Exploiting problem structure in derivative free optimization

M. Porcelli∗† and Ph. L. Toint‡

12 I 2021

Abstract

A structured version of derivative-free random pattern search optimization algorithms is introduced which is able to exploit coordinate partially separable structure (typically associated with sparsity) often present in unconstrained and bound-constrained optimization problems. This technique improves performance by orders of magnitude and makes it possible to solve large problems that otherwise are totally intractable by other derivative-free methods. A library of interpolation-based modelling tools is also described, which can be associated to the structured or unstructured versions of the initial pattern search algorithm. The use of the library further enhances performance, especially when associated with structure. The significant gains in performance associated with these two techniques are illustrated using a new freely-available release of the BFO (Brute Force Optimizer) package firstly introduced in [42], which incorporates them. An interesting conclusion of the numerical results presented is that providing global structural information on a problem can result in significantly less evaluations of the objective function than attempting to building local Taylor-like models.

Keywords: derivative-free optimization, direct-search methods, structured problems, interpolation models.

Mathematics Subject Classification: 65K05, 90C56, 90C90.

1 Introduction

Derivative-free methods have enjoyed a continued popularity and attention from the very early days of numerical optimization [6, 30, 40, 52] to nowadays [3, 15, 32]. As their name indicates, such methods are aimed at solving nonlinear optimization problems without using derivatives of the objective function, which can be either too costly to compute, or simply unavailable (as is often the case in simulation-based applications). Among

∗Università di Bologna, Dipartimento di Matematica, Piazza di Porta S. Donato, 5, 40126, Bologna, Italy. Email: margherita.porcelli@unibo.it
†Institute of Information Science and Technologies “A. Faedo”, ISTI-CNR, Pisa, Italy.
‡Namur Center for Complex Systems (NAXYS), University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium. Email: philippe.toint@unamur.be
all possible methods, we focus here on two broad classes: model-based techniques and pattern-search algorithms. In the first class, an explicit local model of the objective function is built, which is then used for finding any better approximation of the sought minimizer. Early proposals in this direction focused on unconstrained problems and include [17, 29, 36, 44, 45, 46, 47, 60], and were later analyzed and/or extended to the constrained case (see [7, 13, 14, 11, 16, 21, 48, 28, 55, 53, 51] among many others). The second class, also initially for unconstrained problems [1, 2, 6, 22, 30, 40, 41, 52], was later analyzed in [18, 19, 20, 23, 27, 31, 58, 59], and also extended to more general contexts (see [4, 33, 34, 42, 49, 35] for example).

Most of the contributions on derivative-free methods focus on small-scale unstructured problems. Indeed, building large-scale multivariate models is very expensive as it requires a number of data points (and costly function evaluations) which quickly grows with dimension: for the widely used quadratic polynomial models, this number grows like the square of the problem’s size. As a consequence, model-based approaches are in general unrealistic for moderate or large-scale applications. Similarly, the cost of the sampling strategies at the heart of pattern-search techniques also explodes with problem dimension and in practice restrict the application of these algorithms to optimization in very few variables (a few tens).

The situation fortunately improves if, instead of pure black-box optimization, one now considers the “gray box” case where one is allowed to exploit some underlying problem structure (while still avoiding the use of any derivative), see e.g. for an application [37]. Known sparsity in the objective function’s Hessian was considered in [9, 10] for methods based on quadratic models, where it was shown that the growth in function evaluations with size is, for a large class of sparse problems, essentially linear rather than quadratic. However, knowledge of the Hessian sparsity pattern is rarely directly obtained in practice without the knowledge of derivatives. What is much more common, for instance in discretized problems, is that optimization is performed on an application involving interconnected and possibly overlapping subsystems. While the detailed analytic expression of the subsystems’ models are often unavailable, it is not unusual for their connectivity pattern to be known. Such problems then often fall in the class of coordinate partially separable (or CPS) problems, where one attempts to solve

$$\min_{x \in \mathbb{R}^n} f(x) = \sum_{i=1}^{q} f_i(x)$$

where each $f_i : \mathbb{R}^{n_i} \rightarrow \mathbb{R}$ is a possibly nonsmooth (or even non continuous) function depending on a subset $\{x_i\}_{i \in \mathcal{X}_i}$ of the vector of variables $x$, for some index sets $\mathcal{X}_i \subseteq \{1, \ldots, n\}$ ($i = 1, \ldots, q$). The cardinality of these $q$ subsets is denoted by $n_i$ and is typically much smaller than the dimension $n$ of $x$. In practice it is indeed common to see values of $n_i$ of the order of ten or less, even for problems involving thousands of variables or more, see e.g. problems described in Section 4. Knowing the subsystems’ connectivity pattern then typically amounts to knowing the sets $\mathcal{X}_i$ and being able to evaluate the “element functions” $f_i$ individually. While we have introduced the concept for the unconstrained formulation (1.1), this is by no means restrictive, as constraints
may be added without affecting the structure. In the rest of this paper, we will assume that the problem’s variables are subject to simple bounds, that is
\[ \ell \leq x \leq u, \quad (1.2) \]
where \( \ell \) and \( u \) are vectors in \( \mathbb{R}^n \) such that \( \ell \leq u \), all inequalities being understood componentwise.

Optimizing coordinate partially separable problems\(^{(1)}\) with model-based algorithms was considered in [11], where it was shown that the requirements in terms of function evaluations now depend on the maximum value of the \( n_i \), which is often independent of the true problem’s dimension\(^{(2)}\). The use of the same problem structure was also initiated in [49], where a first pattern-search algorithm was described which could solve respectable size problems with \( n \) up to 5625.

The purpose of the present paper is to build on these contributions and to propose (in Section 2) a new derivative-free algorithm which is able to exploit the problem structure at pattern-search level (known as a “poll step”) as well as a library of partially separable multivariate polynomials to be combined with the structured poll step in a user-controlled “search step” (in Section 3). The dramatic impact of these features on numerical performance is then illustrated using a new version of the BFO (Brute Force Optimizer) package [42] incorporating them (in Section 4). In particular, it is shown that the use of the global structural CPS information is often more efficient than that of local approximations of the objective function’s Taylor expansion. It also shown that a combined approach may also be advantageous, at least for problems of moderate size. Some conclusions and perspectives are finally presented in Section 5.

**Notation**

The \( j \)-th component of a vector \( x \) is denoted by \( x_j \), while \( x_I \) denotes the subvector defined by considering the components of \( x \) indexed by the index set \( I \).

## 2 Exploiting coordinate partial separability in the poll step

For the purpose of introducing our structure-exploiting techniques, we consider a generic pattern-search method consisting of a succession of iterations, each containing a search step (discussed in Section 3), a poll step, in which new (hopefully better) functions values are generated by taking steps along randomly generated orthogonal sets of search directions, and an update of the stepsizes as the iterations proceed. A description of such a simple pattern search for unconstrained problems is given in Algorithm 2.1. Comments on the constrained case are postponed to the end of this section.

\(^{(1)}\)Coordinate partially separable problems are a subclass of the more general partially separable problems, where the objective function takes the form \( f(x) = \sum_{i=1}^n f_i(U_i x) \), where the \( U_i \) are \( n \times n \) low-rank matrices.

\(^{(2)}\)Coordinate separability has been extensively considered also in the context of derivative-based algorithms, see e.g. [50, 61], but, to our knowledge, in this setting the problem structure is assumed to be unknown and therefore is not exploited in the algorithm development.
Algorithm 2.1: Outline of a simple pattern search algorithm

**Initialization:** The initial iterate $x$, the initial stepsize $\alpha$ and a convergence threshold $\epsilon > 0$ are given. The parameters $\gamma \geq 1$, $\beta, \eta \in (0, 1)$ are given. Define a set of orthonormal polling directions $\{d_i\}_{i=1}^n$.

