BFO, a trainable derivative-free brute force optimizer for nonlinear bound-constrained optimization and equilibrium computations with continuous and discrete variables

BFO, a trainable derivative-free Brute Force Optimizer for nonlinear bound-constrained optimization and equilibrium computations with continuous and discrete variables

MARGHERITA PORCELLI*, Università degli Studi di Firenze
PHILIPPE L. TOINT, University of Namur

A direct-search derivative-free Matlab optimizer for bound-constrained problems is described, whose remarkable features are its ability to handle a mix of continuous and discrete variables, a versatile interface as well as a novel self-training option. Its performance compares favourably with that of NOMAD, a well-known derivative-free optimization package. It is also applicable to multilevel equilibrium- or constrained-type problems. Its easy-to-use interface provides a number of user-oriented features, such as checkpointing and restart, variable scaling and early termination tools.

CCS Concepts: • **Mathematics of computing** → **Mathematical software performance**; **Mixed discrete-continuous optimization**; *Numerical analysis*;

Additional Key Words and Phrases: Derivative-free optimization, direct-search methods, mixed integer optimization, bound constraints, trainable algorithms.

1 INTRODUCTION

The efficient solution of optimization problems arising in real applications increasingly calls for the development of efficient and easy-to-use implementations of derivative-free algorithms. In applicative contexts such as engineering design [Conn et al. 1998; Leonetti et al. 2012], medical image registration [Oeuvray and Bierlaire 2007] and design of algorithms [Audet and Orban 2006] (amongst many others), optimization problems are often defined by functions computed by costly simulation. A single simulation performed to evaluate the costly function may, for instance, require the solution of large systems of partial differential equations or a even a costly measurement campaign, and hence, may take from a few minutes to many hours or days depending on the particular application. Functions have therefore to be treated as expensive black-boxes and due to the high computational cost involved, it is important to use optimization algorithms that produce reasonably good solutions with a limited number of function evaluations. Moreover, optimization variables can be of different nature: continuous (e.g., geometrical parameters), integer (e.g., on/off element of a structure) or more generally categorical variables, which are discrete variables which identify an element of an ordered or unordered set (e.g., colors, shapes, or materials). It is also fairly common to have restrictions on the expected size of each variable which can be formulated as bound constraints. In some situations, the presence of bound constraints can prevent the computation of solutions which have no physical meaning. Furthermore, it is often helpful to introduce reasonable bounds on the variables when there is a good guess of the domain where solutions are expected.

In order to model problems encompassing that complexity, we start by considering the following bound-constrained mixed variables nonlinear programming problem

$$\min_{x\in\Omega} f(x) \tag{1}$$

where $f : \Omega \to \mathbb{R}$ is a possibly nonsmooth (or even non continuous) function and the domain Ω is partitioned into continuous, ordinal discrete and fixed variables Ω^c, Ω^d and Ω^f of dimension n_c, n_d

*This author wishes to dedicate this work to her little Alice.

and n_f, respectively. The domains are bound constrained, i.e., $\Omega^c = [l^c, u^c]$, where $l^c, u^c \in \mathbb{R}^{n_c} \cup \{\pm\infty\}$, $l^c \leq u^c$, $\Omega^d = [l^d, u^d]$, where $l^d, u^d \in \mathbb{Z}^{n_d} \cup \{\pm\infty\}$, $l^d \leq u^d$, and Ω^f can be interpreted as $\Omega^f = [l^f, u^f]$, where $l^f, u^f \in \mathbb{R}^{n_f}$ are equal. Specifically, discrete variables are required to have values in an ordered and equally-spaced set. Let l, u be n-dimensional vectors such that $l^T = ({l^c}^T, {l^d}^T, {l^f}^T)$ and $u^T = ({u^c}^T, {u^d}^T, {u^f}^T)$, with $n = n_c + n_d + n_f$. Let $C = \{i \in \{1, \dots n\} \mid x_i \text{ is continuous } \}$ and $\mathcal{D} = \{i \in \{1, \dots n\} \mid x_i \text{ is discrete } \}$ be index sets of the continuous and the discrete variables ($n_c = |C|, n_d = |\mathcal{D}|$). We assume that the evaluation of the objective function is time-consuming, and that no derivative is available. Moreover, no convexity assumption is made.

We also consider the multilevel problem of the min-max type

$$\min_{x_1 \in \Omega_1} \max_{x_2 \in \Omega_2} \dots \min_{x_m \in \Omega_m} f(x_1, \dots, x_m) \tag{2}$$

where x_1 to x_m form a partition of the variables in m "levels", for which specific feasible sets $\Omega_1, \dots, \Omega_m$ are given and where the objective function may be minimized or maximized (at the user's choice) at each level. Furthermore, we allow that each of the set Ω_ℓ ($1 < \ell \leq m$) to depend on the values of the variables in $x_1, \dots, x_{\ell-1}$. This problem arises in the solution of non-smooth mixed-integer leader-follower equilibrium problems. As an example, consider the case where a producer of two goods optimizes the values and prices of these goods, given production costs depending on price and limits on prices depending on values, and given a demand for the two goods. The demand is determined by a class of consumers who, in turn, optimize at their level the perceived values of the goods they buy under a budget constraint. The first good can only be bought in integer quantities while the second good is bulk and can be bought in real quantities.

The very general nature of the problem(s) effectively limits the choice among algorithm's classes for its solution to that of "direct-search methods", that is methods based on the (somewhat brute force) exploration of the variables' domain based on local sampling (for instance using pre-specified geometric patterns) and restricted to comparing objective function values (without interpolation or other type of modeling). A good introduction to direct-search methods may be found in Chapter 7 of the book by Conn, Scheinberg and Vicente [Conn et al. 2009]. These methods have a long history in the optimization literature (see Nelder and Mead [Nelder and Mead 1965], Hookes and Jeeves [Hooke and Jeeves 1961], Box and Wilson [Box and Wilson 1951]) and have proved to be very popular among users, mostly because of their ease of use and robustness. A few direct-search methods can be applied to problem (1) in continuous variables although they involve some elements of objective function modeling: this is the case of NOMAD by Abramson et al. [Abramson et al. 2009, 2008], SID-PSM by Custódio and Vicente [Custódio et al. 2010; Custódio and Vicente 2007] and NMLSR/NMDFU by Grippo and Rinaldi [Grippo and Rinaldi 2015]. On the other hand, to our knowledge, existing literature for solving (1) where the variables are both continuous and discrete is not very extensive. Papers by Audet and Dennis [Audet and Dennis 2001] and by Lucidi, Piccialli and Sciandrone [Lucidi et al. 2005] as well the paper by Abramson et al. [Abramson et al. 2009] (describing NOMAD) consider the mixed case and [Liuzzi et al. 2012, 2014] deals with problems whose variables are mixed-integer. Other methods of interest include HOPSPACK (Hybrid Optimization Parallel Search Package) [Gray et al. 2008, 2010], which provides a general set of tools for pattern search exploiting parallel computing, but does not handle multilevel problems explicitly and MAPS (Model-Assisted Pattern Search) [Siefert et al. 1997] which uses pattern search on purely continuous problems.

We finally mention the interesting paper by Gratton et al. [Gratton et al. 2015], which provides convergence and complexity analysis for a wide class of direct-search methods applied to smooth problems with continuous variables. As is the case in our proposal, the methods of this class use random choices of search directions.

The purpose of the present paper is to present a new algorithm of the direct-search class for finding local solutions of problems (1) or (2) which handles mixed variables and also has the novel feature to be trainable by users to (typically application dependent) families of problems for improved efficiency.

We discuss a Matlab code associated with this algorithm that is accessible through the web site https://sites.google.com/site/bfocode/ together with examples of use [1]. It is important to note that the BFO code can be used to tune general algorithms depending on a modest number of parameters and whose performance can be measured. Moreover, the authors are not aware of any other software package aimed at solving min-max problems (2), although bi-level min-max problems in continuous variables have been considered in [Bard 1998; Colson et al. 2007; Conn and Vicente 2012; Dempe 2002] and in [Gümüş and Floudas 2005] when variables are allowed to be mixed-integer, and linear min-max problems have been studied in [Tang et al. 2015; Vicente et al. 1996].

The paper is organized as follows. Section 2 presents the algorithm itself in the context of the optimization problem (1). Section 3 is dedicated to the numerical validation of BFO and discusses how the algorithm can be used to optimize algorithmic parameters in itself or other numerical methods (Subsection 3.2). It also discusses a numerical comparison with NOMAD (Subsection 3.3). Section 4 describes how the parameter tuning feature of BFO can be used to make BFO itself trainable by users. Section 5 then considers how the algorithm can be adapted to problems of the form (2). Finally, some conclusions and perspectives are outlined in Section 6.

Notation. Let I_q denote the Identity matrix of dimension $q \times q$. For any $v \in \mathbb{R}^q$ and $\mathcal{K} \subset \{1, \ldots, q\}$, we write $v_{\mathcal{K}}$ for the subvector of v having components v_i, $i \in \mathcal{K}$. Furthermore, if V is a $q \times q$ matrix, we denote by V_{ij} the ij-th element of V and by $V_{:,i}$ the i-th column. Given the matrices $B \in \mathbb{R}^{m \times p}$ and $C \in \mathbb{R}^{m \times q}$, let $[B\ C] \in \mathbb{R}^{m \times (p+q)}$ denote the matrix concatenation by columns and let $W_{p,q}$ denote a $p \times q$ matrix of uniformly distributed random numbers on the (0,1) interval.

2 THE BFO ALGORITHM

We start by outlining the general features of the new BFO[2] algorithm for solving the optimization problem (1). BFO generates a sequence of feasible iterates whose objective function value is decreasing. Its underlying structure is that of a direct-algorithm: at any given iteration, the objective function is evaluated at a finite number of points on a local[3] mesh in the neighbourhood of the current iterate, in an attempt to find a new point with a lower objective function value, which then becomes the next iterate. It is hoped that the sequence of such iterates approaches a local minimizer of problem (1).

