Multiple-Gradient Descent Algorithm (*MGDA*) for Pareto-Front Identification

Jean-Antoine Désidéri

Abstract This article compounds and extends several publications in which a *Multiple-Gradient Descent Algorithm (MGDA)*, has been proposed and tested for the treatment of multi-objective differentiable optimization. Originally introduced in [8], the method has been tested and reformulated in [9]. Its efficacy to identify the Pareto front [4] has been demonstrated in [22], in comparison with an evolutionary strategy. Recently, a variant, *MGDA-II*, has been proposed in which the descent direction is calculated by a direct procedure [10] based on a Gram-Schmidt orthogonalization process (*GSP*) with special normalization. This algorithm was tested in the context of a simulation by domain partitioning, as a technique to match the different interface components concurrently [11]. The experimentation revealed the importance of scaling, and a slightly modified normalization procedure was proposed ("*MGDA-IIb*"). Two novel variants have been proposed since. The first, *MGDA-III*, realizes two enhancements. Firstly, the *GSP* is conducted incompletely whenever a test reveals that the current estimate of the direction of search is adequate also w.r.t. the gradients not yet taken into account; this improvement simplifies the identification of the search direction when the gradients point roughly in the same direction, and makes the directional derivative common to several objective-functions larger. Secondly, the order in which the different gradients are considered in the *GSP* is defined in a unique way devised to favor an incomplete *GSP*. In the second variant, *MGDA-IV*, the question of scaling is addressed when the Hessians are known. A variant is also proposed in which the Hessians are estimated by the Broyden-Fletcher-Goldfarb-Shanno (*BFGS*) formula. Lastly, a solution is proposed to adjust the step-size optimally in the descent step.

Key words: multi-objective optimization, descent direction, convex hull, Gram-Schmidt orthogonalization process, *BFGS* quasi-Newton method

---

Jean-Antoine Désidéri
Institut National de Recherche en Informatique et en Automatique (INRIA), Centre de Sophia Antipolis - Méditerranée, BP 93, 2004 Route des Lucioles, F-06902 Sophia Antipolis cedex (France), e-mail: Jean-Antoine.Desideri@inria.fr

1 Introduction

Multi-objective optimization, particularly when constrained by the solution of a partial-differential equation (PDE), is an essential methodological element of Multi-Disciplinary Optimization (MDO) over which a large community has been focusing attention (see e.g. [20]-[21]-[18] for extensive reviews, and [13] for a short introduction). Modern (finite-volume/finite-element-type) PDE-simulation tools, by discrete or continuous adjoint approaches, more systematically provide functional gradients as well as the mere evaluation of the performance, and this reinforces the value of differentiable-optimization algorithms. However, in multi-criterion design optimization, evolutionary strategies that are simple to apply and undeniably very robust, are still the most commonly-used methods to identify Pareto fronts (e.g. [1] and [6]) although numerous alternatives have been proposed in the literature, in particular:

- the normal boundary intersection [7] aiming to produce evenly-distributed points on the Pareto set, and related weights;
- the normalized normal constraint method [19], which incorporates an additional filter for a more proper identification;
- the Pareto-front interpolation [16], in which the authors construct a sub-complex of a Delaunay triangulation of a finite set of Pareto optimal outcomes, and and devise special rules for checking the inherent non-dominance of complexes; the method, was further developed in various publications, e.g. [17], and is supported by a surrogate model to alleviate the high computational cost of function(al) evaluations.

Here, we consider the simultaneous minimization or reduction of n objective-functions, $\{J_i(\mathbf{y})\}$ $(i = 1,\ldots,n)$, assumed to be smooth (say $\mathscr{C}^2$) functions of the design-vector $\mathbf{y} = (y_1, y_2, \ldots, y_N) \in \mathbb{R}^N$. In this new publication, the restriction $n \leq N$, previously made, is abandoned.

Our analysis is developed to identify an appropriate direction of search ω to update the design vector from a given initial design-point $\mathbf{y}^0$, center of an open ball $\mathscr{B}$ in which the objective-functions are well-defined, smooth and convex:

$$\mathbf{y}^1 = \mathbf{y}^0 - \rho\,\omega \qquad (\rho > 0,\ \text{step-size}). \tag{1}$$

For the above iteration to be a descent step, two conditions should be met. Firstly, the directional derivatives of the objective-functions should all be strictly-positive:

$$\forall i = 1,\ldots,n:\ \left(\nabla J_i(\mathbf{y}^0), \omega\right) > 0\,. \tag{2}$$

Then, $-\omega$ is a descent direction common to all objective-functions. Secondly, the step-size ρ should be adjusted appropriately. The important question of step-size adjustment is approached in Subsection 3.3 when additionally Hessians are known; presently, we focus on the first condition, (2).

In [8] and [9], we have introduced the notion of "Pareto-stationarity": the design-point $\mathbf{y}^0$ is said to be Pareto-stationary if there exists a convex combination of the gradients, $\nabla J_i(\mathbf{y}^0)$, equal to 0:

$$\exists \alpha = \{\alpha_i\}\ (i = 1,\ldots,n) \text{ such that : } \alpha_i \geq 0\ (\forall i);\ \sum_{i=1}^{n} \alpha_i = 1;\ \sum_{i=1}^{n} \alpha_i \nabla J_i(\mathbf{y}^0) = 0\,. \quad (3)$$

We have shown that Pareto-stationarity is a necessary condition to Pareto-optimality. Originally in [8]-[9], this result was established under the assumption $n \leq N$; however, the result has been recently extended to arbitrary dimensions n and N, using a different, more rigorous argument and assuming convexity (see [8], version 3). Thus, hereafter, we examine the case where the initial design-point $\mathbf{y}^0$ is not Pareto-optimal or Pareto-stationary.