**Until convergence**

**Search step:** Ask the user to provide a new (potentially improved) approximate minimizer of $f$, typically using problem specific modelling techniques;

**Poll step:** For $j = 1, \ldots, n$ or until “sufficient decrease” in $f$ is obtained

- define a step of the form $\alpha d^j$
- evaluate $f(x + \alpha d^j)$ and, if necessary, $f(x - \alpha d^j)$,

If $f(x + \alpha d^j) < f(x)$ or $f(x - \alpha d^j) < f(x)$, replace $x$ by $x \pm \alpha d^j$, depending on which step gave decrease;

Terminate the poll step if sufficient decrease is obtained, i.e. if

$$f(x \pm \alpha d^j) - f(x) < \eta \alpha^2.$$  \hfill (2.1)

**Termination step:**

- If $f$ has decreased in the course of the poll step, increase the stepsize $\alpha$ by setting $\alpha \leftarrow \gamma \alpha$, generate a new random set of orthonormal polling directions $\{d_i\}_{i=1}^n$ (successful iteration) and start a new iteration;
- otherwise check for convergence (unsuccessful iteration): if $\alpha > \epsilon$, \hfill (2.2)

  decrease $\alpha$ by setting $\alpha \leftarrow \beta \alpha$, generate a new random set of orthonormal polling directions $\{d_i\}_{i=1}^n$ and start a new iteration; otherwise declare convergence.
A few comments are useful at this stage. We first note that when sufficient decrease is obtained for one of the polling direction, as tested in (2.1), the algorithm stops using the current set of directions and directly updates the stepsize before starting a new iteration.

When (2.2) fails, that is when
\[ \alpha \leq \epsilon, \]  
(2.3) it would be acceptable to terminate the optimization entirely. However, it is useful to continue the effort for finding a better point by applying a user-defined number of additional poll steps, each using a new random orthonormal basis \( \{ d^i \}_{i=1}^n \). If condition (2.3) is met at every such step, final convergence is then declared.

We assume, as indicated by (1.1), that all variables are continuous and that the polling directions are given by the canonical coordinate vectors \( \{ e^i \}_{i=1}^n \).

Two key observations allow to set up a strategy to exploit the partial separable structure in (1.1) that yields a new improved algorithm inspired by [49]. The first is that if a step along \( e^j \) is made, then only a subset of the element functions \( f_i \) will be affected and thereby need recomputation: only the \( f_i \) such that \( j \) appears in \( X_i \) are concerned. Secondly, if variable \( x_k \) does not occur in any of the \( f_i \)'s involving variable \( x_j \), a step of the form \( \alpha_j e^j + \alpha_k e^k \) (\( \alpha_j \) and \( \alpha_k \) are stepsizes) can be computed involving completely disjoint sets of element functions: those involving variable \( x_j \) and those involving variable \( x_k \). Crucially, the cost of this combined step is potentially much less than an unstructured step along direction \( e^j \) where the complete \( f \) would be evaluated.

**Example.** As an example, consider the partial separable function
\[ f(x_1, x_2, x_3, x_4, x_5) = f_1(x_1, x_2) + f_2(x_2, x_3) + f_3(x_1, x_2, x_4, x_5) + f_4(x_4, x_5) + f_5(x_4, x_5). \]  
(2.4)

Then, a step along \( e_1 \), say, only involves the element functions \( f_1 \) and \( f_3 \) while, since variable \( x_3 \) does not occur in \( f_1 \) and \( f_3 \), steps along \( \pm e_3 \) can be made without affecting computations along \( e_1 \). We say that \( x_1 \) and \( x_3 \) belong to disjoint sets of variables. Remarkably, the evaluation at points of the form \( \alpha_1 e_1 + \alpha_3 e_3 \) is less costly than a full function evaluation since it does not involve the evaluation of \( f_4 \) and \( f_5 \).

Exploiting these observations in a systematic way is the key to the new structured poll step. More specifically, once the user provides the sets \( X_i \) \((i = 1, \ldots, q)\) indicating the subset of \( \{1, \ldots, n\} \) defined by the components of \( x \) which appear in the domain of the element function \( f_i(i = 1, \ldots, q) \), then this structure is automatically first pre-processed as described in Algorithm 2.2.

As in [49], it can be verified that, for each \( h \in \{1, \ldots, t\} \),
\[ \mathcal{M}_h \overset{\text{def}}{=} \bigcup_{k \mid I_k \subseteq C_h \subseteq 1, \ldots, q} \mathcal{Y}_k \subseteq \{1, \ldots, q\}. \]  
(2.5)
Algorithm 2.2: Subspace structure detection

1. The structure \( \{ X_i \}_{i=1}^q \) is first inverted in that sets \( \{ E_j \}_{j=1}^n \) are built such that
   \[ i \in E_j \text{ if and only if } j \in X_i. \]

   The sets \( E_j \) contains the indices of all element functions \( f_i \) involving variable \( x_j \).

2. Using lexicographic sorting, sets of variables’ indeces \( \{ I_k \}_{k=1}^r \) corresponding to identical lists of elements \( E_j \) are constructed, that is
   \[ j_1 \in I_k \text{ and } j_2 \in I_k \text{ if and only if } E_{j_1} = E_{j_2} \overset{\text{def}}{=} Y_k. \]

   The indeces of the element functions in each \( Y_k \) are thus indistinguishable as far as their dependence of the problem’s variables is concerned.

3. Independent collections \( \{ C_h \}_{h=1}^t \) of \( I_k \) are then constructed by applying a greedy algorithm, independence being understood as the property that
   \[ I_{k_1} \subseteq C_h \text{ and } I_{k_2} \subseteq C_h \text{ if and only if } Y_{k_1} \cap Y_{k_2} = \emptyset. \]

Thus steps in variables belonging to different sets \( I_k \) in \( C_h \) can be computed independently (and in parallel). Moreover, this process only involves a subset \( M_h \) of the set of all element functions. Overall, the following collections are generated as a result of the procedure described in Algorithm 2.2: \( \{ X_i \}_{i=1}^q \) and \( \{ I_k \}_{k=1}^r \) containing variable indeces, \( \{ E_j \}_{j=1}^n \) and \( \{ M_h \}_{h=1}^t \) containing element function indeces and \( \{ C_h \}_{h=1}^t \) containing indeces of sets of variables.

**Example.** Returning to the the objective function described in (2.4), we have that \( q = 5, n = 5 \) and
\[
X_1 = \{1, 2\}, \ X_2 = \{2, 3\}, \ X_3 = \{1, 2, 4, 5\}, \ X_4 = \{4, 5\}, \ X_5 = \{4, 5\}
\]
and the corresponding sets of element functions’ indeces defined in Step 1 of Algorithm 2.2 are given by
\[
E_1 = \{1, 3\}, \ E_2 = \{1, 2, 3\}, \ E_3 = \{2\}, \ E_4 = \{3, 4, 5\}, \ E_5 = \{3, 4, 5\}.
\]

Step 2 aims at finding repeated lists of element functions, i.e. \( E_4 = E_5 \) in this example, which are eliminated in the new sets of variables’ and element functions’ indeces. Therefore proceeding by lexicographic sorting, we get \( r = 4 \) sets
\[
I_1 = \{1\}, \ I_2 = \{2\}, \ I_3 = \{3\}, \ I_4 = \{4, 5\},
\]
and
\[ Y_1 = \{1,3\}, \ Y_2 = \{1,2,3\}, \ Y_3 = \{2\}, \ Y_4 = \{3,4,5\}. \]
We then define the subspaces of independent variables using a simple greedy algorithm as follows. Initializing \( C_1 = \{I_1\} \), we start checking if \( Y_1 \cap Y_2 = \emptyset \). Since, \( Y_1 \cap Y_2 \neq \emptyset \) in our example, we now check if \( Y_1 \cap Y_3 = \emptyset \). Since this is true, \( I_3 \) is added to \( C_1 \). Then we check condition \( (Y_1 \cup Y_3) \cap Y_4 = \emptyset \). Since this condition fails, the definition of the set \( C_1 \) is completed and the next set \( C_2 \) is initialized as \( C_2 = \{I_2\} \). Then one checks the intersection of \( Y_2 \) with \( Y_4 \) (\( I_3 \) has already been assigned). Since the intersection is not empty, we finally obtain \( t = 3 \) collections of sets of independent variables
\[ C_1 = \{I_1, I_3\}, \ C_2 = \{I_2\}, \ C_3 = \{I_4\}. \quad (2.6) \]
The corresponding sets of element functions’ indeces are defined by
\[ M_1 = Y_1 \cup Y_3 = \{1,2,3\}, \ M_2 = Y_2 = \{1,2,3\}, \ M_3 = Y_4 = \{3,4,5\}, \]
In this examples, steps for variables in each collection \( C_i, i = 1,\ldots,3 \) involves \( |M_i| = 3 \) element function evaluations.