More specifically, each iteration of the algorithm is initiated with the current iterate $\bar{x}$ and the current function value $\bar{f} = f(\bar{x})$ as well as with an enumerable set $\mathcal{P}$ of polling directions for both continuous and discrete variables (there are no polling directions for fixed variables). These directions implicitly define the current local mesh as the set of all points in $\Omega^c \times \Omega^d$ which can be reached from $\bar{x}$ by a move along one of the directions in $\mathcal{P}$ with a fixed stepsize. Note that stepsizes and directions are fixed by the problem definition for discrete variables (for instance, the direction must be a coordinate vector and the step size has to be integer if the variable is integer), but both can be varied for continuous variables in the course of the numerical solution. Therefore, the set $\mathcal{P}$

[1] The code version used in this paper is: BFO version 1.00, 23 IX 2015.
[2] For Brute Force Optimizer.
[3] A local mesh is a mesh around the current iterate whose geometry may vary from iteration to iteration both in orientation and spacing.

contains the columns of the matrix

$$P \begin{pmatrix} Q & 0 \\ 0 & I_{n_d} \end{pmatrix} P^T, \tag{3}$$

where P is a permutation matrix ordering the continuous variable first and Q is a matrix of dimension $n_c \times n_c$ whose columns are updated at each iteration. For the sake of simplicity, we assume in what follows that P is the identity matrix.

At the first iteration of BFO, an initial solution $\bar{x}$ and an initial set $\mathcal{P}$ of polling directions is given, as well as a vector of stepizes $\delta = \delta^0 \in \mathbb{R}^{n_c} \times \mathbb{Z}^{n_d}$. The algorithm then proceeds by exploring the local mesh specified by these elements until no further improvement is possible. The continuous stepsizes δ_C are then decreased, the search directions for continuous variables are possibly updated and the process repeated.

Exploration of a given local mesh is conducted in one or two phases. In the first phase (called the POLL STEP and detailed in Section 2.1), a search is performed by computing forward and backward steps along all the current search directions from the current iterate. If $n_c \neq 0$, this search is terminated prematurely if sufficient decrease in the objective function is detected after a move in a direction corresponding to a continuous variable. If this first phase does not succeed in improving the current point $\bar{x}$ and $n_d \neq 0$, a second phase is entered. In this phase, a further search is performed by exploring the subproblems defined by fixing successively each of the discrete variables to a value neighbouring (by a move with the proper stepsize) that present in $\bar{x}$. This second phase is performed by recursively calling the algorithm itself for the solution of each such subproblem. This phase is called the RECURSIVE STEP and is detailed in Section 2.2.

These (possible) two phases are followed by the TERMINATION STEP (see Section 2.3). If a point with a better objective value than $\bar{f}$ is found in either phase, then the iteration is declared *successful*, otherwise the iteration is declared *unsuccessful*. In the successful case, the better point becomes the current iterate, and a new iteration is started with a coarser local mesh in the continuous variables (EXPANSION SUBSTEP). In the unsuccessful case, either the grid may be further refined so that a new iteration is initiated at the same current solution, but with a finer local mesh on the continuous variables (REFINEMENT SUBSTEP), or a check is performed to declare convergence on the finest local mesh (CHECK-CONV SUBSTEP).

A general outline of BFO is given in Algorithm 1 and its steps are described in detail in the following subsections. Each iteration of the algorithm consists in performing the three steps.

Note that it is also possible to include an additional "surrogate search step" before the TERMINATION STEP of this algorithm. A possibility is to exploit any available approximation of the objective function constructed either from the available function values (see the "BFGS-finish" option in Section 2.4) or from additional specific evaluations.

2.1 The POLL STEP

The POLL STEP is a standard feature of direct-search algorithms: given the current iterate and function value $(\bar{x}, \bar{f})$, the objective function is evaluated at forward and backward mesh points in search of a better pair (x^{best}, f^{best}) such that $f^{best} < \bar{f}$. The algorithm therefore performs a loop on both continuous and discrete variables, moving along the current direction as follows. Consider the i-th variable and assume first that $i \in C$. Let $Q \in \mathbb{R}^{n_c \times n_c}$ be a matrix whose columns form an orthonormal basis in Ω^c and let δ_i be the current stepsize. The continuous components of the forward poll point are then given by

$$x_C^{fwd} = \bar{x}_C + \alpha_f Q_{:,i}$$

Algorithm 1 An outline of the BFO algorithm for solving (1)

Initialization. Let $\bar{x} \in \Omega^c \times \Omega^d \times \Omega^f$ and $\bar{f} = f(\bar{x})$, the initial stepsize $\delta = \delta^0 \in \mathbb{R}^{n_c} \times \mathbb{Z}^{n_d}$ and the initial set of polling directions $\mathcal{P}$ whose columns are chosen according to (3). Set the initial best value $x^{best} = \bar{x}$, $f^{best} = \bar{f}$, the initial decrease $\Delta_f = \infty$ and $\eta \in (0, 1)$.

Until convergence

(1) POLL STEP. Perform a search loop on the variables, moving forward and backward along the directions in the set $\mathcal{P}$ with stepsize δ. If a poll point x^p constructed in this way is found such that $f(x^p) < f^{best}$, then set $x^{best} = x^p$ and $f^{best} = f(x^p)$. Terminate the search loop as soon as a poll point corresponding to a continuous variable satisfies

$$\bar{f} - f(x^p) \geq \eta \, \Delta_f. \tag{4}$$

If the loop is not terminated by (4) and $f^{best} < \bar{f}$, update

$$\Delta_f = \bar{f} - f^{best}. \tag{5}$$

If $f^{best} < \bar{f}$ or $n_d = 0$, go to the TERMINATION STEP.

(2) RECURSIVE STEP. If requested, apply the BFO algorithm to solve the subproblems defined by fixing each of the discrete variables to a value differing from that in $\bar{x}$ by plus or minus the corresponding stepsize. If a point x^r is found such that $f(x^r) < f^{best}$, then set $x^{best} = x^r$ and $f^{best} = f(x^r)$.

(3) TERMINATION STEP. If $f^{best} < \bar{f}$ [successful iteration], update $\bar{x} = x^{best}$ and $\bar{f} = f^{best}$, increase δ_C, update the set $\mathcal{P}$ for continuous variables (i.e., the columns of Q) and go to the POLL STEP (EXPANSION SUBSTEP).
Else [unsuccessful iteration], check for convergence (CHECK-CONV SUBSTEP).
If convergence is not declared, decrease δ_C, update the matrix Q and go to the POLL STEP (REFINEMENT SUBSTEP).

where

$$\alpha_f = \min_{i \in C} \alpha_i, \text{ with } \alpha_i = \begin{cases} \min\{u_i - \bar{x}_i, \delta_i\}/Q_{ii} & \text{if } Q_{ii} > 0, \\ \min\{l_i - \bar{x}_i, -\delta_i\}/Q_{ii} & \text{if } Q_{ii} < 0, \\ \infty & \text{else.} \end{cases}$$

If $i \in \mathcal{D}$, then the i-th component of the forward poll point is simply given by

$$x_i^{fwd} = \min\{\bar{x}_i + \delta_i, u_i\}.$$

The backward poll point associated with the i-th variable is computed analogously, setting

$$x_C^{bwd} = \bar{x}_C - \alpha_b Q_{\cdot,i}$$

where

$$\alpha_b = \min_{i \in C} \alpha_i, \text{ with } \alpha_i = \begin{cases} \min\{\bar{x}_i - l_i, \delta_i\}/Q_{ii} & \text{if } Q_{ii} > 0, \\ \min\{\bar{x}_i - u_i, -\delta_i\}/Q_{ii} & \text{if } Q_{ii} < 0, \\ \infty & \text{else,} \end{cases}$$

if $i \in C$, and

$$x_i^{bwd} = \max\{\bar{x}_i - \delta_i, l_i\}$$

if $i \in \mathcal{D}$. The current best function value f^{best} is then updated if $f^{fwd} = f(x^{fwd}) < f^{best}$ or $f^{bwd} = f(x^{bwd}) < f^{best}$.

Condition (4) results in a greedy/opportunistic type of polling where the search terminates as soon as significant decrease is found.

In order to avoid unnecessary computation, BFO also keeps track of 'fixed variables', that is variables whose lower and upper bounds are equal. Such variables are simply skipped in the poll-step loop.

Whenever the loop over all continuous non-fixed variables is completed, the POLL STEP is also exploited to compute an estimate of the projected gradient size for the continuous variables (whether this gradient exists or not) using the formula

$$\epsilon_{cr} = \| \max(l_C, \min(x_C - Q^T g_{dif}, u_C)) - x_C \|, \tag{6}$$

where, for $i \in C$,

$$(g_{dif})_i = \frac{f_i^{fwd} - f_i^{bwd}}{\|x_i^{fwd} - x_i^{bwd}\|}.$$

The value of ϵ_{cr} is not used by the algorithm but is supplied as an informational output to the user.

2.2 The RECURSIVE STEP

BFO allows the user to require further exploration of subspaces defined by fixing discrete variables. As is common in many techniques for exploring fixed subsets of integer variables, we model this exploration by a tree, where a child subset (or, in our case, subspace) is obtained from its father subset by fixing an additional variable, the root subset being that with no fixed variables at all. BFO allows the user to choose between the "depth-first" and "breadth-first" strategies to recursively explore the subspace tree. In the latter, all subspaces corresponding to potentially interesting values of the discrete variables are explored before grid refinement. In contrast, grid refinement is performed as soon as possible (before exploring other possible subspaces for the same mesh size) when depth-first search is chosen.

More specifically, let j be the index of a discrete variable and assume that a recursive subspace exploration is entered starting from either x^{fwd} or x^{bwd}. Let also F be the index of the discrete variables which are already fixed at $\bar{x}$. BFO is then called to solve

$$x^r = \operatorname*{argmin}_{l \le x \le u} f(x) \quad \text{subject to} \quad x_i \text{ fixed for } i \in \{j\} \cup F,$$

where we use the 'fixed variable' feature mentioned in the previous paragraph.

In our implementation, the index $j \in \mathcal{D}$ is selected in increasing order starting from 1, but this choice is admittedly arbitrary. If breadth-first search is chosen, the recursion is called every time the POLL STEP did not produce an improved f^{best} (and discrete variables are present) and the mesh size of the child subproblem is inherited from the father calling problem. On the other hand, if depth-first search is employed, the recursion starts if both the POLL STEP is unfruitful and the search in the current subspace has terminated with a CHECK-CONV SUBSTEP. In this case, the mesh size of the subproblem is reset to the user-defined initial one. In both strategies, the recursive call ends when convergence is declared in the CHECK-CONV SUBSTEP.

