kriging [7] as a state-of-the-art discrete surrogate model, and use the R package CEGO[6] as an implementation. As an error measure, we compute the mean absolute error (mae) and the mean squared error (mse) of $\widehat{f}$ w.r.t Q. When the proposed Walsh model is experimented, we additionally record the Walsh coefficients estimate ($\widehat{w}_k$), and compute their R^2 coefficient.

4.2 Training with CG *vs.* LARS

First, we consider the accuracy of the model when using the two fitting techniques (CG and LARS) for estimating the Walsh coefficients (see Section 3). In Fig. 1, we show the evolution of the mean absolute error as a function of the size of the training set S_t, using nk-landscapes with $n = 20$ and $k = 1$. Notice that similar results holds for the mean squared error and are omitted due to lack of space.

Fig. 1 (left) shows that the Walsh model computed with LARS leads to a significantly better fit than the CG based technique, while requiring a sample of significantly lower size. Actually, the error of both methods converges to 0 with LARS being significantly faster in the size sample. This means that it requires much fewer function evaluations to converge to a high fidelity Walsh approximate. In Fig. 1 (right), we show a scatter plot rendering the relative distribution of $\hat{f}(x)$ and $f(x)$ using the LARS for nk-landscapes with $n = 15$, and $k = 1$ trained on a random sample set of size 60, and tested on a set composed by the whole search space. The quality of the regression visually approximates the original function for all fitness values. Indeed, the residues can be bounded by a constant independent of the fitness value: for any $x \in \{0,1\}^n$, $|\widehat{f}(x) - f(x)| = |\sum_k (\widehat{w}_k - w_k)\varphi_k(x)|$. Given that $|\varphi_k(x)| = 1$, we obtain the upper bound $|\widehat{f}(x) - f(x)| \leqslant \sum_k |\widehat{w}_k - w_k|$ which interestingly does not depend on x.

4.3 Walsh vs. Kriging

In Fig. 2, we show results comparing the proposed Walsh model to Kriging. Two main tightly related observations can be extracted. On the one hand, for

[6] `https://cran.r-project.org/package=CEGO`

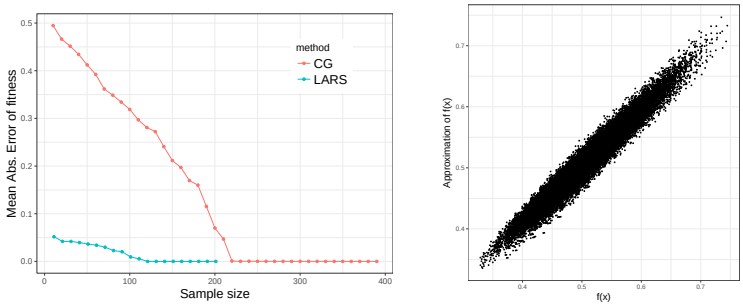

Fig. 1. Mean Absolute Error of fitness values using CG and Gradient techniques as a function of the training set size for $n = 20$, and $k = 1$ (left). Scatter plot (right) of the regression using LARS for $n = 15$, $k = 1$, and a random sample of size 60.

$k = 0$, both surrogates provide consistently similar accuracy. However, as the test function is no more linear ($k = 1$ and $k = 2$), the difference is substantial in favor of the proposed Walsh surrogate. From a fitness landscape analysis perspective, higher values of k lead to more rugged (non-smooth) multi-modal functions. In this case, and although Kriging has a better accuracy with very few samples (very few function evaluations), the Walsh model is able to converge much faster to a high quality fit. On the other hand, the difference becomes even more substantial when scaling the dimension of the pseudo-boolean function. In fact, for the highest values of k and n, it is clear that the performance of Kriging drops very significantly since it is even no more able to provide a high accuracy within a reasonable budget. This is to contrast with the Walsh model, which converges to a zero absolute error within a few hundreds of function evaluations. The high quality of the Walsh surrogate can be explained by the fact that it is a deterministic model (for noise-free function) which provides a quasi-exact modeling of the pseudo-boolean function once the coefficient value estimates are close enough to their true values. This claim can be supported by a more focused analysis on the quality of the coefficients estimates, which is discussed in the next section while commenting on the impact of the Walsh expansion order.

4.4 Impact of the Walsh Expansion Order

In the previous results, the order of the Walsh decomposition was fixed to $d = k+1$ where k is the number of the epistatic interaction in the considered nk-landscapes. However, one might ask what happens if the order is fixed to a different value. This is what is depicted in Fig. 3, showing the coefficient of determination (R^2) to render the relative quality of the approximated coefficients. Notice that the exact values of the Walsh coefficients at any order are computed by the sum of Walsh decomposition of the component functions [9].