Once the \( \{C_h\}_{h=1}^t \) and \( \{M_h\}_{h=1}^t \) are known, we may define the subspaces
\[ S_k \overset{\text{def}}{=} \text{span} \{e^j \mid j \in I_k\}, \quad (2.7) \]
x_{S_h} the projection of the current iterate \( x \) onto \( S_k \) and the “inactive” index sets for the \( h \)-th collection as
\[ N_h = \{1,\ldots,n\} \setminus \bigcup_{k|I_k \subseteq C_h} I_k \quad \text{for} \ h \in \{1,\ldots,t\}. \]
The structured poll step then consists in performing poll steps in the subspaces \( S_k \) generated in each collection \( C_h \) of sets of independent variables and checking the sufficient decrease condition (2.1) in \( f \) by only evaluating elements functions in the corresponding set \( M_h \) (for \( k = 1,\ldots,r \) and \( h = 1,\ldots,t \)).

Our updated algorithm using a structured poll step is then given by Algorithm 2.3.
Algorithm 2.3: Pattern-search algorithm for partial separable problems

Initialization: The initial \( x \), the initial stepsize \( \alpha \), a convergence threshold \( \epsilon > 0 \) and the sets \( X_i \subseteq \{1, \ldots, n\} \) \( (i = 1, \ldots, q) \) specifying the connectivity pattern for problem (1.1) are given. The parameters \( \gamma, \iota \geq 1, \beta, \eta \in (0, 1) \) and \( n_2 < n \) are given.

Structure analysis: Compute a collection of sets of independent variables \( \{C_h\}_{h=1}^t \) using Algorithm 2.2. Initialize the stepsizes \( \alpha_k = \alpha \) for each set \( I_k \), \( k = 1, \ldots, r \) and define a set of orthonormal polling directions in \( S_k \), \( k = 1, \ldots, r \) defined in (2.7).

Until convergence
1. Search step: Ask the user to provide a new (potentially improved) approximate minimizer of \( f \), typically using problem specific modelling techniques;
2. Structured poll step: For \( h = 1, \ldots, t \) or until “sufficient decrease” in \( f \) is obtained
2.1 Poll step in \( C_h \): For each \( k \) such that \( I_k \subseteq C_h \) perform a standard poll step (using random orthogonal directions in \( S_k \) and stepsize \( \alpha_k \)) starting from \( x_{S_k} \) on the restricted function
\[
f_{Y_k} = \sum_{i \in Y_k} f_i,
\]
producing a potentially improved \( x_{S_k}^+ \) and decrease \( f_{Y_k}(x_{S_k}) - f_{Y_k}(x_{S_k}^+) \geq 0 \);
If sufficient decrease in \( f_{Y_k} \) is obtained, i.e. if \( f_{Y_k}(x_{S_k}^+) - f(x) < \eta \alpha_k^2 \), increase the stepsize \( \alpha_k \) by setting \( \alpha_k \leftarrow \gamma \alpha_k \); otherwise decrease \( \alpha_k \) by setting \( \alpha_k \leftarrow \beta^\iota \alpha_k \). Generate a new set of orthonormal polling directions in \( S_k \).
2.2 Iterate and function decrease update: Define a new iterate \( x^+ \) by
\[
x_{I_k}^+ = x_{S_k}^+ \quad \text{for each} \quad I_k \subseteq C_h
\]
and
\[
x_{N_h}^+ = x_{N_h} \quad \text{otherwise},
\]
and the corresponding objective function decrease by
\[
f(x^+) - f(x) = \sum_{k \mid I_k \subseteq C_h} \left[ f_{Y_k}(x_{S_k}) - f_{Y_k}(x_{S_k}^+) \right].
\]
If sufficient decrease in the objective function, i.e. if \( f(x^+) - f(x) < \eta \alpha_k^2 \), has been obtained, replace \( x \) by \( x^+ \) and terminate the structured poll step.
3. Termination step: Update the value of the “global” stepsize \( \alpha = \min_k \alpha_k \). If sufficient decrease in \( f \) was not obtained in the structured poll step and \( \alpha \) is not below the prescribed accuracy \( \epsilon \), go to Step 1. Otherwise enter the second pass (Step 4).
4. Second pass: Generate a set of orthonormal polling directions \( \{d_i\}_{i=1}^{n_2} \) in the full space and compute a complete poll loop along these directions with stepsize \( \alpha \). If sufficient reduction has been obtained, go to Step 1; else, declare convergence.
As is standard in a poll step, an objective function decrease may not be obtained in the structured poll step with the current choice of stepsize, in which case $x_S^k = x_S^k$ and $f_Y^k(x_S^k) = f_Y^k(x_S^k)$.

Once again we stress that using the successive collections $\{C_h\}_{h=1}^l$ is not mandatory (even if it clarifies the overall evaluation cost) and that poll steps can be performed on the subspaces $\{S_k\}$ using the restricted functions $\{f_Y^k\}$ completely independently and in parallel.

Compared with the method described in [49], the algorithm described in Algorithm 2.3 is both simpler and more efficient, as it performs complete poll steps using random orthogonal directions in each of the subspaces $S_k$ (instead of only retaining the best increment for steps along a fixed positive basis). Moreover, the mechanism used to adapt the stepsizes is more elaborate than the somewhat adhoc technique described in [49, Section 2.3]. It was also observed in practice that the stepsize reduction can be slightly faster than in the standard unstructured case. This faster reduction is translated in the algorithm by a stepsize shrinking factor $\beta \in (0, 1)$ (used in the unstructured case) which is raised to some power $\iota \geq 1$ (see Step 2.1 of Algorithm 2.3).

While the structured poll step is extremely efficient (as will be seen in the example below and in Section 4), it still has a potential drawback. Because the random poll directions are constrained to remain in each of the $S_k$, they are not random directions in the complete space $\mathbb{R}^n$, and convergence is obtained, following [27], to points where no further decrease can be obtained in each of these subspaces. This unfortunately does not imply that no further decrease can be obtained for $f$. A second pass is therefore necessary for obtaining this desirable property (Step 4 in Algorithm 2.3). This second pass does not use structure and is therefore considerably less efficient. However, this is mitigated by the observation that most (and sometimes all) the decrease in objective function value is obtained in the first pass, the second pass often only playing the role of a (possibly mildly expensive) convergence check. Moreover, if sufficient decrease is identified during the second pass, a return is made to the structure-using mechanism of the first, in an attempt to efficiently improve the decrease obtained. Fortunately, the overall efficiency of the structured minimization remains, in all examples we have seen, orders of magnitude better than that of the unstructured one. Moreover, it is not unusual for applications of derivative-free algorithms that the user is above all interested in obtaining a significant decrease in the objective function and not so much in extracting the last carat of decrease, let alone in checking local optimality. In this case, the second unstructured pass may often be unnecessary, bringing the optimization cost further down.

In order to reduce the cost of the second pass, our implementation of Algorithm 2.3 allows the user to specify a small number $n_2$ of random directions (typically much lower than $n$) for the second pass.