2.3 The TERMINATION STEP

After the POLL STEP and (possibly) the RECURSIVE STEP, the algorithm evaluates the set

$$\mathcal{S} = \{i \in C,\ x_i^{best} - l_i \le \delta_i \text{ or } u_i - x_i^{best} \le \delta_i\},$$

of nearly saturated bounds at x_{best}, sets $n_s = |\mathcal{S}|$ and determines the matrix N of normals of these nearly saturated constraints. Then, if the iteration is successful, i.e., $f^{best} < \bar{f}$, it performs the EXPANSION SUBSTEP, itself consisting of three parts. Firstly, the mesh size for continuous variables

is increased by a constant factor by setting

$$\delta_C \leftarrow \min\left[(u-l)_C,\ \min(\alpha\delta_C, \gamma\delta_C^0)\right], \tag{7}$$

where $\alpha \geq 1$ and $\gamma > 0$ are grid expansion factor and maximum grid expansion factor, respectively. Secondly, given an integer parameter `inertia` and the "progress direction" defined by

$$\Delta^{avg} = \sum_{j=1}^{\texttt{inertia}} \Delta x_{:,j}^{acc} \tag{8}$$

where Δx^{acc} is the $n_c \times$ `inertia` matrix whose columns contains the directions of descent $x_{best} - \bar{x}$ accumulated over the last `inertia` iterations at the same recursion level, a new set of orthonormal directions $Q^{new} \in \mathbb{R}^{n_c \times n_c}$ is computed from the QR factorization

$$[N\ \ \Delta^{avg}\ \ W_{n_c, n_c - n_s - 1}] = Q^{new} R$$

for some upper-triangular matrix R. We recall that n_c denotes the number of continuous variables and W is a matrix of uniformly distributed random numbers over (0,1) of dimension $n_c \times (n_c - n_s - 1)$ which is generated at each iteration. This change of basis has the effect of redefining the continuous variables, of projecting the progress direction onto the nullspace of the nearly saturated bounds and of ensuring that the normals of the nearly saturated constraints belong to the new basis. This latter property is ensured by the fact that the nearly active bound constraints normals are the first columns of the above matrix, and thus that their orthonormal nature is preserved by the QR factorization. Thirdly, the new current iterate is redefined by $\bar{x} = x^{best}$, $\bar{f} = f^{best}$ and a new iteration started.

If, by contrast, the iteration is unsuccessful, the algorithm enters the CHECK-CONV SUBSTEP and checks for termination in the sense that convergence is (preliminarily) declared if

$$|(\delta_C)_i| \leq \epsilon, \quad i \in C, \tag{9}$$

where $\epsilon > 0$ is a mesh-size threshold. In this case, further attempts to reduce the objective function are performed by a user-specified number of poll steps, each using a new randomly drawn orthonormal basis Q^{new} obtained from the QR factorization

$$[N\ W_{n_c, n_c - n_s}] = Q^{new} R. \tag{10}$$

If condition (9) is met every time, final convergence of BFO is declared and x^{best} is returned to the user as the best approximation solution found.

If this convergence test fails, the REFINEMENT SUBSTEP is then entered, where, given a grid shrinking factor $\beta \in (0, 1)$, the grid for the continuous variables is refined by setting

$$\delta_C \leftarrow \max\{\epsilon/2,\ \beta\delta_C\} \tag{11}$$

and Δ_f in (5) is reduce by the factor β. Analogously to the procedure used in the EXPANSION SUBSTEP, the new basis Q^{new} is defined from the QR factorization (10) and a new iteration is started.

The BFO parameters are summarized in Table 1. The same parameters are used at each level of the recursion.

Despite obvious similarities between BFO and the class of algorithms considered by Gratton at al. [Gratton et al. 2015], convergence of the BFO algorithm for unconstrained continuous problems is not guaranteed by their analysis. Indeed their framework requires that a decrease in objective function value which is is not sufficient (in terms of a given forcing function) is not accepted, while BFO accepts any decrease. However, the theory of grid-based methods (see Coope and Price [Coope

$p_\#$	Parameter	Type	Description
p_1	α	c	The grid expansion factor (see (7))
p_2	β	c	The grid shrinking factor (see (11))
p_3	γ	c	The maximum grid expansion factor (see (7))
p_4	δ	c	The initial stepsize vector (see Algorithm 1)
p_5	η	c	The sufficient decrease fraction in the poll step (see (4))
p_6	`inertia`	i	The inertia for continuous step accumulation (see (8))
p_7	`stype`	i	The discrete tree search strategy {BF, DF, none} (see Section 2.2)

Table 1. Table of BFO parameters.

and Price 2001]) can be invoked to deduce convergence of the BFO algorithm in this case, provided the successive sets of polling directions $\mathcal{P}$ become dense in the unit sphere.

Another difference between the framework of [Gratton et al. 2015] and the present proposal follows. The BFO strategy can be interpreted as a sequential conditional choice where a random direction is selected first and, if this direction does not produce sufficient descent, its opposite is most likely to provide descent. Moreover, if neither of these moves is successful, an obvious choice it to look in their orthogonal complement. The framework of [Gratton et al. 2015] does not cover this last option.

Finally, the BFO algorithm presents obvious opportunities for exploiting parallel computations. One can, for instance, perform all the evaluations corresponding to the poll step in parallel (the condition (4) becomes irrelevant in this case). This is especially useful when integer variables are present because it allows to run the recursive steps in parallel. Parallelism can also be exploited in the training phase, where the computation of the performance measures can be improved by solving test problems in parallel. Although these developments are of interest, we do not discuss them in this paper.

2.4 Additional BFO features

The BFO code also provides a few additional facilities for the user, which we briefly describe.

termination on objective function target: In some applications, the cost of complete optimization is simply too high, especially in the context of derivative-free methods where asserting convergence may itself be a reasonably costly algorithmic phase. Many users are therefore more interested in obtaining a decent decrease/increase of the objective function from a starting value than in pursuing optimization to its conclusion. BFO allows the user to terminate optimization before convergence specifying a "target value" for the objective function, and the algorithm then terminates as soon as this target is attained.

cheap unsuccessful objective evaluation: Many relevant optimization problems occur in the form where the objective function consists of a sum of positive terms (such as nonlinear least-squares). In these cases, it is possible to save significant computational effort by determining, in the course of the evaluation of the objective function itself, if the accumulated value for a given number of terms already results in a value too large for the evaluation to be considered successful by the algorithm. BFO provides the necessary interface to allow stopping evaluation by ignoring further terms.

We note that algorithm training as described in Section 4 is by nature a problem where this feature can be applied. Indeed its use results in a 10% saving in evaluations when training BFO with the AO strategy, compared to the naive version ignoring this structure. When combined with the use of objective function targets (just described), the computational

cost of RO training strategy typically decreases by an order of magnitude. For instance, in the nonlinear least-squares example described in Section 4.1, the total cost (in time and operations) of computing the RO objective function decreases by a factor 10 because of this strategy. This situation is typical in our experience.

graceful handling of undefined function values: If the objective function value is undefined at a point x where evaluation is attempted, returning NaN to BFO will allow the code to proceed and continue optimization, excluding x from the set of possible solutions.

variable-dependent scaling: It may happen that variables in a problem have different scalings, resulting in ill-conditioning and termination difficulties if this property is ignored. BFO allows the user to explicitly specify variable scalings in order to avoid this type of detrimental numerical behaviour. This is achieved by specifying a scaling vector (`xscale`) whose values are used to (statically) scale the variables before problem solution is attempted. A change in `xscale` yields a corresponding change in the initial stepsize δ.

user-specified discrete lattice: By default, BFO considers discrete variables as integer, but also provides the possibility for the user to specify a discrete lattice on which optimization must be carried on. This is done by passing to BFO a matrix whose columns corresponding to discrete variables contain a basis of this lattice. Optimization on the i-th (discrete) variable is then interpreted as optimization along fixed multiples of the i-th column of the given matrix.

user-specified search-step: At the end of each poll step, BFO allows the user the possibility to suggest a guess for a good descent point. Typically, this point is the minimizer of some kind of model (e.g., interpolation, regression, Kriging, RBF, etc.). As an option, the user may specify a function, that, given a set of points where the objective function as been evaluated in the poll step, (possibly) returns a better point.

checkpointing and restart: Because optimization with costly objective function may be time consuming, it is useful for an algorithm to provide checkpointing and restart facilities. This is the case in BFO, where the user may specify the checkpointing frequency and the name of the checkpointing file(s).

BFGS finish: When convergence is approached on smooth problems, the grid is refined following iterations where no improvement can be obtained in the polling loop. However, a complete polling loop provides enough function evaluations to allow for a central difference estimation of the gradient at the current iterate. This in turn can be exploited at successive iterates of this type, the associated differences in (estimated) gradient being used to build a BFGS [Broyden 1970; Fletcher 1970; Goldfarb 1970; Shanno 1970] variable-metric approximation of the (assumed) second derivatives, at least in the presence of positive curvature. More specifically, let x and x^+ be two successive iterates at which the polling loop has terminated after considering all directions in $\mathcal{P}$ and let g_{dif} and g^+_{dif} be the corresponding gradient approximations. The quasi-Newton step may then be computed to enforce the secant condition

$$B(x^+ - x) = g^+_{dif} - g_{dif},$$

where B is the resulting approximate Hessian matrix. This facility is provided as an option in BFO, and often results in significantly higher accuracy of the solution for a moderate increase in function evaluations.

a CUTEst **interface:** In order to facilitate comparison with other packages, an interface to the CUTEst testing environment [Gould et al. 2015] is also provided.

3 NUMERICAL EXPERIMENTS

This section is devoted to the numerical validation of the Matlab implementation of BFO. This issue is carried out through two series of experiments performed on a given set of benchmark problems: the first series is used for fine-tuning the BFO parameters and the second for the comparison of the practical behavior of BFO with that of a competitor solver chosen to represent the class of derivative-free solvers.

The comparative computational analysis is carried out by using performance and data profiles proposed in [Moré and Wild 2009] for benchmarking derivative-free optimization algorithms. Following this reference, we compare different solvers and algorithmic versions declaring that the problem is solved as soon as

$$f(x_0) - f(x) \geq (1 - \tau)(f(x_0) - f_*) \tag{12}$$

where x_0 is the starting point for the problem, x is the solution returned by a solver, f_* is computed for each problem as the smallest value of f obtained by any solver within a given number μ_f of function evaluations, $\tau \in [0, 1]$ is a tolerance that represents the the percentage decrease from the starting value $f(x_0)$. In practice (12) measures the function value reduction $f(x_0) - f(x)$ achieved by x relative to the best possible reduction $f(x_0) - f_*$.