Remark 1. Following classical publications [3] [5], Fliege and Svaiter [14] have been using the notion of Pareto critical points characterized as follows:

$$range(A) \cap (-\mathbb{R}++)^N = \emptyset \quad (4)$$

where, in their notations, A is the Jacobian matrix,

$$A = \begin{bmatrix} \frac{\partial J_1}{\partial y_1} & \cdots & \frac{\partial J_1}{\partial y_N} \\ \vdots & & \vdots \\ \frac{\partial J_n}{\partial y_1} & \cdots & \frac{\partial J_n}{\partial y_N} \end{bmatrix} \quad (5)$$

$\mathbb{R}++$ denotes the set of strictly-positive numbers, and the power a Cartesian product. This condition excludes the existence of a direction along which the directional derivatives of all the objective functions are strictly positive. The Pareto-stationarity condition (3) is therefore equivalent to it, but expressed differently, in our view, more simply. From this definition, in [14] they have introduced a variational formulation that define Pareto critical points as solutions of the following $\min - \max$ problem:

$$\min_{v} f_{\mathbf{y}}(v) + \tfrac{1}{2}||v||^2 \quad (6)$$

where $f_{\mathbf{y}}(v) = max\big((Av)_i,\ i = 1,\ldots,n\big)$. Evidently, if $\mathbf{y}$ is not a Pareto critical (or stationary) point, for certain directions v, $f_{\mathbf{y}}(v) < 0$, and the $\min - \max$ itself is strictly negative. This formulation is thus equivalent to choosing v such that all the directional derivatives are strictly negative, and the smallest in absolute value is as large as possible; i.e. equivalent to maximizing the minimum descent. From there, they have constructed algorithms that accumulate at Pareto critical points, and relaxed the condition using different norms. We put momentarily the comparison between their formulation and ours, and point out that they later extended their theory quite technically in [15] and developed classes of steepest-descent methods different from ours which is devised from a simpler, but very general geometric property. Note that from a design-point that is not Pareto critical, or stationary, infinitely many directions exist along which the directional derivatives of all the objective functions are of a strict given sign, and many practical algorithms can be constructed to be appropriate in the application context.

Clearly, the above condition (2), as it only involves scalar products, can be applied to projected gradients, in case of constrained minimization. More specifically, suppose that the active scalar constraints at $\mathbf{y} = \mathbf{y}^0$ are the following:

$$g_1(\mathbf{y}^0) = g_2(\mathbf{y}^0) = \cdots = g_K(\mathbf{y}^0) = 0\,, \tag{7}$$

and define the vectors

$$v_k = \nabla g_k(\mathbf{y}^0) \qquad (k = 1,\ldots,K)\,, \tag{8}$$

normal to the constraint surfaces, and assumed to be linearly-independent. Apply the Gram-Schmidt orthogonalization process (*GSP*) to them to get a family $\{w_k\}$ ($k = 1,\ldots,K$) of orthonormal vectors that collectively span the same subspace. Define the following projection matrix:

$$\mathbf{P} = \mathbf{I}_N - \sum_{k=1}^{K} [w_k]\,[w_k]^T\,, \tag{9}$$

where the bracketed vector $[w_k]$ stands for the column-vector of its components viewed as a $N \times 1$ matrix, and the superscript ^{T} indicates transposition. Then, the forthcoming *MGDA* construction is meant to apply after the original gradients, $\nabla J_i(\mathbf{y}^0)$, have been replaced by their projections onto the subspace tangent to the constraint surfaces, that is by $\mathbf{P}\nabla J_i(\mathbf{y}^0)$. Current research developments are focused on a more systematic treatment of constraints and will be the main topic of a future publication. Presently, without great loss of generality, we are considering thereafter the unconstrained formulation.

In the original formulation of *MGDA* [8]-[9], the vector ω has been defined as the minimum-norm element in the convex hull of the gradients:

$$\omega = \arg\min_{u\in\overline{\mathrm{U}}} \|u\|\,,\ \overline{\mathrm{U}} = \left\{ u \in \mathbb{R}^N \,/\, u = \sum_{i=1}^{n} \alpha_i \nabla J_i(\mathbf{y}^0)\,;\ \alpha_i \geq 0\ (\forall i);\ \sum_{i=1}^{n} \alpha_i = 1 \right\}. \tag{10}$$

This definition is the most general; in particular, it is applicable whether the gradient vectors are linearly independent or not. The element ω can be identified by numerical minimization in the convex hull, which can be parameterized isomorphically to the hypercube $[0,1]^{n-1}$ (see [9]). This minimization can however be numerically delicate, and in fact, not necessary, as the subsequent versions of our construction demonstrate.

Remark 2. Restricting the search in (10) to convex combinations plays the same role as penalizing the norm in the min – max formulation of (6). But, while in [14] the solution of the min – max problem is automatically a descent direction, and in the sense defined by the normalization through the norm-penalty term, the best solution, we construct a descent direction from a purely-geometrical property and optimize the step-size by a similar min – max solution (see Subsection 3.3).