We now discuss the cost of applying the structured poll step of Algorithm 2.3 compared with that of using Step 2 of Algorithm 2.1. Consider an unsuccessful poll step first, that is a poll step during which sufficient decrease is not obtained. (Note that such
steps must occur as the stepsizes have to become sufficiently small for the algorithm to terminate.) Because of (2.5), we see that, for a given \( h \), the cost of evaluating the element function \( f_i \) for \( i \in M_h \) once cannot exceed that of evaluating \( f \), which we denote by \( c_f \). As forward and backward moves are considered for every polling direction (at an unsuccessful poll step), the total evaluation cost of the unsuccessful structured poll step is at most than \( 2tc_f \). By comparison, the cost of an unsuccessful unstructured poll step is equal to \( 2nc_f \). Since it is very often the case that \( t \ll n \) (see Table 4.1), the evaluation cost of the structured step is typically only a small fraction of that of the unstructured one. Because the poll step is terminated as soon as sufficient decrease is obtained, the cost of successful structured and unstructured poll steps is slightly more difficult to compare. As discussed in [27], the expected number of polling directions considered in a single successful (unstructured) poll step is small (typically 2 or 3). Two variants are however possible for the structured step. In principle, it can be terminated as soon as sufficient decrease is obtained on a given subspace \( S_k \). In the implementation discussed in Section 4, the loop on the subspaces associated to a collection \( C_h \) is always completed before sufficient decrease triggers poll-step termination. This choice appears to be efficient and allows for parallel execution of the subspace-restricted poll steps for different subspaces. Its evaluation complexity therefore depends on the number of subspace collections \( C_h \) examined (which is at most \( t \)). If this number is also a very small integer (as is often the case), the evaluation cost of the structured poll step is very similar to that of the unstructured one. However, the objective function decrease obtained is typically much larger, as it corresponds to a number of unstructured successful poll steps given by the total number of subspaces \( S_k \) considered in the calculation. Of course, further gains may be obtained in the structured case if the parallelism between the subspaces \( S_k \) is also exploited.

When bound constraints are present, the backward and forward steps within the poll step are truncated to prevent computation of the objective function at infeasible points, but the rest of the calculation is essentially unchanged.

**Example.** We now give a first taste of the effectiveness of the structured poll step. Consider an unconstrained problem with objective function structured as in (2.4) and whose element functions are given by

\[
\begin{align*}
    f_1(x) &= \sqrt{x_1^2 + x_2^2}, \\
    f_2(x) &= (\sin x_1 - 23x_2x_1)^2, \\
    f_3(x) &= (x_3^3 - 56x_2x_3)^2 - x_4, \\
    f_4(x) &= \max\{|x_4|, |x_5|\}, \\
    f_5(x) &= \sqrt{x_4^2 + x_5^2}.
\end{align*}
\]

(2.8) (2.9) (2.10) (2.11) (2.12)

We observe that the full objective \( f \) is bounded below but the element \( f_3 \) is unbounded in variable \( x_4 \). However, the collection of independent sets in (2.6) ensures that the
restricted function \( f_{y_4}(x_{S_4}) \) is still bounded below when computing poll steps in the sub-
space \( S_4 \) (the only one involving \( x_4 \)). In fact, the elements functions involved include \( f_3 \)
together with \( f_4 \) and \( f_5 \). We then apply both the unstructured and structured variants of
the method described above with a tolerance \( \epsilon \) on the stepsize equal to \( 10^{-5} \). Figure 2.1
shows the two resulting convergence histories in terms of the number of evaluations of
the complete \( f \). The advantage of using the structured poll step is striking.

![Figure 2.1: The evolution of \( f \) as a function of the number of (complete) evaluations for
the objective defined in (2.8)-(2.12). Zoom on the function evaluations involved in the
second pass when the partially separable structure is exploited (black curve).]

The effect of the second pass is visible in the magnified part of the black curve (blue
circle). The first flat segment indicates that the second pass is entered, but sufficient
reduction is soon detected along a direction in the full space and the structured poll step
is performed again (vertical segment). Then, the second pass is entered again (second
flat segment) and convergence is finally declared.

3 A callable library of structured models

After discussing how to exploit the coordinate partially separable structure (1.1) in the
poll step, we now briefly describe a multivariate interpolation technique that permits its
exploitation in the user-controlled search step (i.e. Step 1 of Algorithms 2.1 and 2.3).
As is common for such steps (see [33], for instance), the idea is to provide a surrogate
model of the objective function in the neighbourhood of the current iterate \( x_{best} \), which
is built using information gathered in the course of the algorithm. This model can then
be minimized (typically within some trust region) to provide an improved guess of the
minimizer. We now describe a library named BFOSS for computing such a step while
exploiting structure. As in [11, 53], it considers a surrogate model whose structure
mirrors that of (1.1) in that it shares the same coordinate partially separable definition.
This is achieved by constructing separate multivariate polynomial interpolation models
for each of the \( f_i \) (restricted to their free variables). General multivariate polynomial
interpolation models follow the principles of [13, 55, 16], while their use in the context of partially separable problems is inspired by [11, 53] and the necessary adaptations used here for handling the bound constraints are similar to those discussed in [28]. While the general idea of structured models is not new, our experience suggests that its practical performance does depend on a number of more detailed decisions at the implementation level. We therefore focus, in what follows, on the algorithmic aspects which are specific to our approach and relevant to the numerical comparison to be conducted in Section 4.

We first describe our strategy for the unstructured case \((q = 1)\) and then specialize it to the partially separable case \((q \geq 1)\).

Let \(Y = \{y_i\}_{i=1}^p\) be a sample set and let \(m(x)\) denote a polynomial of degree \(d (d = 1, 2)\) interpolating \(f\) at the points in \(Y\), that is satisfying the interpolation conditions

\[
m(y_i) = f(y_i), \quad i = 1, \ldots, p. \tag{3.1}
\]

The polynomial \(m(x)\) can be expressed as a linear combination of elements of the natural basis \(\phi\) for the space of polynomial of degree at most 2, i.e. the basis of monomials, that we assumed to be ordered as follows

\[
\phi = \{1, x_1, x_2, \ldots, x_n, \frac{1}{2}x_1^2, \ldots, \frac{1}{2}x_n^2, x_1x_2, \ldots, x_{n-1}x_n, x_1x_3, \ldots, x_{n-2}x_n, \ldots, x_1x_n\}.
\]

Finding a model that satisfies conditions (3.1) is equivalent to solving a potentially underdetermined linear system of the form

\[
M_\phi z = f(Y) \tag{3.2}
\]

with coefficient matrix \(M_\phi\) of dimension \(p \times \bar{p}\), \(z \in \mathbb{R}^\bar{p}\) and \((f(Y))_i = f(y_i), i = 1, \ldots, p\), where \(p \leq \bar{p}\) and \(\bar{p}\) is the chosen number of elements from the basis \(\phi\).

If the coefficient matrix \(M_\phi\) is square and nonsingular, then the model \(m\) is unique. If \(p < \bar{p}\) the linear system is underdetermined, the resulting interpolating polynomials may not exist or may no longer be unique, and different approaches to construct the model \(m(x)\) are possible [16]. BFOSS allows the user to choose between two possibilities: a model can be built by taking the minimum 2-norm solution of (3.2) [21] or, alternatively, a sub-basis \(\tilde{\phi}\) of \(p\) elements is extracted from \(\phi\) yielding the corresponding square matrix \(M_{\tilde{\phi}}\) (the components of \(z\) corresponding to the removed columns are set equal to zero).

As a result of the search step, a new tentative iterate \(x_+\) is then computed by minimizing the model \(m\) in the intersection between the trust-region and the feasible set, that is

\[
\{x_{\text{best}} + s : \max(l - x_{\text{best}}, -\Delta) \leq s \leq \min(u - x_{\text{best}}, \Delta)\} \tag{3.3}
\]

where \(\Delta > 0\) is the current trust-region radius, and \(f(x_+)\) is evaluated. Computing \(x_+\) within the trust-region (3.3) is a well-understood question and there are several algorithms available for the task (see [12, Chapter 7] for an overview). BFOSS allows the user to choose between a bound-constrained variant of the Moré-Sorensen trust-region algorithm and a projected gradient trust-region algorithm to solve (3.3) [38, 8, 57, 56].
The trust-region radius $\Delta$ is updated on the basis of the ratio of the achieved to predicted reduction $\rho = (f(x_+) - f(x_{\text{best}}))/(m(x_+) - m(x_{\text{best}}))$ as follows:

$$\Delta = \begin{cases} 
\min\{\alpha_1 \Delta, \Delta_{\text{max}}\} & \text{if } \rho > \eta_2 \\
\Delta & \text{if } \rho \in [\eta_1, \eta_2] \\
\max\{\epsilon, \alpha_2 ||s||\} & \text{else}
\end{cases}$$

where $\alpha_1 > 1$, $\alpha_2 \in (0, 1)$, $\Delta_{\text{max}} > 0$ and $0 < \eta_1 < \eta_2 < 1$ are trust-region parameters [12, Chapter 17] and $\epsilon$ denotes the machine precision. When BFOSS is called the first time in Step 1 of Algorithm 2.1, the value of $\Delta$ is set equal to the current step size. At all subsequent calls, if $\Delta$ is too small, i.e. $\Delta < \Delta_{\text{min}}$ defined by the user, the radius $\Delta$ is restarted to the current step size. If $\Delta$ is below $\Delta_{\text{min}}$ after this restart, BFOSS is terminated. Finally, all the newly evaluated points ($x_+$ and, when relevant, the new interpolation points) are returned to Algorithm 2.1 with their associated function values.