Let $\mathtt{fe}_{P,S}$ denote the total number of function evaluations needed for the solver S to solve problem P, that is to satisfy (12) for a given tolerance τ, and let $\mathtt{fe}_P$ be the total number of function evaluations employed by the best solver to solve problem P.

We consider the classical *performance profile* function π_S defined as

$$\pi_S(t) = \frac{\text{number of problems s.t. } \mathtt{fe}_{P,S} \leq t\, \mathtt{fe}_P}{\text{number of problems}}, \quad t \geq 1,$$

which is interpreted in [Dolan and Moré 2002] as the probability for solver S that a performance ratio $\mathtt{fe}_{P,S}/\mathtt{fe}_P$ is within a factor t of the best possible ratio.

We also consider a further performance measure, the *data profile* function [Moré and Wild 2009], which computes the percentage of problems that can be solved (for a given tolerance τ) within a certain number of function evaluations ν (the "budget"). The data profile function is defined as

$$\delta_S(\nu) = \frac{\text{number of problems s.t. } \mathtt{fe}_{P,S} \leq \nu\,(n+1)}{\text{number of problems}}, \quad \nu > 0.$$

With the scaling $n + 1$, $\delta_S(\nu)$ can be interpreted as the percentage of problems that can be solved with the equivalent of ν simplex gradient estimates.

We note that the performance profile $\pi_S(t)$ measures how well the solver S performs relative to the other competitive solvers on a given set of problems, while the data profile $\delta_S(\nu)$ for a given solver S is independent of other solvers.

In our experiments, we allowed a maximum number of 10000 function evaluations and considered two levels of accuracy $\tau = 10^{-4}$, and 10^{-8}. In order guarantee the satisfaction of the condition (12) for these values of τ, we used tight tolerances in the solver converging tests. Finally, performance and data profiles are plotted in the following sections using an horizontal logarithmic scale and selecting $t > 1$ and $\nu \in [0, 2500]$, respectively. Experiments were carried out using Matlab R2012a on Intel Core 2 Duo U7006 @1.2GHx2, 1.5 GB RAM.

3.1 The benchmark problems

We consider problems from the CUTEst test collection [Gould et al. 2015] both for testing and training BFO. The test examples are selected using the CUTEst interactive select tool in order to

locate the subset of bound constrained problems and picking the problems with $n \leq 12$ and those that could be modified to reduced their dimension below 12 without losing their meaning[4].

The resulting testing set consists of the 55 problems listed in Table 2 with their name, dimension n, number of free variables n_{fr}, fixed variables n_f, lower bounded variables n_l, upper bounded variables n_u and number of variables with both lower and upper bounds n_{lu}.

We consider two versions of this set of problems: the first, denoted as `Set-cont`, contains problems where variables are continuous (original problem formulation) and the second, `Set-mix`, which consists of problems with mixed-integer variables. `Set-mix` is built modifying the CUTEst problems imposing that some variables can only assume integer values. In particular, we imposed that all variables with even indexes are integers and rounded accordingly the corresponding bounds, i.e., $x_{2i}, l_{2i}, u_{2i} \in \mathbb{Z}$ for all i [5].

The Matlab interfaces provided in [Gould et al. 2015] were used to test solvers on CUTEst problems.

Name	n	n_{fr}	n_l	n_u	n_{lu}	n_f	Name	n	n_{fr}	n_l	n_u	n_{lu}	n_f
ALLINIT	4	1	1	1		1	KOEBHELB	4	1	2			
BDEXP	10		10				LINVERSE	9	4	5			
BIGGSB1	10	1			9		LOGROS	2		2			
CAMEL6	2				2		MAXLIKA	8				8	
CHARDIS0	10				10		MCCORMCK	10				10	
CHEBYQAD	4				4		MDHOLE	2	1	1			
CVXBQP1	10				10		NCVXBQP1	10				10	
EG1	3	1			2		NCVXBQP2	10				10	
EXPLIN	12				12		NCVXBQP3	10				10	
EXPLIN2	12				12		NONSCOMP	10				10	
EXPQUAD	12	6			6		OSLBQP	8		3		5	
HADAMALS	4				2	2	PALMER1A	6	4	2			
HARKERP2	10		10				PALMER2B	4	2	2			
HART6	6				6		PALMER3E	8	7	1			
HATFLDA	4		4				PALMER4A	6	4	1			
HATFLDB	4		3		1		PALMER4	4	1	3			
HATFLDC	9	1			8		PENTDI	5		5			
HIMMELP1	2				2		POWELLBC	6			6		
HS110	10				10		PROBPENL	10				10	
HS1	2	1	1				PSPDOC	4	3		1		
HS25	3				3		QUDLIN	12				12	
HS2	2	1	1				S368	8				8	
HS38	4				4		SIMBQP	2	1				1
HS3	2	1	1				SINEALI	4				4	
HS3MOD	2	1	1				SPECAN	9				9	
HS45	5				5		WEEDS	3		2		1	
HS4	2		2				YFIT	3	2	1			
HS5	2				2								

Table 2. The benchmark problem set.

[4]We excluded the MINSURFO, the TORSION*, JNLBRNG* and the OBSTCLA* family because they arise from the discretized problems and are therefore meaningless in small dimensions. Moreover, we arbitrarily chose only 5 problems within the PALMER* family.

[5]The only exceptions are problems HATFLDB, MAXLIKA, KOABHELB for which we considered variables 'icic', 'cccicici', 'cci', respectively, to obtain problems with meaningful bounds ('c'and 'i' stand for continuous and integer variables).

3.2 The BFO parameters self-tuning

The BFO algorithm depends on a set of algorithmic parameters whose value may influence its numerical performance. Ideally, one would like to set these parameters to those values which gives the best numerical performance of BFO and set them as the default ones. In practice, a default parameter configuration can be computed by approximating the parameters which gives the best performance of BFO on a set of test problems that is chosen to be representative enough to "ensure" good performance of the solver on further problems.

We focus on the 7 BFO parameters reported in Table 1 together with their description and type: 5 parameters are continuous c, one is integer i and one (`stype`) is a set of labels {BF, DF, none} which identifies the discrete tree search strategy, to which we associate the integer values $\{0, 1, 2\}$ in the range $[0, 2]$. Note that this parameter is not necessary if a testing problem has only continuous variables.

The aim of this section is to estimate the best BFO parameters configuration with respect to the benchmark problem set. To address this issue we consider two parameter optimization problems formulated as a bound-constrained black-box optimization problem of the form (1) where the variables are the BFO parameters, and we use the BFO itself to solve it.

The first formulation, first proposed in [Audet and Orban 2006] and used later in [Audet et al. 2010, 2014], is based on defining the number of objective function evaluations as measure of (negative) performance of the algorithm and use a derivative-free solver to minimize it. This technique is implemented as follows. Let $\mathcal{T}$ be the set of test problems described in Section 3.1 and let $p = (p_1, \ldots, p_7)$ be the BFO parameters listed in Table 1. We choose reasonable default values p_0 for the parameters and associated lower/upper bounds l_p and u_p. Then, we use BFO to solve the "Average Objective" (AO) problem

$$\min_{l_p \le p \le u_p} \ \phi_{BFO}(p) \tag{13}$$

where $\phi_{BFO}(p)$ counts the total number of evaluations of f to solve all the problems in $\mathcal{T}$ with parameters p. We note that this formulation falls in the class of problems where the objective function is the sum of positive terms.

The second formulation relies on robust optimization which provides a tool for protecting against strong local variation of performance by looking for a safe worst-case scenario [Conn and Vicente 2012]. We follow this approach by allowing perturbations of each continuous algorithmic parameter by at most 5% around each tested value and defining the local box

$$\mathcal{B} = \prod_{i=1}^{5} [0.95\, p_i, 1.05\, p_i] \times \prod_{i=6}^{7} [p_i, p_i],$$

and use BFO to solve the problem "Robust Objective" (RO) problem

$$\min_{l_p \le p \le u_p} \ \max_{\bar{p} \in \mathcal{B}} \ \phi_{BFO}(\bar{p}) \tag{14}$$

where ϕ_{BFO}, l_p, u_p are defined as above.

In the experiments, we set the starting parameter p_0 and the bounds l_p, u_p as given in Table 3 and the initial scaling $\delta_p = (u_p - l_p)/10$ for continuous parameters and $\delta_p = 1$ for the integer ones.

Moreover, we used the value $\epsilon = 10^{-13}$ in the convergence test (9) in the solution of a single problem in $\mathcal{T}$ required to evaluate ϕ_{BFO}. On the other hand, based on our experiments, we set $\epsilon = 10^{-2}$ in the same test when solving the outer minimization problem in both (13) and (14), and $\epsilon = 10^{-1}$ in the solution of the inner minimization problem (14). Finally, we set an upper bound of 100 parameter configuration trials.

	α	β	γ	δ	η	inertia	stype
p_0	2	0.5	5	1	10^{-3}	10	0
l_p	1	0.01	1	0.25	10^{-5}	5	0
u_p	2	0.95	10	10	0.5	30	2

Table 3. Parameter setting for the BFO self-tuning.

In Table 4 we give the estimated parameters p_{AO} and p_{RO} computed using the AO formulation (13) and the RO formulation (14), respectively. Values of p_{AO} and p_{RO} slightly differ and both suggest to use the depth-first strategy (stype = 1) in the RECURSIVE STEP. We also report the value of the seed of the random number generator (labeled as rseed) which we found experimentally suitable. Figure 1 shows the corresponding performance profiles which reveals that: using p_{RO} yields the most efficient version of BFO on Set-cont; the performance of BFO with p_{RO} and p_{AO} is comparable for $t \approx 1$ on Set-mix and both outperform BFO with p_0; the robustness of the three versions of BFO is comparable (see percentage values in brackets).