The convex hull can also be viewed as an affine structure, since:

$$\forall u \in \overline{\mathrm{U}} : \; u - u_n = \sum_{i=1}^{n} \alpha_i u_i - \left(\sum_{i=1}^{n} \alpha_i \right) u_n = \sum_{i=1}^{n-1} \alpha_i u_{n,i} \qquad (u_{n,i} = u_i - u_n) \,. \quad (11)$$

Hence, $\overline{\mathrm{U}} \subseteq \mathscr{A}_{n-1}$ (or using affine-space notations, $\dot{\overline{\mathrm{U}}} \subseteq \dot{\mathscr{A}}_{n-1}$), where $\mathscr{A}_{n-1}$ is a set of vectors pointing onto an affine sub-space $\dot{\mathscr{A}}_{n-1}$ of dimension at most $n-1$.

Let us examine these affine and vector structures, with the support of Fig. 1 drawn in the case $n = 3$. Here vectors are represented in the $\mathbb{R}^3$ affine space with a given origin O. The gradient vectors are here denoted $\{u_i\}$ $(i = 1,2,3)$. The convex hull of the gradients is the set of vectors of origin O pointing onto the triangle made of the 3 endpoints of $\{u_i\}$. This triangle lies in a plane (generally speaking a subspace of dimension at most $n-1$) denoted $\dot{\mathscr{A}}_2$. The orthogonal projection of O onto the plane $\dot{\mathscr{A}}_2$ is denoted $O^\perp$. The figure has been drawn in the case where $O^\perp \notin \dot{\overline{\mathrm{U}}}$.

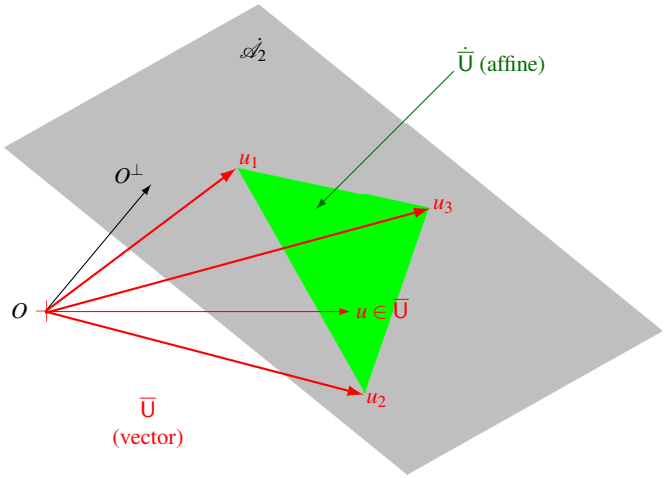

Fig. 1 Affine and vector structures: here, three vectors $\{u_i\}$ $(i = 1,2,3)$ are considered to define the convex hull $\overline{\mathrm{U}}$; the endpoints of their representatives of origin O are the vertices of the green triangle, $\dot{\overline{\mathrm{U}}}$, affine structure associated with the convex hull $\overline{\mathrm{U}}$; u is an arbitrary element in $\overline{\mathrm{U}}$; $\dot{\overline{\mathrm{U}}}$ lies in the plane $\dot{\mathscr{A}}_2$; $O^\perp$ is the orthogonal projection of O onto $\dot{\mathscr{A}}_2$; the figure illustrates the case where $O^\perp \notin \dot{\mathscr{A}}_2$.

Now, consider the inverse, highly-favorable situation in which $O^\perp \in \dot{\overline{\mathrm{U}}}$, or equivalently, $\overrightarrow{OO^\perp} \in \overline{\mathrm{U}}$. Since $\overrightarrow{OO^\perp} \perp \dot{\overline{\mathrm{U}}}$, $\omega = \overrightarrow{OO^\perp}$, and by orthogonality:

$$\left(u_i, \omega\right) = \|\omega\|^2 \qquad (\forall i) \,. \quad (12)$$

As a result, the directional derivatives of all objective-functions are equal.

The element ω being defined, the *MGDA* iteration is a form of generalization of the classical steepest-descent method [2] to multi-objective optimization in which

the vector $-\omega$ is used as the direction of search. Under certain weak provisions on the problem formulation, if the step-size ρ is adjusted optimally, the iteration accumulates at a Pareto-stationary design-point [8]. Whenever $\omega = 0$[1], the current design-point is Pareto-stationary, and the optimization is interrupted. Hence, in [22], the efficacy of *MGDA* to identify the Pareto front has been demonstrated, and comparisons with an evolutionary strategy (*PAES*) have been made.

More recently, a variant, *MGDA-II*, has been proposed in which an alternate descent direction is calculated by a direct procedure [10] based on a *GSP* with special normalization. In the basic version of the algorithm, the gradient vectors are required to be linearly independent. Additionally, due to the numerically observed importance of scaling, user-supplied scaling factors $\{S_i\}$ $(i = 1,\ldots,n)$, are assumed to be given, and the following scaled gradients are defined:

$$J_i' = \frac{\nabla J_i(\mathbf{y}^0)}{S_i} \tag{13}$$

($S_i > 0$; e.g. $S_i = J_i$ for logarithmic gradients). The *GSP* is performed as follows:

- Set $u_1 = J_1'$
- For $i = 2,\ldots,n$, set:

$$u_i = \frac{J_i' - \sum_{k<i} c_{i,k} u_k}{A_i}\,, \tag{14}$$

where for some arbitrary but small ε_i:

$$c_{i,k} = \frac{\left(J_i', u_k\right)}{\left(u_k, u_k\right)}\ (\forall k < i)\,,\ \text{and}\ A_i = \begin{cases} 1 - \displaystyle\sum_{k<i} c_{i,k} & \text{if nonzero,} \\ \varepsilon_i & \text{otherwise.} \end{cases} \tag{15}$$