It is well-known that the fact that the model $m$ is well-defined not only depends on the number of points in $Y$, but that a further geometric condition (known as poisedness) is also required (see [16] for details). We now describe the specific strategy used in BFOSS to define the interpolation set $Y$, which combines the use of models of various degrees and types and ensures the necessary poisedness condition.

When using BFOSS, the user selects a value of $\bar{p}$ in the set \{n + 1, 2n + 1, (n + 1)(n + 2)/2\} corresponding to a linear model, a quadratic one with a diagonal Hessian and a full quadratic one. A first tentative set $Y$ of $\bar{p}$ points is built around $x_{\text{best}}$ extracting from the recorded history (past points and function values), the closest points to $x_{\text{best}}$ in the 2-norm and their associated function values (including $x_{\text{best}}/f(x_{\text{best}})$). If the history is too short in that it only contains $p < \bar{p}$ points, BFOSS uses all the available $p$ points according to one of the strategy described above for the case $p < \bar{p}$.

Then matrix $M_{\phi}$ in (3.2) is then built, together with its generalized inverse $M_{\phi}^\dag$. The generalized inverse $M_{\phi}^\dag$ is constructed from a truncated SVD of $M_{\phi}$, where “redundant” singular values in a standard SVD are zeroed. We say that a singular value is redundant if its value is lower than the maximum singular value scaled by a user-defined parameter $k_{\text{ill}}$ that measure the maximum ill-conditioning allowed. (3) Finally, the pseudoinverse is computed from this regularized SVD. When redundant singular values are detected, the corresponding points in $Y$ are progressively removed, replaced by random points and $M_{\phi}^\dag$ rebuilt. Importantly, this crucial phase does not require the computation of additional function values.

Once $M_{\phi}$ is considered sufficiently well-conditioned (within $k_{\text{ill}}$) the poisedness of the current set $Y$ is measured by computing the maximum absolute value of the Lagrange polynomials in the neighbourhood. Based on this measure, following [54], some points in $Y$ can be replaced by some “far” points available in the history but not used so far. Then, $Y$ is possibly further improved by performing exchanges until the improvement

---

(3) By default $k_{\text{ill}} = \infty$ and $k_{\text{ill}} = 10^{12}$ for structured and unstructured problems, respectively. These values resulted from our numerical experiments.
in poisedness becomes moderate enough. Finally, the obtained (reasonably conditioned) $M^*_q$ matrix is used to define the linear combination of the Lagrange polynomials which interpolates function values at the interpolation points.

The adaptation of this strategy to the partially separable case is straightforward: one simply applies the same technique to define interpolation models $m_i$ for each of the element functions $f_i(x_i)(i = 1, \ldots, q)$. Note that each model $m_i$ has at most $n_i$ variables and approximate $f_i$'s around the projection of $x_{\text{best}}$ onto the subspace $\mathbb{R}^{n_i}$. The final trial point $x_+$ is then computed by minimizing the global quadratic model

$$m(x_{\text{best}} + s) = \sum_{i=1}^{q} m_i(x_{\text{best}} + s),$$

in the box (3.3).

4 Numerical illustration

We now report numerical experiments comparing and combining the two techniques described above. The results were obtained by implementing the structured poll-step described in Algorithm 2.3 with the BFO (the Brute Force Optimizer) package [42] and exploiting the BFOSS library in combination with this upgraded version of BFO\textsuperscript{(4)}. BFO is a random pattern search algorithm proposed by the authors in [42], whose structure is identical to that of Algorithm 2.1. Since a full description of this package is somewhat involved and many of its features irrelevant for our present discussion, we avoid restating it here in detail, and refer the reader to [42] for an in-depth description.

The main objectives of the numerical tests are

- to illustrate the impact of structure usage both at rather global level (via the use of the structured poll step) and a local level (by building local interpolation/regression models using BFOSS),
- to discuss the relative merits of the two techniques,
- to investigate the effectiveness of their combination.

We first provide some numerical illustration of our claim that the structured poll step is more efficient than the unstructured one. For this purpose, we compare the structured\textsuperscript{(5)} and unstructured version of BFO\textsuperscript{(6)} on a set of variable dimension test problems extracted from CUTEst [26] and/or already used in [49] for the most part.

\textsuperscript{(4)}The BFOSS library is distinct from the BFO package itself –they come in different files– although it interacts with it. But its use or even presence is not necessary for running the main package.

\textsuperscript{(5)}In the BFO implementation of the new poll step, $\iota$ is a newly introduced algorithmic parameter which, like all such parameters [42], can be (and has been) trained for improved performance.

\textsuperscript{(6)}It may be recalled that the unstructured BFO was shown in [42] to be quite competitive, in particular when compared with NOMAD [33].
For each problem, we considered dimensions ranging from around 10 to 10000, whenever solvable in reasonable time\(^{(7)}\). They are partitioned in four dimension-dependent test sets: small, smallish, medium and large. Variables are all considered to be continuous although the structured poll step is implemented in BFO for handling integer and categorical variables as well. The performance analysis with these variables are out of the scope of this work and will be considered in the future. Their main characteristics are detailed in Table 4.1, but we now give some additional information.

- The original BEALE problem from CUTEst only has two variables. The variant BEALES used here is obtained by juxtaposing \(n/2\) copies of the original problem, resulting in a totally separable problem (hence the \(S\)). It is useful as it exposes how well a method can exploit such an important structure.

- The function NZF1 is derived from that published in [49]. Its precise formulation is detailed in Appendix C. It reduces to the version of [49] for \(n = 13\).

- The CONTACT problem is a bound-constrained minimum surface problem involving nonlinear surface boundaries and an obstacle from below the surface. It is described in Appendix B.

- The CUTEst minimum-surface problem NLMINSRF only differs from LMINSURF (also in CUTEst) in that the surface boundary is nonlinear.

- The BROWNAL6 problem is that presented in [49] under the name ‘Brown almost linear’.

- In the original MOREBV problem, the distance from the starting point to the solution decreases with dimension, which makes it less interesting for large problems. We use here a version of the problem where the original starting point is multiplied by \(\log_{10}(n)\) to compensate.

- The JNLBRNG1 is the journal bearing problems of MINPACK2 [5].

- Problems BROYDN3D, CONTACT, ENGVAL, JNLBRNG1, LMINSURF, NLMINSRF, MOREBV are discretized problems. They illustrate how frequently the CPS structure appears in the situation: one variable associated to a particular location in the problem typically only depends on the variables associated with neighbouring locations, and not with all of them.

These problems are typical of a very large class of medium/large-scale applications, where, although \(q\) obviously depends on \(n\), both \(t\) and \(\max_k |I_k|\), the maximal dimension of any subspace \(S_k\), do not.