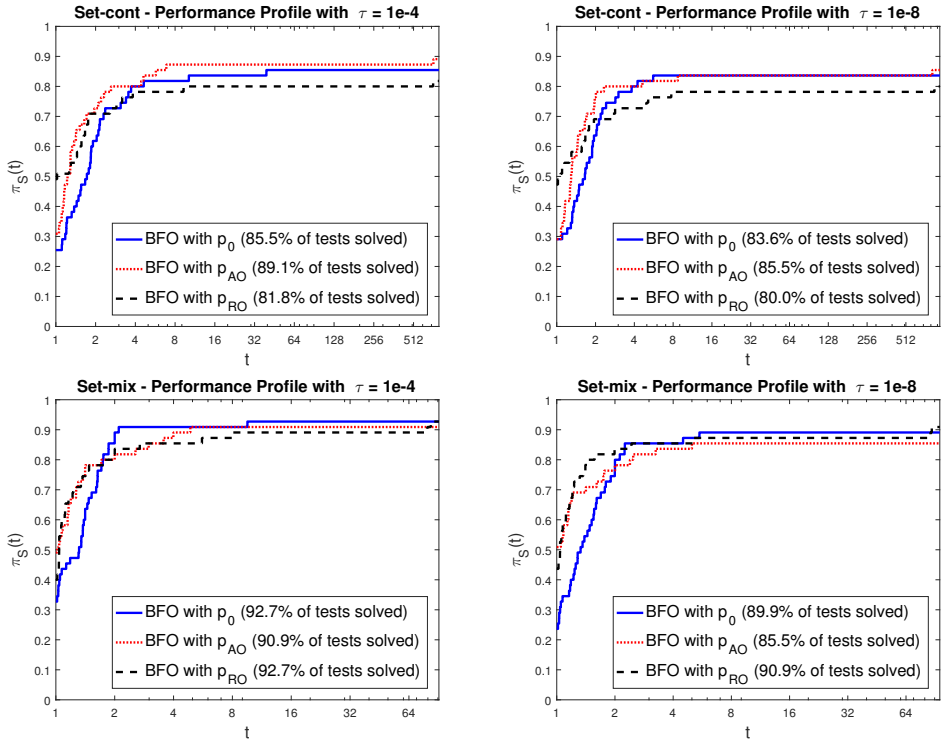

Fig. 1. Performance profiles of BFO with different algorithmic parameters on Set-cont (top) and on Set-mix (bottom). Cutoff $\tau = 10^{-4}$ (left) and $\tau = 10^{-8}$ (right).

We conclude this section by the (important in our view) observation that the same approach can be used to optimize performance of other numerical algorithms, using BFO to solve the associated AO or RO problems.

		α	β	γ	δ	η	inertia	stype	rseed
continuous	p_{AO}	1.6366	0.3426	4.4507	8.5744	0.1142	13	-	64
problems	p_{RO}	1.4248	0.1997	2.3599	1.0368	0.4528	11	-	53
mixed-integer	p_{AO}	2	0.0448	5	1	0.1190	10	1	91
problems	p_{RO}	2	0.3135	5	3.6030	10^{-5}	10	1	91

Table 4. Estimated optimal BFO parameters.

3.3 Comparison with NOMAD

As a competitor solver, we considered NOMAD (Release 3.6.2), a well-known solver for derivative-free mixed variable nonlinear optimization [Abramson et al. 2008; Le Digabel 2011]. NOMAD (Nonsmooth Optimization by Mesh Adaptive Direct Search) belongs to the class of direct-search methods and is based on the the recent development of Mesh Adaptive Direct Search methods [Abramson et al. 2009]. NOMAD is in fact an hybrid method that enhances the efficiency of MADS methods by combining direct-search strategies with different types of surrogate models in a mesh adaptive direct search filter method.

We now report on the numerical comparison between BFO and NOMAD evaluated in both "direct-search" mode and "model-based" mode (denoted as NOMAD-DS and NOMAD-MB, respectively). In our experiments we set $\mu_f = 10000$ and considered two levels of accuracy $\tau = 10^{-i}$, $i = 4, 8$. In both BFO and NOMAD, the internal stopping criterion is based on driving the mesh size below a tolerance ϵ that we set as $\epsilon = 10^{-13}$. All other NOMAD parameters have been set as the default ones [6].

In Figures 2 and 3, we plot the comparison between the three versions of BFO, i.e., with parameters p_0, p_{AO} and p_{RO} of Tables 3 and 4, and the two versions of NOMAD.

Figure 2 shows that BFO with p_{RO} is the most efficient in the 40% of the tests on `Set-cont` while in the solution of `Set-mix` BFO with p_{AO} is the most efficient for $\tau = 10^{-4}$; for $\tau = 10^{-8}$ the performance of the compared solvers is similar for $t \approx 1$. It is also clear in Figure 3, that the tuned versions of BFO are very competitive with NOMAD-MB on `Set-mix` since the plotted curves are very close for $t \geq 1.5$ while, unsurprisingly, the NOMAD model based approach is more efficient than BFO on the smooth continuous problems in `Set-cont`, especially for $\tau = 10^{-8}$.

Remarkably, BFO is on average more robust than NOMAD in that it solves around 10-15% more problems then the competitor (see the percentage values in brackets).

Finally, we report the corresponding data profiles in Figure 4. From these profiles, it is clear that when the computational budget is small, say 100 simplex gradient evaluations, the behaviour of the competing solvers is comparable. On the other hand, for both accuracy levels τ, as the computational budget increases, BFO solves a larger number of problems than NOMAD (both versions) and the difference increases with the number of simplex gradients ν.

4 BFO AS A TRAINABLE ALGORITHM

If the performance of the BFO algorithm can be optimized with itself, the obvious next step is to provide this facility within the code, allowing a user to specify a set of test problems (we used the CUTEst test problems above) and optimizing performance on this class. As a result, we obtain what we call a "trainable algorithm": given a set of test problems, a trainable algorithm can be trained for

[6] When NOMAD is tested in the "direct-search" mode, we disabled the options `MODEL_SEARCH` and `MODEL_EVAL_SORT` in order not to use model based strategies.

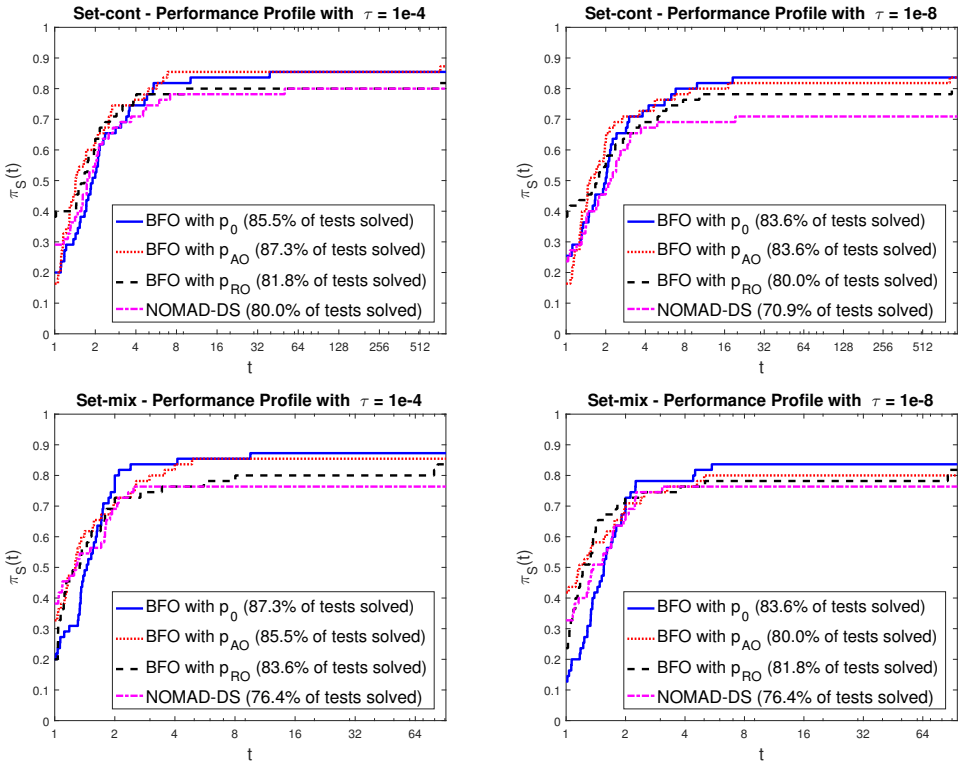

Fig. 2. Performance profiles of BFO against NOMAD-DS on `Set-cont` (top) and on `Set-mix` (bottom). Cutoff $\tau = 10^{-4}$ (left) and $\tau = 10^{-8}$ (right).

(hopefully) improved performance on this set and subsequently applied to further problems (of the same type) using the internal configuration (algorithmic parameters) resulting from this training.

In BFO, this facility is implemented by allowing the user to specify three possible "training modes". The first correspond to the training phase only and solves problem AO or RO on the user-supplied set of training problems. The second mode first perform this training and then immediately uses the resulting optimized algorithmic parameters to solve one or more new problems. The third mode first reads previously trained parameters from a file and then uses them for the solution of a new problem. Various options may be specified, allowing the user to choose between AO and RO, specifying the name of the file where trained parameters are saved and restricting the training to certain meaningful sets of parameters. We refer to the description of the BFO input parameters for more details. Note that BFO being a descent method implies that performance is improved at every training iteration, and therefore that accurate solution of the training problem AO or RO is generally unnecessary (and potentially leading to overfitting).

We now illustrate the potential benefits and pitfalls of training by considering three specific class of minimization problems.

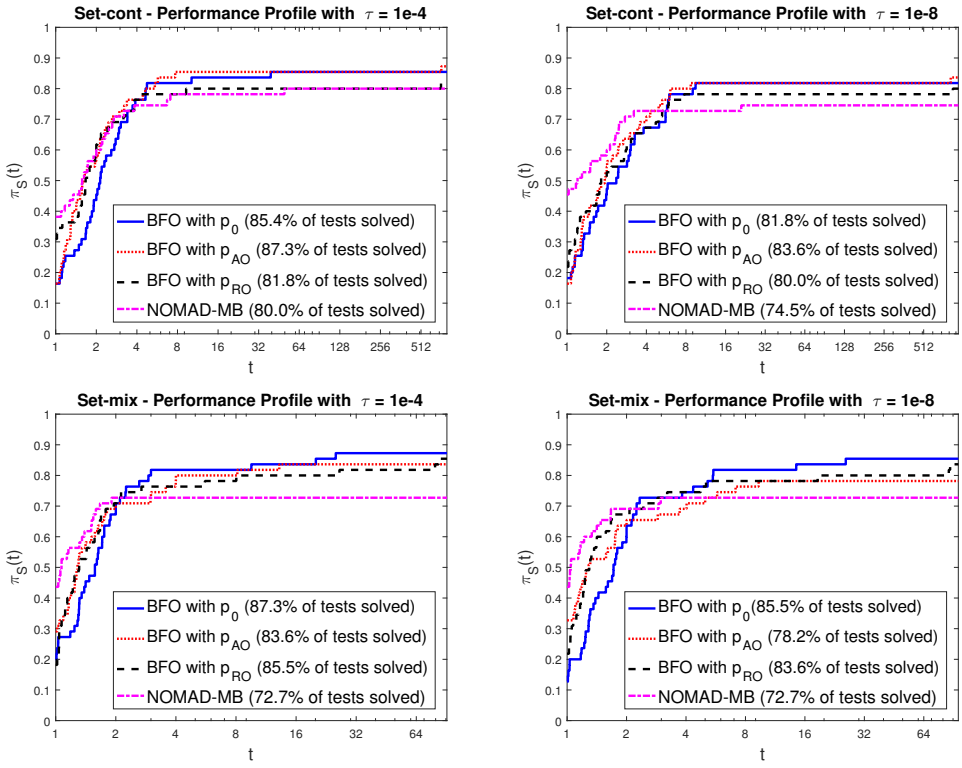

Fig. 3. Performance profiles of BFO against NOMAD-MB on `Set-cont` (top) and on `Set-mix` (bottom). Cutoff $\tau = 10^{-4}$ (left) and $\tau = 10^{-8}$ (right).