As a result of this construction, a new element ω is defined, as the minimum-norm element in the convex hull of the orthogonal vectors $\{u_i\}$ $(i = 1,\ldots,n)$,

$$\omega = \sum_{i=1}^{n} \alpha_i u_i\,, \tag{16}$$

in which the coefficients $\{\alpha_i\}$ are strictly positive and less than 1:

$$\alpha_i = \frac{1}{\|u_i\|^2 \displaystyle\sum_{j=1}^{n} \frac{1}{\|u_j\|^2}} = \frac{1}{1 + \displaystyle\sum_{j\neq i} \frac{\|u_i\|^2}{\|u_j\|^2}} < 1\,. \tag{17}$$

Due to the orthogonality of the family $\{u_i\}$ $(i = 1,\ldots,n)$, $\omega = \overrightarrow{OO^1}$, and (12) holds. This is illustrated by Fig. 2. Consequently

[1] In the numerical implementation, the condition must be relaxed to $\|\omega\| < TOL$.

$$\left(J_i', \omega\right) = \left(A_i + \sum_{k<i} c_{i,k}\right) \|\omega\|^2 = \|\omega\|^2 \,, \tag{18}$$

by definition of the normalization constant A_i. In conclusion, the directional derivatives of all objective-functions are here systematically equal.

It should be emphasized that in general the newly-defined element ω is distinct from the former, except in the particular case of two objective-functions ($n = 2$), when the gradient vectors form an obtuse angle.

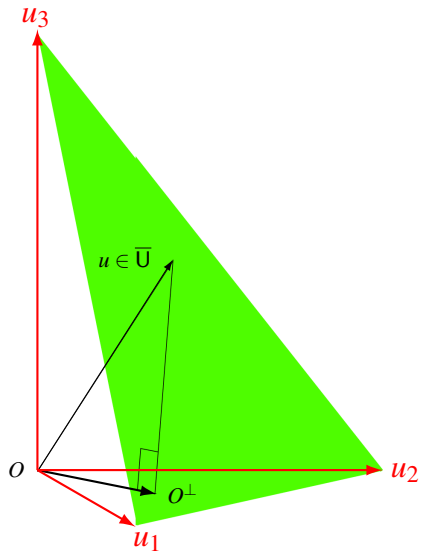

Fig. 2 Construction of orthogonal vectors in *MGDA-II* .

It was later observed [11] that situations in which a normalization constant A_i was negative for some i, was to be avoided. To do this, in *MGDA-II b*, the definition

$$A_i = 1 - \sum_{k<i} c_{i,k} \,, \tag{19}$$

was maintained only if this number is strictly-positive. Otherwise, the difficulty can be circumvented by redefining the corresponding scale to be:

$$S_i' = \left(\sum_{k<i} c_{i,k}\right) S_i \,, \tag{20}$$

so that:

$$c_{i,k}' = \left(\sum_{k<i} c_{i,k}\right)^{-1} c_{i,k} \,, \tag{21}$$

and

$$\sum_{k<i} c'_{i,k} = 1\,, \tag{22}$$

and $A_i = \varepsilon_i$, for some small ε_i. (This procedure was referred to as the "automatic rescale procedure".)

This has led us to the same formal conclusion: the directional derivatives are equal; but the value is much larger, and (at least) one objective-function has been rescaled.

This variant was tested on a somewhat peculiar model problem of domain-partitioning, in which all objective-functions tend to 0, which results in a Pareto set restricted to a single point. In this application rather remote from the context for which *MGDA* was originally devised, using logarithmic scaling of the gradients ($S_i = J_i$) and automatic rescale, *MGDA-II b* was found to converge satisfactorily; in fact, at a rate only twice less than the optimal quasi-Newton method; additionally, the iteration indicated asymptotically an interesting trend to convergence acceleration [11].

Nevertheless, these developments have brought up some open questions to which the following sections bring certain answers. In particular, the following three:

1. Since the element ω provided by *MGDA-II* is in general different from the original one, how can we guarantee the convergence of *MGDA* to a Pareto-stationary design-point?
2. In which order should the gradients be arranged to perform the *GSP* ?
3. Can the scaling of the gradients be adequately devised to mimic quasi-Newton methods?

MGDA-III provides an answer to the first two questions, and *MGDA-IV* to the third.

2 Ordered and economical *GSP* : *MGDA-III*

The steering idea is that in case of numerous gradients, trends might emerge among them, permitting to account for the general direction of a subgroup by a unique vector in the orthogonal basis. Hence the *GSP* could be interrupted as soon as a direction is found to be a descent direction common to all objective functions while being constructed on the basis of only $I < n$ gradients. To achieve this purpose, in the following algorithm, at the stage of computing a new orthogonal vector, the gradient that is elected among those not yet accounted for, is the one for which the scalar product with the current estimate of the element ω is algebraically smallest. In this way, it is the vector for which the construction so far is the least satisfactory. Thus, computational economy is achieved through the specification of the ordering in which the gradients are considered to perform the *GSP* , with the expectation of a rapid interruption of the process. Further comments on the expected gain in efficiency will be made a posteriori.

2.1 Algorithm

Again, one starts from scaled gradients, $\{J'_i\}$, and duplicates, $\{g_i\}$,

$$J'_i = \frac{\nabla J_i(\mathbf{y}^0)}{S_i} \quad (S_i: \text{ user-supplied scale}), \tag{23}$$

and proceeds in three steps: A, B and C.