Our experiments were conducted using a new release of BFO, which contains both the new structure-exploiting poll step and the BFOSS library. In order to train the

\(^{(7)}\)Eight hours using Matlab R2017b on a Intel(R) Xeon(R) CPU E3-1245 v5 @ 3.50GHz machine with 64 GB RAM.
Table 4.1: Characteristics of the test problem instances

<table>
<thead>
<tr>
<th>Problem</th>
<th>small</th>
<th>smallish</th>
<th>medium</th>
<th>large</th>
<th>q</th>
<th>max_k</th>
<th>t</th>
<th>max_k</th>
<th>I_k</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARWHEAD</td>
<td>10</td>
<td>50, 100</td>
<td>500, 1000</td>
<td>5000, 10000</td>
<td>n − 1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>BDARWHD</td>
<td>10</td>
<td>50, 100</td>
<td>500, 1000</td>
<td>5000, 10000</td>
<td>n − 2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>BEALEX</td>
<td>10</td>
<td>50, 100</td>
<td>500, 1000</td>
<td>5000, 10000</td>
<td>n − 4</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>BROYDN3D</td>
<td>10</td>
<td>50, 100</td>
<td>500, 1000</td>
<td>5000, 10000</td>
<td>n/2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>BROWNAL6</td>
<td>10</td>
<td>50, 102</td>
<td>502, 1002</td>
<td>5002, 10002</td>
<td>(n − 2)/4</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>CONTACT</td>
<td>15</td>
<td>64, 144</td>
<td>400, 900</td>
<td>2500, 4900</td>
<td>(\sqrt{n} − 1)^2</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ENVAL</td>
<td>10</td>
<td>50, 100</td>
<td>500, 1000</td>
<td>5000, 10000</td>
<td>n − 1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>DIXMAANA</td>
<td>15</td>
<td>51, 102</td>
<td>501, 1002</td>
<td>5001, 10002</td>
<td>n</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>DIXMAANI</td>
<td>15</td>
<td>51, 102</td>
<td>501, 1002</td>
<td>5001, 10002</td>
<td>n</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>FREUROTH</td>
<td>10</td>
<td>50, 100</td>
<td>500, 1000</td>
<td>5000, 10000</td>
<td>n − 1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>HELIX</td>
<td>11</td>
<td>21, 101</td>
<td>501, 1001</td>
<td>5001, 10001</td>
<td>(n − 1)/2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>JNLBLRNG1</td>
<td>24</td>
<td>64, 144</td>
<td>400, 900</td>
<td>2500, 4900</td>
<td>4(n_t + 1)^2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>LMINSURF</td>
<td>16</td>
<td>64, 144</td>
<td>400, 900</td>
<td>2500, 4900</td>
<td>(\sqrt{n} − 1)^2</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>MOREBV</td>
<td>12</td>
<td>52, 102</td>
<td>502, 1002</td>
<td>5002, 10002</td>
<td>n</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NLINMSRF</td>
<td>16</td>
<td>64, 144</td>
<td>400, 900</td>
<td>2500, 4900</td>
<td>(\sqrt{n} − 1)^2</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>NZF1</td>
<td>13</td>
<td>39, 130</td>
<td>650, 1300</td>
<td>6500, 13000</td>
<td>(7n/13) − 2</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>POWSING</td>
<td>20</td>
<td>52, 100</td>
<td>500, 1000</td>
<td>5000, 10000</td>
<td>n/4</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>ROSENBR</td>
<td>10</td>
<td>50, 100</td>
<td>500, 1000</td>
<td>5000, 10000</td>
<td>n/2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>TRIDIA</td>
<td>10</td>
<td>50, 100</td>
<td>500, 1000</td>
<td>5000, 10000</td>
<td>n</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>WOODS</td>
<td>20</td>
<td>40, 200</td>
<td>400, 2000</td>
<td>4000, 10000</td>
<td>n/4</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Characteristics of the test problem instances

parameter \( \eta \) defining the faster stepsize decrease in the structured optimization pass, we selected, for each of the above problems, the instance of third smallest dimension (mostly \( n \approx 100 \)) and performed training to improve the resulting data profile (see [43]). The experimental guess of \( \eta = 1.25 \), was only very marginally improved\(^{(8)}\) to 1.2550. Note that other BFO algorithmic parameters were set to their default values. For small problems, the reported results are an average of 30 independent runs, for smallish 10 runs for medium ones, 5 runs and a single run for the large ones. All the results discussed in this section are reported in Table 1.3 in Appendix A.

4.1 Exploiting structure in the poll step

The results obtained by running the new structure-exploiting version of BFO (with the trained \( \eta \)) are presented in Table 4.2. This table reports the number of complete function evaluation to obtain an approximate\(^{(9)}\) minimizer for each of the instances of Table 4.1. For each instance, we give the number of required objective-function (full) evaluations for structure exploiting (first) and standard (second, no structure exploitation) versions.

\(^{(8)}\) We used the BFO default training accuracy requirement \( \epsilon_t = 0.01 \).

\(^{(9)}\) We used the BFO default accuracy requirement \( \epsilon = 0.0001 \). We also stress that the algorithmic parameters in BFOSS have been trained for improved performance (as all other parameters of BFO) on the small test-set, using the data profile performance measure [43]. The values of the trained parameters are set as default in BFO and BFOSS.
It clearly results from Table 4.2 that it is possible to solve large partially-separable problems without using derivatives at an acceptable cost in number of function evaluations. Using the structure is crucial if the problem size exceeds ten or so. In fact, a significant fraction of the problems of that size can’t be solved at all in a reasonable number of evaluations if structure is neglected: direct derivative-free methods like BFO proceed by sampling, and this technique badly suffers from the curse of dimensionality.

A comparison of the above results with those of Table 2 in [49] also shows that the structured BFO significantly outperforms the simpler algorithm presented in that reference.

We now illustrate our comments of Section 2 about the relative efficiency of the two polling passes. Figures 4.2 to 4.4 show, for the BROYDEN3D problems in dimensions ten to one thousand, the objective function decreases obtained by the new BFO using structure (black line) and that obtained by the standard version of BFO which ignores structure (magenta line). These decreases are expressed as a function of the number of (complete) objective-function evaluations.

The rate of decrease of the structured BFO is very clearly much faster than that of BFO \(^{(10)}\).

When partially separable structure is provided, the number of full function evaluation per BFO iteration is retrieved by summing the number of element function evaluations and dividing by the number of elements (and rounding to integer).

Table 4.2: The numbers of objective function evaluations required by the structured/unstructured versions of BFO for the problem instances of Table 4.1, as a function of increasing problem size (∞ meaning that more than 100000 evaluations were needed).

<table>
<thead>
<tr>
<th>Problem</th>
<th>small</th>
<th>smallish</th>
<th>medium</th>
<th>large</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARWHEAD</td>
<td>79/962</td>
<td>91/11589</td>
<td>97/36085</td>
<td>146/∞</td>
</tr>
<tr>
<td>BDARWHD</td>
<td>101/2152</td>
<td>70/∞</td>
<td>71/∞</td>
<td>81/∞</td>
</tr>
<tr>
<td>BDEXP</td>
<td>1661/21122</td>
<td>13218/∞</td>
<td>34409/∞</td>
<td>∞/∞</td>
</tr>
<tr>
<td>BDQRTIC</td>
<td>290/2468</td>
<td>301/∞</td>
<td>298/∞</td>
<td>393/∞</td>
</tr>
<tr>
<td>BEALES</td>
<td>275/∞</td>
<td>275/∞</td>
<td>275/∞</td>
<td>275/∞</td>
</tr>
<tr>
<td>BROYDN3D</td>
<td>308/1225</td>
<td>199/2244</td>
<td>273/7458</td>
<td>304/∞</td>
</tr>
<tr>
<td>BROWNAL6</td>
<td>15773/6171</td>
<td>4325/∞</td>
<td>∞/∞</td>
<td>∞/∞</td>
</tr>
<tr>
<td>DEXP</td>
<td>101/2152</td>
<td>221/30452</td>
<td>442/∞</td>
<td>1009/∞</td>
</tr>
<tr>
<td>DIXMAAN6</td>
<td>185/2358</td>
<td>189/17357</td>
<td>240/55748</td>
<td>226/∞</td>
</tr>
<tr>
<td>DIXMAAN1</td>
<td>185/17702</td>
<td>186/∞</td>
<td>184/∞</td>
<td>230/∞</td>
</tr>
<tr>
<td>ENVAL</td>
<td>143/1567</td>
<td>155/30400</td>
<td>157/∞</td>
<td>159/∞</td>
</tr>
<tr>
<td>FREUROTH</td>
<td>57/83101</td>
<td>57/∞</td>
<td>57/∞</td>
<td>57/∞</td>
</tr>
<tr>
<td>HELIX</td>
<td>265/268</td>
<td>221/30452</td>
<td>442/∞</td>
<td>1009/∞</td>
</tr>
<tr>
<td>JNLBLNG1</td>
<td>101/734</td>
<td>301/27567</td>
<td>427/∞</td>
<td>944/∞</td>
</tr>
<tr>
<td>LMINSURF</td>
<td>101/734</td>
<td>301/27567</td>
<td>427/∞</td>
<td>944/∞</td>
</tr>
<tr>
<td>MOREBV</td>
<td>412/415</td>
<td>579/29409</td>
<td>1082/∞</td>
<td>8853/∞</td>
</tr>
<tr>
<td>NLMINSRF</td>
<td>177/1480</td>
<td>225/61772</td>
<td>625/∞</td>
<td>684/∞</td>
</tr>
<tr>
<td>NORMA</td>
<td>177/1480</td>
<td>225/61772</td>
<td>625/∞</td>
<td>684/∞</td>
</tr>
<tr>
<td>POWSING</td>
<td>716/20605</td>
<td>824/∞</td>
<td>849/∞</td>
<td>988/∞</td>
</tr>
<tr>
<td>ROSENSB</td>
<td>361/13241</td>
<td>361/∞</td>
<td>384/∞</td>
<td>436/∞</td>
</tr>
<tr>
<td>TRIDIA</td>
<td>440/3073</td>
<td>316/∞</td>
<td>314/∞</td>
<td>345/∞</td>
</tr>
<tr>
<td>WOODS</td>
<td>1609/∞</td>
<td>1747/∞</td>
<td>1924/∞</td>
<td>2241/∞</td>
</tr>
</tbody>
</table>