First, we can see that the R^2 converges relatively quickly to 1 when the order of the expansion is $k + 1$. Hence, one can reasonably suggest that for other highly multi-modal functions, it might hold that only a restricted number of (low) order coefficients have non-zero values. In this case, only a restricted number of

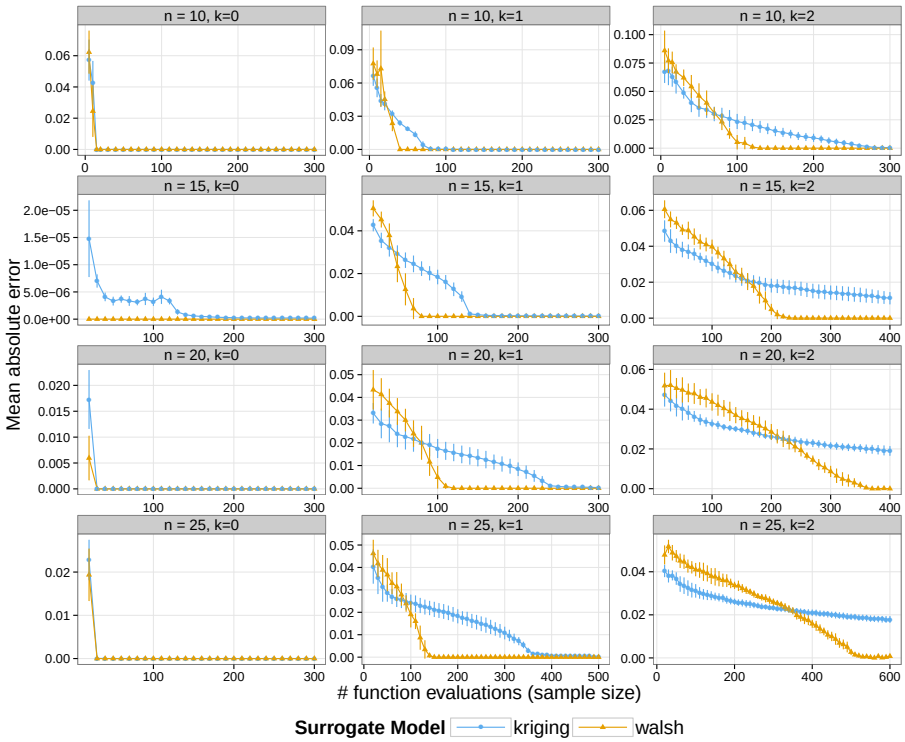

Fig. 2. Mean Absolute Error of fitness values on test set (1000 solutions) as a function of the size of the training set. $n \in \{10, 15, 20, 25\}$ (rows), $k \in \{0, 1, 2\}$ (columns), and the order of the Walsh expansion is $k + 1$.

coefficient could impact the accuracy of the Walsh surrogate model, which would make it easier to build. Moreover, Fig. 3 shows that the coefficients of lower orders can still be accurately estimated although the value of the order chosen for the fit does not match with the maximum order of non-zero coefficients in the exact Walsh expansion. This suggests that for other highly multi-modal functions, it might hold that a small value of the order considered when fitting the Walsh surrogate is still sufficient to provide a high fidelity rendering of the original function. This property is of special interest since the lower the considered order, the lower the number of coefficient to be estimated, and the lower the cost of building the Walsh model. The cost of computing the surrogate model can in fact constitute a critical issue, especially if it comes to dominate the cost of the (expensive) function evaluation. In this respect, our LARS implementation of the Walsh surrogate was found to have a number of orders of magnitude better CPU runtime compared against the Kriging implementation.