A: Initialization

- Set[2]
$$k = \arg\max_i \min_j \frac{\left(J'_j, J'_i\right)}{\left(J'_i, J'_i\right)} \text{ and } u_1 = g_1 := J'_k. \tag{24}$$
- Set $n \times n$ lower-triangular matrix $c = \{c_{i,j}\}$ $(i \geq j)$ to 0[3].
- Set, conservatively, $I := n$[4].
- Assign some appropriate value to a cut-off constant $a : (0 \leq a < 1)$.

B: Main *GSP* loop

For $i = 2, 3, \ldots,$ (at most) n, do:

1. Calculate the $i-1$st column of coefficients:
$$c_{j,i-1} = \frac{\left(g_j, u_{i-1}\right)}{\left(u_{i-1}, u_{i-1}\right)} \qquad (\forall j = i, \ldots, n), \tag{25}$$
and update the cumulative row-sums:
$$c_{j,j} := c_{j,j} + c_{j,j-1} = \sum_{k<i} c_{j,k} \qquad (\forall j = i, \ldots, n). \tag{26}$$
2. Test:
   - If the following condition is satisfied
   $$c_{j,j} > a \qquad (\forall j = i, \ldots, n), \tag{27}$$
   set $I := i-1$, and interrupt the *GSP* (go to 3).
   - Otherwise, compute next orthogonal vector u_i as follows (steps a-b-c-d):

[2] The choice made for u_1 will be justified afterwards.

[3] The main diagonal of matrix c is to contain cumulative row-sums.

[4] The integer $I \leq n$ is the expected number of computed orthogonal basis vectors.

a - *Identify index* $\ell = \arg\min_j \ \{c_{j,j} \ / \ i \le j \le n\}$.[5]
b - *Permute information associated with* i *and* ℓ :

$$g\text{-vectors: } g_i \rightleftharpoons g_\ell\,,$$
rows i and ℓ of array c and corresponding cumulative row-sums,
$$c_{i,i} \rightleftharpoons c_{\ell,\ell}\,.$$

c - *Set* $A_i = 1 - c_{i,i} \ge 1 - a > 0$ *and calculate*[6]

$$u_i = \frac{g_i - \sum_{k<i} c_{i,k} u_k}{A_i}\,. \tag{28}$$

d - *If* $u_i \neq 0$*, return to 1. with incremented* i*; otherwise:*

$$g_i = \sum_{k<i} c_{i,k} u_k = \sum_{k<i} c'_{i,k} g_k\,, \tag{29}$$

where the $\{c'_{i,k}\}$ *are calculated by backward substitution.*
Then, if $c'_{i,k} \le 0$ $(\forall k < i)$*:*

Pareto-stationarity detected: STOP MGDA iteration;

otherwise (exceptional ambiguous case):

STOP GSP *; compute* ω *according to original definition and go to C.*

(end of a-b-c-d)

3. Calculate ω as the minimum-norm element in the convex hull of $\{u_1, u_2, \ldots, u_I\}$:[7]

$$\omega = \sum_{i=1}^{I} \alpha_i u_i \neq 0\,, \tag{30}$$

where:

$$\alpha_i = \frac{1}{\|u_i\|^2 \sum_{j=1}^{I} \frac{1}{\|u_j\|^2}} = \frac{1}{1 + \sum_{j\neq i} \frac{\|u_i\|^2}{\|u_j\|^2}}\,. \tag{31}$$

C: Descent step or termination

If $\|\omega\| < TOL$, STOP *MGDA* iteration; otherwise, perform descent step and return to Step B.

[5] Note that necessarily $c_{\ell,\ell} \le a < 1$.

[6] $c_{i,i}$ = former-$c_{\ell,\ell} \le a$; g_i = former-$c_{\ell,k}$.

[7] Note that ω is calculated on the basis of a smaller number of gradients if $I < n$; here all computed $u_i \neq 0$, and $0 < \alpha_i < 1$.

2.2 Properties

Case $\mathbf{I} = \mathbf{n}$.

In this case, the *GSP* is performed completely, and the algorithm is equivalent to the former *MGDA-II* with the enhancement that the rescale of the *b-version* is no longer ever necessary, since the specified ordering implies that

$$\forall i : A_i \geq 1 - a > 0\,. \tag{32}$$

Case $\mathbf{I} < \mathbf{n}$ (incomplete *GSP*).

Here, the directional derivatives satisfy different bounds according to two subcases:

- First I directional derivatives:

$$\big(g_i, \omega\big) = \big(u_i, \omega\big) = \|\omega\|^2 > 0 \qquad (\forall i = 1, \ldots, I)\,. \tag{33}$$

- Subsequent ones ($i > I$):

By construction, the vectors $\{u_1, u_2, \ldots, u_I\}$ are orthogonal, and ω is given by (30), so that:

$$g_i = \sum_{k=1}^{I} c_{i,k} u_k + v_i\,, \tag{34}$$

where $v_i \perp \{u_1, u_2, \ldots, u_I\}$. Consequently,

$$\big(g_i, \omega\big) = \sum_{k=1}^{I} c_{i,k} \big(u_k, \omega\big) = \sum_{k=1}^{I} c_{i,k} \|\omega\|^2 = c_{i,i} \|\omega\|^2 > a \|\omega\|^2 > 0\,. \tag{35}$$

Note that this bound is only slightly less favorable than (33), and this depends on the chosen cut-off constant a.