\(^{(10)}\)When partially separable structure is provided, the number of full function evaluation per BFO iteration is retrieved by summing the number of element function evaluations and dividing by the number of elements (and rounding to integer).
Figure 4.2: The evolution of $f$ as a function of the number of (complete) evaluations for \textsc{broyden3d} for $n = 10$ and $n = 50$

Figure 4.3: The evolution of $f$ as a function of the number of (complete) evaluations for \textsc{broyden3d} for $n = 100$ and $n = 500$
the unstructured variant. Moreover, the two passes of the structure-exploiting algorithm are very noticeable, the first pass being significantly faster and yielding most of the final decrease. We observe that the length of the second pass increases with size, which is expected because the space to sample in the second pass increases in dimension. However, because the first pass already made such good progress, the effort spent in the second pass remains acceptable, at least for the problem sizes considered here. By contrast, the performance of the standard unstructured version of BFO quickly degrades with size: for small dimensions, one notices the staircase-like decrease which is typical of pattern search methods. In this test problem, the first poll step is unsuccessful, leading to a full $2n$ function evaluations before the stepsize is reduced. While this can be acceptable for small $n$, it becomes more problematic as $n$ grows. For larger problems, the structured variant has already terminated before the first poll step is completed in the unstructured method. For the instance in 10 variables, one also notices (in the left picture of Figure 4.2) that it can be beneficial to reapply the structured poll-step mechanism of the first pass after a sufficient decrease in the second pass (cfr. Figure 2.1).

The behaviour of the structured variants on problem BDEXP finally merits a comment. This problem features a very flat objective function near the solution and, while the objective function value is decreased quickly, the mechanism of the pattern search method takes many iterations to declare optimality and terminate.

### 4.2 Exploiting structured models in the search step

We now turn to illustrating the performance which can be obtained using the BFOSS model library described above, and also compare it with the pure (structured and unstructured) sampling strategy of BFO alone. In the experiments reported next, we use fully quadratic models, the Moré-Sorensen method for maximizing the Lagrange polynomials in the trust-region and the truncated conjugate-gradient algorithm to solve (3.3).

In this section we use performance and data profiles [24, 39] to compare different
variants of BFO. For this purpose, we measure performance in terms of numbers of full objective function evaluations necessary for termination that is declared when the following condition holds
\[ f(x_0) - f(x) \geq (1 - \tau)(f(x_0) - f_\ast). \] (4.1)
Here \(x_0\) is the starting point for the problem, \(x\) is the solution returned by a solver, \(f_\ast\) is computed for each problem as the smallest value of \(f\) obtained by any solver within a given number \(\mu_f\) (\(\mu_f = 100000\) in our tests) of function evaluations, \(\tau \in [0, 1]\) is a tolerance that represents the the percentage decrease from the starting value \(f(x_0)\) (we used the standard value \(\tau = 10^{-4}\)). The stopping criterion (4.1) is suggested in [39] to generate profiles and differs from the default stopping criterion on the minimum step size used in BFO by default.

In Figure 4.5 four variants of BFO are compared on the small test. These variants are

- **unstructured**: the standard unstructured BFO algorithm without using the BFOSS models,
- **models**: the standard unstructured BFO algorithm where a full-dimensional BFOSS search step is attempted at every iteration,
- **ps**: the version of BFO using the coordinate partially-separable structure, but without using the BFOSS models,
- **ps & models**: the version of BFO using the coordinate partially-separable structure, using BFOSS models for each element function at every iteration.

This figure shows that, for small problems, the combined use of models and structure is the best algorithmic choice, but also indicates that using structure without models is clearly preferable to using models without structure. An interpretation of this observation is that providing global information on the problem (structure) outperforms approximating local one (models). The situation is less clear when the size of the problems increases, as is shown in Figures 4.6 and 4.7, where one compares the performance of the 'ps' and 'ps & models' variants on the smallish and medium test sets (the two other variants fail for a large proportion of the smallish problems). Profiles on the large test-set have not been generated since 'ps' was the only variant that could solve the test set in the maximum time allowed (8 hours).

In Figures 4.6 and 4.7, we see that the relative advantage obtained by the use of models for small problems progressively vanishes to disappear completely when the problem size grows.

While comparing the number of function evaluations, as we have done above, is most natural for derivative-free problems (the cost of an evaluation in real world applications often dominating that of all computations internal to the algorithm), it is also interesting to consider memory usage and internal computing effort. The pure unstructured variant is clearly the most economical from both points of view (but at the price of being the
Figure 4.5: Performance and data profiles for unstructured/structured variants, with or without models (small test set).

Figure 4.6: Performance and data profiles for structured variants, with or without models (smallish test set).
Performance profiles with $\tau = 10^{-4}$

Data profiles with $\tau = 10^{-4}$

Figure 4.7: Performance and data profiles for structured variants, with or without models (medium test set).

less efficient in function evaluations). Because of the need to store the model itself, memory usage can significantly increase for the variant using unstructured models (at least for fully quadratic ones as discussed above). This is also the case for the structured variants because they have to store, analyze (once) and exploit the structure, which requires using additional pointers and lists. However, the additional memory necessary to use models in the structured case remains typically modest, as only a collection of small matrices needs being stored. We finally note that variants using models require a significantly higher internal computational effort, mostly in the solution of the many trust-region subproblems involved in managing the interpolation set and computing the search step. This is apparent in Figure 4.7 and in Table 1.3 in Appendix A where we report the number of function evaluations taken by all variants on all test sets using the internal default stopping criterion. It is clear from the table that the worse reliability of the 'ps & models' variant is nearly entirely due to exhausting the maximum allowed cpu-time (symbol ‘†’ in the table). Numerical experiments were also carried out using linear and diagonal models in the search step (both in the structured and unstructured cases). Unfortunately, it was observed that, despite the solution of the involved trust-region problem being considerably cheaper than when using quadratic models, the use of the simpler models resulted in a considerable increase of the number of function evaluations, an thus worse overall performance.

5 Conclusions and perspectives

We have introduced a structured version of derivative-free random pattern search algorithms which is able to exploit coordinate partially separable structure (typically as-
associated with sparsity) present in unconstrained and bound-constrained optimization problems. This technique improves performance by orders of magnitude and makes it possible to solve problems that otherwise are totally intractable by standard derivative-free methods.