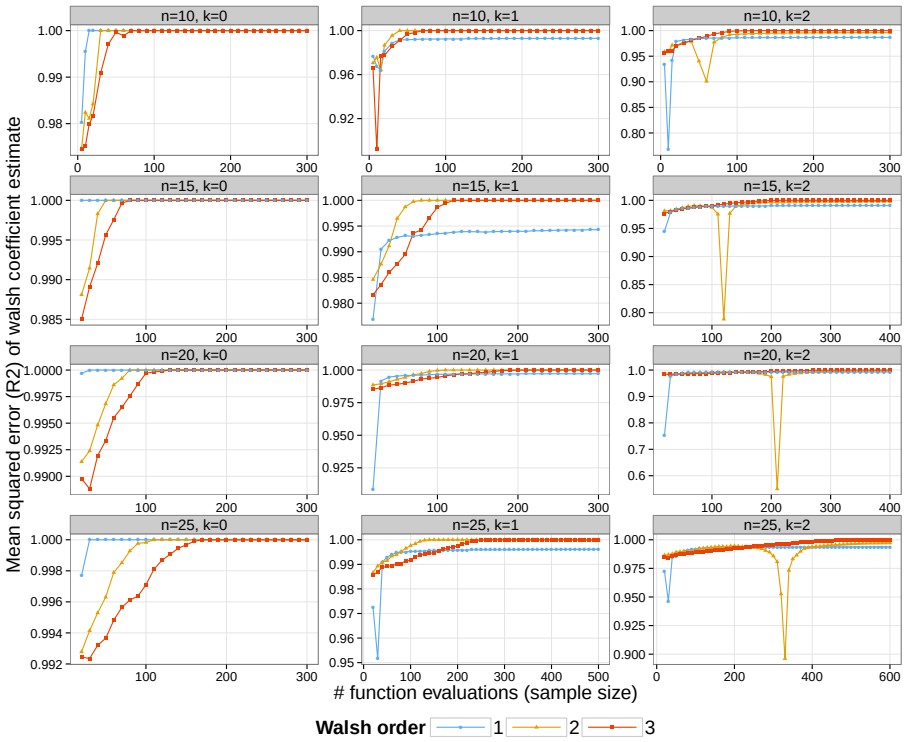

Fig. 3. R^2 of Walsh coefficients as a function of the size of the training set. $n \in \{10, 15, 20, 25\}$ (rows), $k \in \{0, 1, 2\}$ (columns), and the order of the expansion (color).

5 Conclusions

In this paper, we introduced a framework allowing to apply the Walsh functions basis in order to construct a novel discrete surrogate model for expensive pseudo-boolean functions, which is shown to be highly accurate on a set of nk-landscapes. Unlike previous distance/similarity based discrete surrogates, the proposed model is based on a deterministic pseudo-exact approximation. As such, it has some advantages and some shortcomings that, hopefully, will provide as much new research opportunities as new scientific challenges. Besides, e.g., applying the proposed model to other (real-world) pseudo-boolean functions, or improving the quality of the approximation using more advanced optimization techniques or heuristics, embedding the Walsh-based model within a conventional surrogate-assisted optimization framework would provide a highly effective approach to expensive problem solving. In fact, not only building the Walsh surrogate is extremely fast compared against Kriging, but its solving to optimality using the recently-proposed grey-box optimization techniques [4] is fully plausible even for large-scale problems. This can for instance be a relevant alternative to the use of EGO-like selection criteria at a reduced cost. Additionally, generalizing

the model to other non-necessarily pseudo-boolean functions, like permutation problems, would be a major advance. Finally, we believe that accommodating the deterministic nature of the model, for instance by taking into account the error in the coefficients approximation based on a probabilistic modeling, while being a very challenging issue, can increase its potential in tackling a wide range of optimization problems of large dimensions.

References

1. T. Bartz-Beielstein and M. Zaefferer. Model-based methods for continuous and discrete global optimization. *Applied Soft Computing*, 55:154–167, 2017.
2. A. D. Bethke. *Genetic algorithms as function optimizers*. PhD thesis, University of Michigan, 1980.
3. F. Chicano, D. Whitley, and E. Alba. Exact computation of the expectation surfaces for uniform crossover along with bit-flip mutation. *Theoretical Computer Science*, 545:76–93, 2014.
4. F. Chicano, D. Whitley, G. Ochoa, and R. Tinós. Optimizing one million variable nk landscapes by hybridizing deterministic recombination and local search. In *GECCO*, pages 753–760, 2017.
5. B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
6. S. Forrest and M. Mitchell. What makes a problem hard for a genetic algorithm? some anomalous results and their explanation. *Ma. Learn.*, 13(2-3):285–319, 1993.
7. A. Forrester, A. Keane, et al. *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons, 2008.
8. D. E. Goldberg. Genetic algorithms and walsh functions: Part i, a gentle introduction. *Complex systems*, 3(2):129–152, 1989.
9. R. B. Heckendorn. Embedded landscapes. *Evo. Comp.*, 10(4):345–369, 2002.
10. Y. Jin. Surrogate-assisted evolutionary computation: Recent advances and future challenges. *Swarm and Evolutionary Computation*, 1(2):61–70, 2011.
11. D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.
12. S. A. Kauffman. *The Origins of Order*. Oxford University Press, 1993.
13. A. Moraglio and A. Kattan. Geometric generalisation of surrogate model based optimisation to combinatorial spaces. In *European Conference on Evolutionary Computation in Combinatorial Optimization*, pages 142–154. Springer, 2011.
14. J. R. Shewchuk et al. An introduction to the conjugate gradient method without the agonizing pain, 1994.
15. R. Tibshirani, M. Wainwright, and T. Hastie. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.
16. J. L. Walsh. A closed set of normal orthogonal functions. *American Journal of Mathematics*, 45(1):5–24, 1923.
17. M. Zaefferer and T. Bartz-Beielstein. Tabular survey: Surrogate models in combinatorial optimization - version 5. Technical report, 05 2017.
18. M. Zaefferer, J. Stork, and T. Bartz-Beielstein. Distance measures for permutations in combinatorial efficient global optimization. In *International Conference on Parallel Problem Solving from Nature*, pages 373–383. Springer, 2014.
19. M. Zaefferer, J. Stork, M. Friese, A. Fischbach, B. Naujoks, and T. Bartz-Beielstein. Efficient global optimization for combinatorial problems. In *GECCO*, 2014.