2.3 A posteriori justification of the choice of u_1

At initialization, we have set $u_1 = g_1$ according to (24). We now see that this was equivalent to maximizing $c_{2,1} = c_{2,2}$, that is, maximizing the least cumulative row-sum, at first estimation. Hence, at start, the worst case is less severe. One anticipates that the favorable situation for which all cumulative row-sums are positive (or $> a$), is more likely to occur.

2.4 Expected benefits

According to the section above, the specified ordering has been devised to permit the *GSP* to be performed incompletely. When gradients exhibit a general trend, ω is found in fewer steps and this realizes a computational economy.

Secondly, the rescale procedure, no longer ever necessary, is abandoned.

Thirdly, an incomplete *GSP* results in an element ω of larger norm since it realizes the minimization in a smaller subset, namely the convex hull of an incomplete orthogonal basis. This corresponds to larger directional derivatives, since

$$\left(g_i, \frac{\omega}{\|\omega\|}\right) = \|\omega\| \text{ or } a\,\|\omega\|\,, \tag{36}$$

and to the greater efficiency of the subsequent *MGDA* descent step.

3 Using Hessians to better scale the gradients: *MGDA-IV*

3.1 Addressing the question of scaling when Hessians are known

In single-objective optimization, when both gradient and Hessian are known, Newton's method is the most effective unless additional information is provided.

For the optimization of the objective $J_i(\mathbf{y})$ alone, Newton's method writes:

$$\mathbf{y}^1 = \mathbf{y}^0 - p_i\,, \tag{37}$$

where the vector p_i is given by the solution of the system:

$$H_i p_i = \nabla J_i(\mathbf{y}^0)\,, \tag{38}$$

where H_i is the Hessian matrix of objective function J_i at $\mathbf{y} = \mathbf{y}^0$. Hence the preconditioning by the inverse Hessian realizes a form of optimal scaling. However, in general, the vector p_i is not parallel to the gradient itself. Thus to ensure that the iteration remains a descent step, only its projection should be retained.

Thus, we propose to split the vector p_i into orthogonal components

$$p_i = q_i + r_i\,, \tag{39}$$

where:

$$q_i = \frac{\left(p_i, \nabla J_i(\mathbf{y}^0)\right)}{\|\nabla J_i(\mathbf{y}^0)\|^2}\nabla J_i(\mathbf{y}^0)\,, \tag{40}$$

is along the gradient, and $r_i \perp \nabla J_i(\mathbf{y}^0)$, and to define the scaled gradient as follows:

$$J_i' = q_i\,. \tag{41}$$

We thus define *MGDA-IV* as *MGDA-III* applied to the gradients scaled as above. This is equivalent to defining the scaling constant S_i as follows:

$$S_i = \frac{\left\|\nabla J_i(\mathbf{y}^0)\right\|^2}{\left(p_i, \nabla J_i(\mathbf{y}^0)\right)} = \frac{\left\|\nabla J_i(\mathbf{y}^0)\right\|^2}{\left(H_i^{-1}\nabla J_i(\mathbf{y}^0), \nabla J_i(\mathbf{y}^0)\right)} > 0 \tag{42}$$

where the inequality holds if H_i (or H_i^{-1}) is positive-definite (convexity).

3.2 BFGS-inspired variant: MGDAIVb

When the Hessians are not known exactly, they can be approximated by the Broyden-Fletcher-Goldfarb-Shanno (*BFGS*) iterative estimate (see e.g. [2], Section 4.5):

$$(\forall i = 1, \ldots, n) \quad \widetilde{H}_i^{(0)} = Id\,, \tag{43}$$

$$\widetilde{H}_i^{(k+1)} = \widetilde{H}_i^{(k)} - \frac{1}{s^{(k)T}\widetilde{H}_i^{(k)}s^{(k)}}\widetilde{H}_i^{(k)}s^{(k)}s^{(k)T}\widetilde{H}_i^{(k)T} + \frac{1}{z_i^{(k)T}s^{(k)}}z_i^{(k)}z_i^{(k)T}\,, \tag{44}$$

$$s^{(k)} = \mathbf{y}^{(k+1)} - \mathbf{y}^{(k)}, \quad z_i^{(k)} = \nabla J_i(\mathbf{y}^{(k+1)}) - \nabla J_i(\mathbf{y}^{(k)}) \quad (k\text{: } MGDA \text{ iteration index}). \tag{45}$$

3.3 Recommended step-size

The Taylor's series expansion to second-order of objective-function $J_i(\mathbf{y})$ about $\mathbf{y}^0$ corresponding to the increment $\delta\mathbf{y} = -\rho\,\omega$ of the design variable writes:

$$J_i(\mathbf{y}^0 - \rho\,\omega) = J_i(\mathbf{y}^0) - \rho\left(\nabla J_i(\mathbf{y}^0), \omega\right) + \tfrac{1}{2}\rho^2\left(H_i\omega, \omega\right) + \ldots\,. \tag{46}$$

Neglecting the third-order terms in the above expansion yields the following expression for the expected decrease of the objective-function:

$$|\delta J_i| := J_i(\mathbf{y}^0) - J_i(\mathbf{y}^0 - \rho\,\omega) := \bar{a}_i\rho - \tfrac{1}{2}\bar{b}_i\rho^2\,, \tag{47}$$

where: $\bar{a}_i = S_i a_i$, $\bar{b}_i = S_i b_i$, and:

$$a_i = \left(J_i', \omega\right) \qquad b_i = \left(H_i\omega, \omega\right)/S_i\,. \tag{48}$$

The above coefficients are known numerically, exactly or approximately. The coefficients $\{a_i\}$ are positive by construction, and for the first I of them equal to $\|\omega\|^2$. The coefficients $\{b_i\}$ are also positive by assumption of convexity.