We have also described a library of interpolation-based modelling tools which can be associated to the structured or unstructured versions of the initial pattern search algorithm. For problems of small or moderate size, the use of the library further enhances performance, especially when associated with structure. For larger problems the internal computing costs increase and, while still reducing the number of function evaluations, the use of the library may require a sometimes unrealistic computing time, in particular if the problem is unstructured.

In comparing the benefits of using problem structure in a poll step and building local models using interpolation techniques, we have concluded that the former is likely to be more efficient, in particular for larger problems, even for structure-exploiting models.

A new release of the Matlab BFO package\(^\text{(11)}\) featuring both use of structure and modelling tools (as discussed in this paper) is now freely available online from

\texttt{https://github.com/m01marpor/BFO}

and the main new features are briefly sketched in Appendix D.

The selective use of structured models in conjunction with structured poll steps remains an attractive option. However, the criteria defining the circumstances in which interpolation models should be used, if at all, need further investigation. Many other topics also merit research, including the design of “grayer” optimization tools which could exploit derivatives available for part of the problem while adapting the techniques described here for the rest.

**Acknowledgements**

Both authors are indebted to the University of Florence for its support while the present research was being conducted.

The first author is member of the Gruppo Nazionale per il Calcolo Scientifico (GNCS) of the Istituto Nazionale di Alta Matematica (INdAM) and this work was partially supported by INdAM-GNCS under Progetti di Ricerca 2019 and 2020.

**References**


\(^\text{(11)}\) Although this is unrelated to the subject of the present paper, we emphasize that the new BFO release also features the important capability of handling categorical variables, as well as the new training strategies discussed in [43], see Appendix D.


A Full results

Table 1.3 shows the number of objective function evaluations required by the structured/unstructured versions of BFO with and without interpolation model for the problem instances of Table 4.1 (∞ meaning that more than 100000 evaluations were needed, † meaning that cpu time exceed the maximum time, 8 hours, allowed). Moreover, a blank entry means that the solution of the corresponding problem instance was not attempted either because the solver version failed to solve a smaller version of the same problem (symbols † and ∞) or because the problem size was considered too large for the considered solver. In particular, the unstructured/model implementation was used for solving small problems only and the partially separable/model implementation was not applied to solve large problems).

B Details for problem CONTACT

The problem CONTACT is a bilinear finite-element discretization of a membrane contact problem defined on the unit square of $\mathbb{R}^2$. This square is discretized in $q = (\sqrt{n} - 1)^2$ element squares and the surface of the membrane “above” this element square is given by

$$\frac{1}{q} \sqrt{1 + (x_{SW} - x_{NE})^2 + (x_{SO} - x_{NW})^2},$$

where $x_{SW}$, $x_{NE}$, $x_{SO}$ and $x_{NW}$ give the height of the membrane at the four corners of the element square. The membrane is fixed on the boundaries of the unit square to the value of the function

$$b(x, y) = 1 + 8x + 4y + 3 \sin(2\pi x) \sin(2\pi y).$$

(2.1)

In addition, the membrane has to lie above a square flat obstacle of height 10 positioned at $[0.4, 0.6]^2$. The starting point is the projection of the vector given by the values of (2.1) at each element square’s corner onto the feasible domain.

C Details for problem NZF1

The NZF1 problem is a variable dimension version of the eponymous nonlinear least-squares problem defined in [49] as an example illustrating the concept of partially sepa-
Table 1.3: The number of objective function evaluations required by the structured/unstructured versions of BFO with and without interpolation model for the problem instances of Table 4.1.
variable functions. If $n = 13\ell$, it is defined by

$$
f(x_1, \ldots, x_n) = \sum_{i=1}^\ell \left[ (3x_i - 60 + \frac{1}{m}(x_{i+1} - x_{i+2})^2)^2 + \left( \frac{x_{i+5}}{1 + x_{i+4}^2 + \sin(\frac{x_{i+4}}{1000})} \right)^2 \right]$$

$$
+ \left( x_{i+6} + x_{i+7} - x_{i+8}^2 + x_{i+10} \right)^2
+ \left( \log(1 + x_{i+10}^2) + x_{i+11} - 5x_{i+12} + 20 \right)^2
+ \left( x_{i+4} + x_{i+5} + x_{i+5} \times x_{i+9} + 10x_{i+9} - 50 \right)^2$$

$$
+ \sum_{i=1}^{\ell-1} (x_{i+6} - x_{i+19})^2.
$$

The starting point is the vector of all ones.

### D BFO 2.0 and its new features

Release 2.0 of the Matlab BFO package is a major upgrade from Release 1 and includes several important new problem-oriented possibilities. Beyond the BFOSS library of model-based search steps described in Section 3 and the exploitation of CPS structure described in Section 2, BFO 2.0 also supports the following new problem features.

**Categorical variables.** In addition to standard continuous and discrete variables, BFO now supports the use of categorical variables. Categorical variables are unconstrained non-numeric variables whose possible 'states' are defined by strings (such as 'blue'). These states are not implicitly ordered, as would be the case for integer or continuous variables. As a consequence, the notion of neighbourhood of a categorical variable is entirely application-dependent, and has to be supplied, in one form of two possible forms, by the user. Moreover, the 'vector of variables' is no longer a standard numerical vector when categorical variables are present, but is itself a *vector state* defined by a value cell array of size $n$ (the problem's dimension), whose $i$-th component is either a number when variable $i$ is not categorical, or a string defining the current state of the $i$-th (categorical) variable. For example, such a 4-dimensional vector state can be given by the value cell array

```matlab
{{ 'blue', 3.1416, 'magenta', 2 }}.
```

Variable $i$ is declared to be categorical by specifying `xtypen(i) = 's'`. If a problem contains at least one categorical variable, it is called a categorical problem and optimization is carried on vector states (instead of vectors of numerical variables). As
a consequence, the starting point and the returned best minimizing point are vector states, and the objective function’s value is computed at vector states (meaning that the argument of the function $f$ is a vector state).

The user must specify the application-dependent neighbours (also called categorical neighbourhoods) of each given vector state with respect to its categorical variables. This can be done in two mutually exclusive ways.

1. The first is to specify static neighbourhoods. This is done by specifying, for each categorical variable, the complete list of its possible states. The neighbourhood of a given vector state $vs$ wrt to categorical variable $j$ (the hinge variable) then consists of all vector states that differ from $vs$ only in the state of the $j$-th variable, which takes all possible values different from $vs(j)$. In this case, all variables of the problem retain their (initial) types and lower and upper bounds.

2. The second is to specify dynamical neighbourhoods. This more flexible technique is used by specifying a user-supplied function whose purpose is to compute the neighbours of the vector state $vs$ ‘on the fly’, when needed by BFO. At variance with the static neighbourhood case, the variable number and types, as well as lower and upper bounds of the neighbouring vector states (collectively called the ‘context’) may be redefined within a framework defined by a few simple rules.

In effect, this amounts to specifying the neighbouring nodes in a (possibly directed) graph whose nodes are identified by the list, types, bounds and values of the variables. As a consequence, the user-supplied definition of the neighbour(s) of one such node may need to take the values of all variables into account. Of course, for the problem to make sense, it is still required that the objective function can be computed for the new neighbouring vector states and that its value is meaningfully comparable to that at $vs$.

The very substantial flexibility allowed by this mechanism of course comes at the price of the user’s full responsibility for overall coherence.

**Performance and data profile training strategies.** Because BFO is a trainable package (meaning that its internal algorithmic constants can be trained/optimized by the user to optimize its performance on a specific problem class), it needs to define training strategies which allow to decide if a particular option is better than another. Release 1 of BFO included the natural “average” training criterion (quality is measured by the average number of function evaluations on the class) and a robust variant of the same idea (see [42] for details). Release 2.0 now includes two new training strategies (introduced in [43]):

**Performance profiling.** When this training option is selected, the performance of two algorithmic variants (i.e. versions of BFO differing by the value of their internal algorithmic parameters) are compared using the well-known performance profile methodology [24, 25].
Data profiling. This option is similar to performance profiling, but uses data profiles [39] instead of performance profiles to compare two variants.

These new options correspond more closely to the manner in which derivative-free packages are compared in the optimization literature.