4.1 Nonlinear least-squares

The first is a class of nonlinear least-squares problems with bounds[7] where one seeks to fit a nonlinear model of a vibrating beam (hence our name of VBEAM for this problem class) to data by minimizing

$$f(x_1, x_2, x_3) = \sum_{j=0}^{16} \left[x_3 \tan\left(x_1 \left(1 - \frac{j}{16}\right) + x_2 \frac{j}{16} \right) - y_j \right]^2$$

where $x_3 \geq 0$. We fixed values for the three variables[8] and generated several classes of 10 problems, where

$$y_j = x_3 \tan\left(x_1 \left(1 - \frac{j}{16}\right) + x_2 \frac{j}{16} \right) (1 + \eta_j) \quad (j = 0, \ldots 16),$$

with $\eta_0 = 0$ and η_j $(j > 0)$ being a realisation of a Gaussian noise with zero mean and a prescribed value of the standard deviation σ, each class of test problems corresponding to a different value of σ. For each such class, we trained BFO using both average and robust training strategies on the 10 generated test problems, and then applied the trained BFO on 20 additional validation problems

[7] Derived from the YFIT problem in CUTEst and corresponding to fitting data to Doppler measures of a vibrating beam. The original problem was proposed by L. Watson (VPI).

[8] $x_1^* = 0.21$, $x_2^* = -0.35$ and $x_3^* = 1$, $x_3^* = 10$ or $x_3^* = 100$.

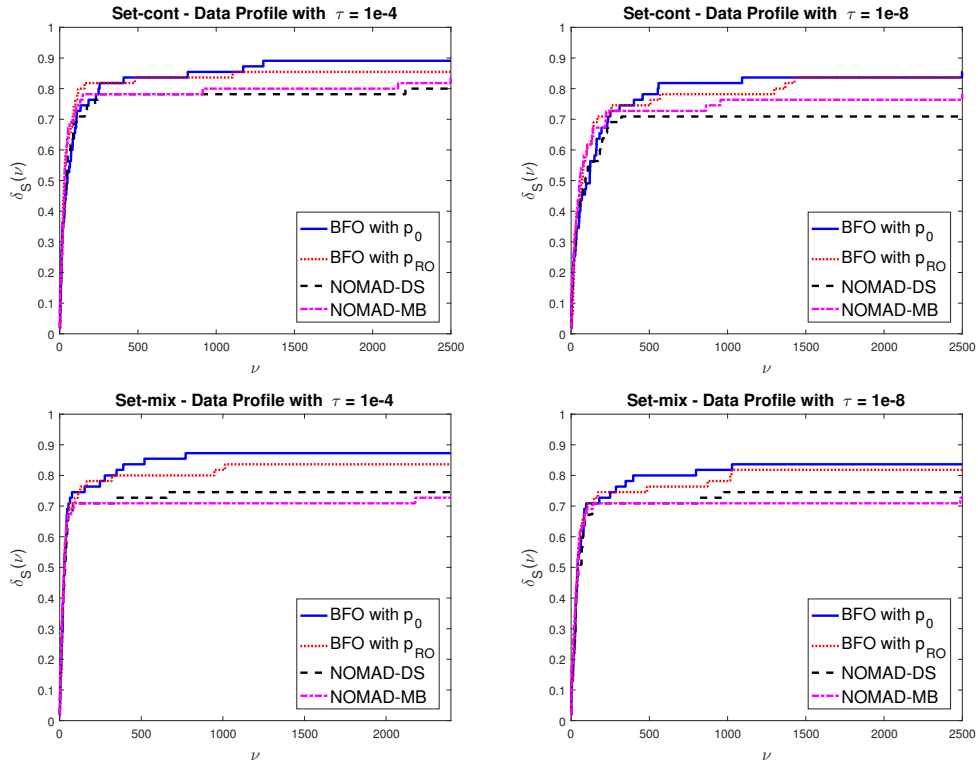

Fig. 4. Data profiles of BFO against NOMAD-DS and NOMAD-MB on `Set-cont` (top) and on `Set-mix` (bottom). Cutoff $\tau = 10^{-4}$ (left) and $\tau = 10^{-8}$ (right).

generated with the same parameters and standard deviation, in order to measure effectiveness of the training on this validation set.

We report in Tables 5 and 6 in appendix some results obtained in this setting. In these tables, ϵ_1 is the accuracy requirement of the outer minimization in the training problem formulation (see (13) and (14)), ϵ_2 is the accuracy requirement in the inner maximization of (14), σ is the standard deviation described in the preceding paragraph, "neval" is the total number of evaluation of problems in the testing set, "t-set" is the gain/loss (in percentage) in the number of problem function evaluations achieved by the training phase on the 10 training problems of each class and "v-set" is the gain/loss of problem function evaluations obtained on the validation set consisting of the 20 additional problems of each class which were not included in the training set.

We now attempt some tentative conclusions on this specific class of problems:

(1) The gains in performance obtained by training using either strategy, although clearly not guaranteed, may be substantial, especially for narrowly defined problems classes (small values of σ). Reported values approach 60% in some cases, but further experimentation not reported here indicate d that even higher gains may sometimes be possible. On average, gains appear to be of the order of 25%.

(2) Asking more accuracy for the outer minimization in training ($\epsilon_1 = 0.001$) is typically[9] more costly.
(3) Unsurprisingly, a very moderate accuracy requirement for this outer minimization often seems perfectly sufficient. Higher accuracy levels may increase the cost of training without any guarantee of improvement in performance on the validation set. This observation appears to hold for both training strategies, with the possible exception of the worse conditioned problems ($x_3 = 100$) with larger standard deviation ($\sigma \geq 0.5$) when the RO strategy is used.
(4) Again unsurprisingly, the RO training strategy is nearly always substantially more costly that the AO one (by a factor often exceeding one order of magnitude). It may however produce benefits in terms of training capacity for problems with relatively high values of σ, that is problems where deviations within the class are proportionally larger.
(5) Increasing the accuracy in the inner maximization (ϵ_2) in the RO strategy again appears to bring few benefits, with the same marginal exception of the worse conditioned problems with larger standard deviation.

4.2 Regularized cubics

The second class of problems (RCUBIC) on which training was experimented is a class of unconstrained problems featuring an regularized cubic objective function of the form

$$\begin{aligned} f(x_1, x_2, x_3, x_4) &= -\mu[x_1(1+\nu_1) + x_2(1+\nu_2) + x_3(1+\nu_3) + x_4(1+\nu_4)] \\ &\quad -x_1^2 - 2x_2^2 - 3x_3^2 - 4x_4^2 + 10(1+\nu_5)[x_1^2 + x_2^2 + x_3^2 + x_4^2]^{3/2} \end{aligned}$$

where μ is a parameter in $\{0.1, 1, 10\}$ and $\{\nu_i\}_{i=1}^5$ are realizations of a Gaussian random noise with zero mean and prescribed values of the standard deviation σ. As for the VBEAM class, 10 training problems were generated for each value of μ and σ. BFO was then trained on this set for different accuracy levels using both the AO and RO strategies and the resulting algorithmic parameters were then used to solve 20 validation problems generated with the same σ. The results of these experiments are reported in Tables 7 and 8 in appendix.

These tables suggest the following comments.

(1) The typical gains in performance are smaller on this problem class than those obtained for the VBEAM class, but they remain non-negligible (from 10 to 30% except for the case $\sigma = 1$).
(2) For both strategies, the evolution of the gains with increasing value of the noise standard deviation σ is somewhat unpredictable, the best results being often obtained for intermediate values of this problem parameter. However, the largest standard deviation typically leads to worse performance.
(3) As for the VBEAM class, moderate training accuracy seems most often sufficient to extract good performance, both for the AO and RO strategies.
(4) Again the RO strategy is considerably more costly, in this case for results comparable on the whole to those obtained for AO.

Summarizing, the experiments reported in this section indicate that training has a non-negligible cost but may yield significant efficiency gains when the class of problems considered is sufficiently well-defined. Moreover, it appears that the RO training strategy, albeit clearly designed for increased

[9]The nonconvexity of the training criteria nevertheless implies that different approximate minimizers can be found for successive training runs with different accuracy requirements. Similarly, the geometry of problems may vary with σ. As a consequence monotonicity of cost and/or performance with respect to increased accuracy requirements or decreasing noise is not always preserved.

robustness, does not (in these two experiments at least) deliver very convincingly on this ground in view of its very significantly higher training cost. However it definitely should be remembered that gains are not guaranteed, and therefore that the encouraging but tentative conclusion (especially for the AO strategy) must be taken with due caution. If the user wishes to solve a large number of problems within a well-defined class of interest, some experimentation with training is probably advisable.

5 TACKLING MULTILEVEL PROBLEMS

We now turn to the description of how BFO has been adapted to solve problems of the form (2), provided each optimization subproblem (at different levels) is well-defined. The fact that forward and backward steps are used by the algorithm makes it easy to restrict minimization (or maximization) to specific subspace. In particular, variable selection is trivial as it is sufficient to fix one or more variables to define a proper subspace. Indeed, assume for simplicity that there are two levels involving vectors of variables x_1 and x_2, respectively. In order to evaluate the objective function for the outer (level 1) optimization problem, BFO recursively calls itself to optimize the objective on x_2 while keeping the variables in x_1 fixed.

Which variable is assigned to which level is specified by the user using an optional argument (it is then required that every variable is assigned a level, and that every level is assigned at least one variable). Another argument allows the specification of the choice of minimization or maximization at each level.