If the scales $\{S_i\}$ are truly relevant for the objective-functions, it makes sense to maximize the relative decrease

$$\delta_i = \frac{|\delta J_i|}{S_i} = a_i\rho - \tfrac{1}{2}b_i\rho^2\,. \tag{49}$$

Hence we may define the optimal step-size to be:

$$\rho^\star = \arg\max_\rho \min_i \delta_i\,. \tag{50}$$

If $I = n$ (case a) in Fig. 3), $a_i = \|\omega\|^2\ \ (\forall i)$, and

$$\rho^\star = \rho_I^\star := \frac{\|\omega\|^2}{b_I}\,, \tag{51}$$

where $b_I = \max_{i\le I} b_i$ is assumed to be positive (convexity).

If instead, $I < n$, a complementary subset of objective-functions satisfy a less favorable bound related to the chosen cut-off constant a. For those, we may conservatively approximate $\big(J_i,\omega\big) \doteq a\,\|\omega\|^2$. For those alone, the optimal step-size is:

$$\rho_{II}^\star := \frac{a\,\|\omega\|^2}{b_{II}}\,, \tag{52}$$

where $b_{II} = \max_{i>I} b_i$ is assumed to be positive (convexity).

Another special value of ρ is $\rho_\times$ for which the two bounding $\delta_i(\rho)$-curves associated with the two subsets intersect, that is the solution of the equation:

$$\|\omega\|^2\,\rho - \tfrac{1}{2}b_I\rho^2 = a\,\|\omega\|^2\,\rho - \tfrac{1}{2}b_{II}\rho^2\,. \tag{53}$$

This gives:

$$\rho_\times = \frac{2(1-a)\,\|\omega\|^2}{b_I - b_{II}}\,. \tag{54}$$

Consider first the case $b_I > b_{II}$ for which $\rho_\times > 0$. One finds the equivalences:

$$\rho_\times < \rho_I^\star \Longleftrightarrow a > \beta_1 := \frac{b_I + b_{II}}{2b_I}\,,\ \text{and}\ \ \rho_\times < \rho_{II}^\star \Longleftrightarrow a > \beta_2 := \frac{2b_{II}}{b_I + b_{II}}\,. \tag{55}$$

Note that $\beta_1 \ge \beta_2$. Then three sub-cases are possible:

1. $a < \beta_2$: $\rho_\times > \rho_I^\star$ and $\rho_\times > \rho_{II}^\star$; $\rho^\star = \max(\rho_I^\star, \rho_{II}^\star)$ as in Fig. 3 b);
2. $\beta_2 \le a \le \beta_1$: $\rho_I^\star \le \rho_\times \le \rho_{II}^\star$; $\rho^\star = \rho_\times$ as in Fig. 3 c);
3. $a > \beta_1$: $\rho_\times < \rho_I^\star$ and $\rho_\times < \rho_{II}^\star$; $\rho^\star = \min(\rho_I^\star, \rho_{II}^\star)$, as in Fig. 3 d).

Lastly, if $b_I \le b_{II}$, $\rho_\times \le 0$ and $\rho^\star = \min(\rho_I^\star, \rho_{II}^\star)$.

In summary, if $I = n$, $\rho^\star = \rho_I^\star$; otherwise ($I < n$), $\rho^\star$ is the element of the triplet $\{\rho_I^\star, \rho_{II}^\star, \rho_\times\}$ which separates the other two.

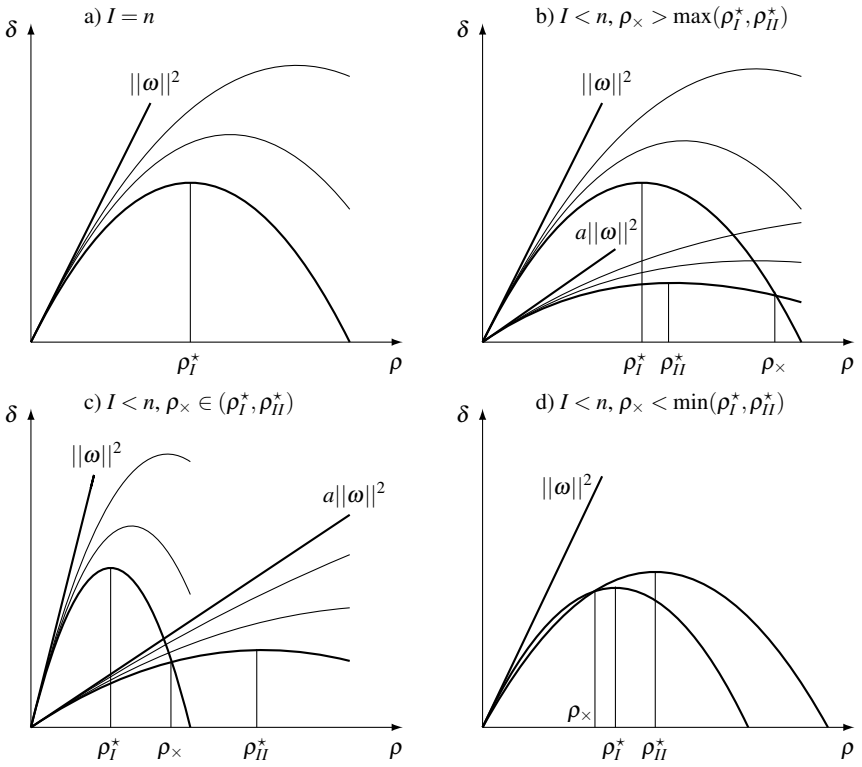

Fig. 3 Cases of interest in adjusting the *MGDA* -iteration step-size ρ. In case a), $I = n$ and $\rho^\star = \rho_I^\star$; in cases b), c) and d), $I < n$, and $\rho^\star$ is the element of the triplet $\{\rho_I^\star, \rho_{II}^\star, \rho_\times\}$ which separates the other two, i.e. $\rho_{II}^\star$, $\rho_\times$ and $\rho_I^\star$ respectively.