As we indicated in the introduction, the variables at each level may be constrained by their own set of bounds, and these bounds may themselves depend on the value of the variables at levels of lower index. This is achieved by calling a user-supplied function defining these "variable bounds" in a reasonably flexible format. This feature has the marginal effect that it allows considering constrained problems of the form

$$\min_x f(x_1, x_2) \quad \text{subject to} \quad g(x_1) \leq x_2 \leq h(x_1)$$

by reformulating them as two levels problem

$$\min_{x_1} \min_{x_2} f(x_1, x_2)$$

where the "variable bounds" on x_2 are defined by $g(x_1) \leq x_2 \leq h(x_1)$. Since the "variable bounds" definition may itself call BFO, it is also possible to tackle problems of the form

$$\min_{x_1 \in \mathbb{R}^n} f(x_1, x_2) \quad \text{subject to} \quad \min_{x_1 \in \mathbb{R}^n} g(x_1) \leq x_2 \leq \min_{x_1 \in \mathbb{R}^n} h(x_1)$$

where x_1 and/or $x_2 \in \mathbb{R}$ may contain a mix of discrete or continuous variables and where additional (fixed) bounds may be imposed on x_1 and x_2.

The multilevel facility coupled with the definition of the variable bounds technique just illustrated therefore shows considerable flexibility, but it must be kept in mind that recursive optimization over two or more levels may be expensive in terms of objective function evaluations, even after training on a specific class of multilevel problems.

6 CONCLUSION

We have presented BFO, a versatile and robust Brute Force Optimizer for small-scale bound-constrained problems based on a simple direct-search strategy, whose distinguishing features are its ability to handle a mix of continuous and discrete variables and its innovative self-training capacity. The fact that it can also handle multilevel problems, although possibly at higher cost, is

also a plus. BFO is written in Matlab. On CUTEst problems, its performance compares favourably with that of NOMAD, a well-known direct-search package, most notably in terms of reliability.

As its name suggests and despite a favourable but limited comparison with NOMAD, BFO has no pretense of superior efficiency, especially for multilevel problems. It is hoped that it will nevertheless turn out to be useful because of its versatility, trainable nature and robustness. In particular, its application for optimizing algorithmic parameters in various numerical methods, in optimization and beyond, is of definite interest.

ACKNOWLEDGMENTS

The work of the first author was supported by *National Group of Computing Science (GNCS-INdAM)* of Italy. This author wishes to thank Francesco Rinaldi for his support and for valuable discussions on derivative-free methods. Part of the research was conducted during a visit of the second author to the Università degli Studi di Firenze, which was partially supported by GNCS-INdAM. Both authors also thank Dominique Orban for helpful discussions, as well as three referees for their careful reading and their suggestions, which lead to significant improvement of the manuscript.

REFERENCES

M. A. Abramson, C. Audet, J. W. Chrissis, and J. G. Walston. 2009. Mesh adaptive direct search algorithms for mixed variable optimization. *Optimization Letters* 3, 1 (2009), 35–47. https://doi.org/10.1007/s11590-008-0089-2

M. A. Abramson, C. Audet, G. Couture, J.E. Jr. Dennis, and S. Le Digabel. 2008. The NOMAD project. (2008). http://www.gerad.ca/nomad

C. Audet, C.-K. Dang, and D. Orban. 2010. Algorithmic Parameter Optimization of the DFO Method with the OPAL Framework. In *Software Automatic Tuning*, K. Naono, K. Teranishi, J. Cavazos, and R. Suda (Eds.). Springer New York, 255–274. https://doi.org/10.1007/978-1-4419-6935-4_15

C. Audet, K.-C. Dang, and D. Orban. 2014. Optimization of algorithms with OPAL. *Mathematical Programming Computation* (2014), 1–22. https://doi.org/10.1007/s12532-014-0067-x

C. Audet and J. Dennis. 2001. Pattern Search Algorithms for Mixed Variable Programming. *SIAM Journal on Optimization* 11, 3 (2001), 573–594. https://doi.org/10.1137/S1052623499352024

C. Audet and D. Orban. 2006. Finding Optimal Algorithmic Parameters Using Derivative-Free Optimization. *SIAM Journal on Optimization* 17, 3 (2006), 642–664. https://doi.org/10.1137/040620886

J. F. Bard. 1998. *Practical bilevel optimization: algorithms and applications.* Vol. 30. Springer Science & Business Media.

G. E. P. Box and K. B. Wilson. 1951. On the Experimental Attainment of Optimum Conditions. *Journal of the Royal Statistical Society. Series B (Methodological)* 13, 1 (1951), pp. 1–45.

Ch. G. Broyden. 1970. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics* 6, 1 (1970), 76–90. https://doi.org/10.1093/imamat/6.1.76

B. Colson, P. Marcotte, and G. Savard. 2007. An overview of bilevel optimization. *Annals of operations research* 153, 1 (2007), 235–256. https://doi.org/10.1007/s10479-007-0176-2

A.R. Conn, K. Scheinberg, and L.N. Vicente. 2009. *Introduction to Derivative-Free Optimization.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

A. R. Conn, K. Scheinberg, and Ph. L. Toint. 1998. A derivative free optimization algorithm in practice. In *Proceedings of 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, St. Louis, MO*, Vol. 48.

A. R. Conn and L. N. Vicente. 2012. Bilevel derivative-free optimization and its application to robust optimization. *Optimization Methods and Software* 27, 3 (2012), 561–577. https://doi.org/10.1080/10556788.2010.547579

I. D. Coope and Ch. J. Price. 2001. On the convergence of grid-based methods for unconstrained optimization. *SIAM Journal on Optimization* 11, 4 (2001), 859–869. https://doi.org/10.1137/S1052623499354989

A.L. Custódio, H. Rocha, and L.N. Vicente. 2010. Incorporating minimum Frobenius norm models in direct search. *Computational Optimization and Applications* 46, 2 (2010), 265–278. https://doi.org/10.1007/s10589-009-9283-0

A. Custódio and L. Vicente. 2007. Using Sampling and Simplex Derivatives in Pattern Search Methods. *SIAM Journal on Optimization* 18, 2 (2007), 537–555. https://doi.org/10.1137/050646706

S. Dempe. 2002. *Foundations of bilevel programming.* Springer Science & Business Media.

E.D. Dolan and J.J. Moré. 2002. Benchmarking optimization software with performance profiles. *Mathematical Programming* 91 (2002), 201–213. https://doi.org/10.1007/s101070100263

R. Fletcher. 1970. A new approach to variable metric algorithms. *Comput. J.* 13, 3 (1970), 317–322. https://doi.org/10.1093/comjnl/13.3.317

D. Goldfarb. 1970. A family of variable-metric methods derived by variational means. *Mathematics of computation* 24, 109 (1970), 23–26. https://doi.org/10.2307/2004873

N.I.M. Gould, D. Orban, and Ph.L. Toint. 2015. CUTEst: a Constrained and Unconstrained Testing Environment with safe threads for mathematical optimization. *Computational Optimization and Applications* 60, 3 (2015), 545–557. https://doi.org/10.1007/s10589-014-9687-3

S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang. 2015. Direct search based on probabilistic descent. *SIAM Journal on Optimization* 25, 3 (2015), 1515–1541. https://doi.org/10.1137/140961602

G. A. Gray, J.D. Griffin, M. Taddy, M. Martinez-Canales, and T.G. Kolda. 2008. *HOPSPACK: Hybrid Optimization Parallel Search Package.* Technical Report SAND2008-8057. Sandia National Laboratories (SNL-CA), Livermore, CA (United States). https://doi.org/10.2172/1130394

G. A. Gray, J.D. Griffin, M. Taddy, M. Martinez-Canales, and T.G. Kolda. 2010. HOPSPACK: Hybrid Optimization Parallel Search Package. (2010). https://software.sandia.gov/trac/hopspack

L. Grippo and F. Rinaldi. 2015. A class of derivative-free nonmonotone optimization algorithms employing coordinate rotations and gradient approximations. *Computational Optimization and Applications* 60, 1 (2015), 1–33. https://doi.org/10.1007/s10589-014-9665-9

Zeynep H Gümüş and Christodoulos A Floudas. 2005. Global optimization of mixed-integer bilevel programming problems. *Computational Management Science* 2, 3 (2005), 181–212. https://doi.org/10.1007/s10287-005-0025-1

R. Hooke and T.A. Jeeves. 1961. "Direct Search" Solution of Numerical and Statistical Problems. *Journal of the ACM (JACM)* 8, 2 (1961), 212–229. https://doi.org/10.1145/321062.321069

S. Le Digabel. 2011. Algorithm 909: NOMAD: Nonlinear Optimization with the MADS Algorithm. *ACM Trans. Math. Software* 37, 4, Article 44 (Feb. 2011), 44:1–44:15 pages. https://doi.org/10.1145/1916461.1916468

M. Leonetti, P. Kormushev, and S. Sagratella. 2012. Combining local and global direct derivative-free optimization for reinforcement learning. *Cybernetics and Information Technologies* 12, 3 (2012), 53–65. https://doi.org/10.2478/cait-2012-0021

G. Liuzzi, S. Lucidi, and F. Rinaldi. 2012. Derivative-free methods for bound constrained mixed-integer optimization. *Computational Optimization and Applications* 53, 2 (2012), 505–526. https://doi.org/10.1007/s10589-011-9405-3

G. Liuzzi, S. Lucidi, and F. Rinaldi. 2014. Derivative-Free Methods for Mixed-Integer Constrained Optimization Problems. *Journal of Optimization Theory and Applications* (2014), 1–33. https://doi.org/10.1007/s10957-014-0617-4

S. Lucidi, V. Piccialli, and M. Sciandrone. 2005. An Algorithm Model for Mixed Variable Programming. *SIAM Journal on Optimization* 15, 4 (2005), 1057–1084. https://doi.org/10.1137/S1052623403429573

J. Moré and S. Wild. 2009. Benchmarking Derivative-Free Optimization Algorithms. *SIAM Journal on Optimization* 20, 1 (2009), 172–191. https://doi.org/10.1137/080724083

J. A Nelder and R. Mead. 1965. A simplex method for function minimization. *Comput. J.* 7, 4 (1965), 308–313. https://doi.org/10.1093/comjnl/7.4.308

R. Oeuvray and M. Bierlaire. 2007. A new derivative-free algorithm for the medical image registration problem. *International Journal of Modelling and Simulation* 27, 2 (2007), 115. https://doi.org/10.1080/02286203.2007.11442407

D. F. Shanno. 1970. Conditioning of quasi-Newton methods for function minimization. *Mathematics of computation* 24, 111 (1970), 647–656. https://doi.org/10.2307/2004840

Ch. Siefert, V. Torczon, and M. W. Trosset. 1997. Model-Assisted Pattern Search Library. (1997). http://www.cs.wm.edu/~va/software/maps/

Y. Tang, J.-Ph. P. Richard, and J. C. Smith. 2015. A class of algorithms for mixed-integer bilevel min–max optimization. *Journal of Global Optimization* (2015), 1–38. https://doi.org/10.1007/s10898-015-0274-7

L. Vicente, G. Savard, and J. Judice. 1996. Discrete linear bilevel programming problem. *Journal of optimization theory and applications* 89, 3 (1996), 597–614. https://doi.org/10.1007/BF02275351

A RESULTS FOR THE TRAINING ON THE VBEAM PROBLEMS

		$\epsilon_1 = 0.1$			$\epsilon_1 = 0.01$			$\epsilon_1 = 0.001$		
x_3^*	σ	neval	t-set	v-set	neval	t-set	v-set	neval	t-set	v-set
1.0	0.05	78016	57%	58%	84676	57%	58%	91376	57%	58%
	0.10	74664	57%	57%	92144	57%	57%	99099	57%	57%
	0.50	76770	55%	51%	90714	55%	51%	97697	55%	51%
	1.00	118002	25%	18%	127859	25%	18%	137598	25%	18%
10.0	0.05	115103	25%	24%	117627	30%	25%	153660	31%	26%
	0.10	91736	30%	27%	101330	30%	27%	116878	31%	27%
	0.50	140689	29%	29%	150873	29%	29%	161938	29%	29%
	1.00	371770	74%	-3%	391895	74%	-3%	469877	75%	-3%
100.0	0.05	1528163	40%	17%	2259418	46%	35%	2409211	46%	35%
	0.10	949534	74%	-63%	990102	74%	-63%	1028300	74%	-63%
	0.50	1058333	44%	37%	1311953	44%	20%	1420676	44%	20%
	1.00	2246316	53%	44%	2282538	42%	28%	2428030	42%	28%

Table 5. Training performance for the VBEAM problems (average strategy).

B RESULTS FOR THE TRAINING ON THE RCUBIC PROBLEMS

			$\epsilon_2 = 0.1$			$\epsilon_2 = 0.01$			$\epsilon_2 = 0.001$		
x_3^*	σ	ϵ_1	neval	t-set	v-set	neval	t-set	v-set	neval	t-set	v-set
1.0	0.05	0.100	628646	34%	32%	1785918	26%	13%	1524726	24%	19%
		0.010	696865	29%	25%	2228106	32%	14%	1861051	24%	19%
		0.001	801217	30%	22%	2388132	31%	21%	2526828	27%	22%
	0.10	0.100	613487	39%	34%	1206625	26%	22%	2082234	30%	35%
		0.010	594036	28%	14%	1602390	30%	29%	1786235	34%	32%
		0.001	832147	28%	14%	2227721	33%	29%	1814774	34%	36%
	0.50	0.100	752798	34%	18%	829479	26%	0%	1291995	25%	11%
		0.010	826141	26%	15%	1569947	32%	15%	1915998	27%	9%
		0.001	893174	27%	3%	1576704	32%	15%	2209387	28%	10%
	1.00	0.100	915320	35%	23%	1428637	34%	26%	1903385	31%	15%
		0.010	1007303	35%	27%	1316097	32%	21%	3275141	33%	16%
		0.001	1057979	36%	28%	1415476	33%	23%	3454439	33%	15%
10.0	0.05	0.100	550136	31%	26%	2231802	29%	25%	2154878	25%	19%
		0.010	742751	33%	27%	2384730	31%	25%	3565739	30%	26%
		0.001	711696	33%	27%	2384978	31%	25%	3425180	29%	29%
	0.10	0.100	356406	34%	27%	1825640	33%	22%	483576	34%	27%
		0.010	415057	34%	27%	1482960	29%	16%	542227	34%	27%
		0.001	437556	34%	27%	1576560	29%	16%	564726	34%	27%
	0.50	0.100	581248	49%	26%	960202	64%	14%	1568987	71%	25%
		0.010	630377	49%	26%	971803	64%	14%	1648454	71%	25%
		0.001	718977	49%	26%	1015169	64%	14%	1828169	71%	25%
	1.00	0.100	1875988	72%	-5%	2012749	69%	0%	4731855	64%	5%
		0.010	2495629	70%	7%	2416306	69%	0%	5870814	65%	3%
		0.001	2687561	70%	7%	2391194	69%	0%	6639303	64%	6%
100.0	0.05	0.100	22800147	49%	53%	23756246	46%	51%	46830361	52%	57%
		0.010	13796099	49%	50%	54424183	53%	55%	41559646	45%	54%
		0.001	16016353	49%	50%	54424183	53%	55%	46977030	45%	54%
	0.10	0.100	12074874	51%	18%	24934693	53%	24%	28315648	50%	20%
		0.010	18603805	51%	20%	26371545	53%	24%	22841398	43%	21%
		0.001	17117883	46%	21%	27787326	53%	24%	27516073	43%	21%
	0.50	0.100	5825325	41%	32%	14433847	34%	44%	15522908	42%	44%
		0.010	5929728	41%	32%	17728327	34%	44%	18043836	42%	44%
		0.001	5943597	41%	32%	19799363	35%	47%	20980871	42%	44%
	1.00	0.100	15984178	39%	43%	50144474	46%	46%	49081955	54%	23%
		0.010	24300473	43%	43%	69466970	49%	13%	57829751	56%	56%
		0.001	26649456	43%	43%	72037240	49%	13%	61968367	56%	56%

Table 6. Training performance for the VBEAM problems (robust strategy).

		$\epsilon_1 = 0.1$			$\epsilon_1 = 0.01$			$\epsilon_1 = 0.001$		
μ	σ	neval	t-set	v-set	neval	t-set	v-set	neval	t-set	v-set
0.1	0.05	171390	21%	18%	215379	21%	18%	236948	21%	18%
	0.10	147549	27%	32%	163969	27%	32%	179003	27%	32%
	0.50	210625	27%	22%	231710	28%	21%	353271	30%	21%
	1.00	1316810	2%	2%	1669344	2%	1%	2420867	2%	1%
1.0	0.05	157465	29%	19%	251780	31%	24%	364525	35%	22%
	0.10	145578	27%	21%	164747	27%	21%	209898	27%	18%
	0.50	167160	33%	20%	213654	33%	20%	275630	34%	19%
	1.00	884294	7%	4%	1117489	7%	4%	1325105	8%	5%
10.0	0.05	170696	24%	17%	190527	24%	17%	210716	24%	17%
	0.10	250660	27%	21%	197321	21%	17%	217332	21%	17%
	0.50	252651	17%	19%	292688	17%	19%	313567	17%	19%
	1.00	365754	9%	6%	577169	11%	7%	621727	11%	7%

Table 7. Training performance for the RCUBIC problems (average strategy).

			$\epsilon_2 = 0.1$			$\epsilon_2 = 0.01$			$\epsilon_2 = 0.001$		
μ	σ	ϵ_1	neval	t-set	v-set	neval	t-set	v-set	neval	t-set	v-set
0.1	0.05	0.100	904337	25%	25%	2666451	20%	18%	1780697	17%	20%
		0.010	955052	25%	25%	2816439	20%	18%	1865210	17%	19%
		0.001	1051446	26%	27%	2969297	20%	18%	2656816	17%	22%
	0.10	0.100	408303	25%	28%	1899920	24%	34%	1975824	22%	27%
		0.010	426958	25%	28%	1910765	22%	27%	3459456	24%	33%
		0.001	486780	25%	28%	2770792	25%	25%	2713023	23%	28%
	0.50	0.100	929811	22%	16%	1038464	22%	20%	1262599	24%	21%
		0.010	776602	22%	20%	1098088	22%	20%	1594679	23%	20%
		0.001	775473	22%	20%	1774483	24%	23%	1740226	23%	20%
	1.00	0.100	5435143	2%	1%	22978138	1%	2%	42813655	1%	1%
		0.010	6475795	2%	1%	25992859	1%	2%	31156577	1%	-1%
		0.001	9047623	2%	1%	43715316	1%	2%	34462404	1%	0%
1.0	0.05	0.100	935440	19%	14%	1395242	23%	19%	1613559	19%	21%
		0.010	1306734	20%	18%	1532900	21%	16%	3032890	22%	20%
		0.001	1382806	20%	18%	1631930	21%	16%	2916307	22%	16%
	0.10	0.100	1036208	20%	21%	723048	20%	20%	2423883	19%	19%
		0.010	1262466	22%	26%	1257674	20%	23%	3574962	21%	20%
		0.001	1485215	23%	21%	1420361	21%	17%	4467090	22%	22%
	0.50	0.100	878268	23%	20%	980388	23%	20%	4097011	22%	20%
		0.010	990193	27%	25%	1141757	26%	21%	2641673	21%	16%
		0.001	1180931	27%	21%	1269361	26%	21%	2810517	21%	16%
	1.00	0.100	4589589	4%	1%	6923078	4%	1%	11537196	4%	2%
		0.010	5064799	4%	1%	8512039	4%	1%	24376570	4%	4%
		0.001	6214429	4%	2%	8925949	4%	1%	21217312	4%	2%
10.0	0.05	0.100	797424	26%	12%	2328438	24%	17%	1578391	25%	17%
		0.010	818112	26%	12%	2358203	27%	20%	1794730	25%	17%
		0.001	951865	26%	12%	2498016	27%	20%	1992326	25%	17%
	0.10	0.100	701318	20%	11%	2566099	20%	18%	2978151	24%	21%
		0.010	1065867	18%	15%	3025223	20%	18%	3711438	23%	18%
		0.001	1183465	18%	15%	3166889	21%	17%	4214931	23%	14%
	0.50	0.100	1383221	19%	21%	1529375	18%	17%	1671648	18%	19%
		0.010	1355317	19%	16%	2017567	19%	23%	2307945	18%	20%
		0.001	1481907	19%	16%	2136717	19%	23%	4720503	19%	20%
	1.00	0.100	2161854	5%	5%	7751704	6%	6%	4880611	7%	4%
		0.010	2254753	5%	5%	10487409	7%	4%	9767706	8%	6%
		0.001	2234031	5%	5%	10908540	7%	4%	10192074	8%	6%

Table 8. Training performance for the RCUBIC problems (robust strategy).