4 Conclusion

The different proposed variants of the MGDA are summarized in Table 1 where the major enhancements realized by each version are indicated, and references to publications provided. An incomplete *GSP* based on an ordered subset of the gradients is recommended to define the support vector ω of the search direction as the minimum-norm element in the convex hull of this subset. When Hessians are available, an estimate of the optimal step-size has also been identified.

A special focus is currently being devoted to a more systematic account for the constraints. Further work will also be directed on experimenting *MGDA-III* and *MGDA-IV* and assessing the actual efficiency improvements in practical engineering applications in which the Pareto fronts correspond to $n > 2$, and possibly involves discontinuities. Scaling with exact or approximate Hessians is a very promising option, but will be verified by cautious experiments, as well as the adequacy of the optimal step-size estimate.

Algorithm	Novel elements	Properties	Tested cases
MGDA	- General constructive principle related to minimum-norm element in convex hull of gradients [8] [9]	- CV proof to Pareto stationary designs - Insensitive to Pareto front convexity	- Multiple quadratics - Fonseca testcase (non-convex Pareto front; comparison *MGDA* vs *PAES*) [22] - DDM for Poisson pb. [11] [12]
b-version	- Meta-model assisted gradient computation [23]	- CV requires a few database enrichments	- Eulerian flow about wing [23] - Navier-Stokes duct flow [23]
MGDA-II	- Direct computation of descent direction by *GSP* [10] [11]	- Requires linearly independent gradients - Modified definition of descent direction, and Pareto-stationarity test necessary - $n!$ possible orderings	- DDM for Poisson pb. (scaling essential; verified CV to unique Pareto-stationary solution) [11]
b-version	- Automatic gradient rescale when normalization coefficient is found < 0	- More efficient (larger directional derivatives)	- (id.)
MGDA-III	- Specific ordering in *GSP* - Incomplete *GSP* - Resort to standard *MGDA* when Pareto stationarity test ambiguous	- Not limited to linearly-independent gradients - Even larger directional derivatives - Pareto-stationary accumulation points	
MGDA-IV	- Scaling inspired from Newton's method using Hessians	- Step-size estimate provided	
b-version	- Uses *BFGS* approximations to Hessians		

Table 1 Variants of *MGDA* with details on progressive enhancements

References

1. *Multiobjective Optimization Using Evolutionary Algorithms*, K. Deb. Wiley, New York, NY (2001).
2. *Practical Optimization*, Ph. E. Gill and W. Murray and M. H. Wright. *Academic Press*, New York London (1986).
3. *Theory of vector optimization*, D.T. Luc. *Springer-Verlag, Berlin* (1988).
4. *Nonlinear Multiobjective Optimization*. R. Miettinen. *Kluwer Academic Publishers*, Boston London Dordrecht (1999).
5. *Vektoroptimierung Theorie, Verfahren und Anwendungen*, A. Göpfert and R. Nehse. *B.G. Teubner Verlagsgesellschaft, Leipzip*, (1990).
6. N. Srinivas and K. Deb, "Multi-objective function optimization using non-dominated sorting genetic algorithms". *Evolutionary Computation, Vol. 2, No. 3, pp. 221 248* (1995).
7. I. Das and J. Dennis, "Normal Boundary Intersection: an alternate method for generating Pareto optimal points in multicriteria optimization problems". *ICASE Report No. 96-62, NASA Langley Research Center, Hampton, Virginia* (1996).
8. J.-A. Désidéri, "Multiple-Gradient Descent Algorithm (MGDA)". *INRIA, Research Report 6953* (2009). Revised version, November 5 (2012). `http://hal.inria.fr/inria-00389811`
9. J.-A. Désidéri, "Multiple-Gradient Descent Algorithm (MGDA) for Multiobjective Optimization". *Comptes rendus - Mathématique*, Vol. 350, Issues 5-6, March 2012, pp. 313-318 (2012). `http://dx.doi.org/10.1016/j.crma.2012.03.014`
10. J.-A. Désidéri, "MGDA II: A direct method for calculating a descent direction common to several criteria". *INRIA, Research Report 7422, April* (2012). `http://hal.inria.fr/hal-00685762`
11. J.-A. Désidéri, "Application of MGDA to domain partitioning". *INRIA, Research Report 7968, May* (2012). `http://hal.inria.fr/hal-00694039`
12. J.-A. Désidéri, "Multiple-Gradient Descent Algorithm for Multiobjective Optimization". *Proc. European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS), September 10-14, 2012 Vienna, Austria* (2012). To be referenced in Scopus.
13. J.-A. Désidéri, "Cooperation and competition in multidisciplinary optimization". *Comput. Optim. Appl. 52:29-68* (2012).
14. J. Fliege and B. F. Svaiter, "Steepest descent methods for multicriteria optimization". *Mathematical Methods of Operations Research*, 51:479-494 (2000).
15. L.M. Graña Drummond and B.F. Svaiter, "A steepest descent method for vector optimization". *J. Comput. & App. Math.* 175:395-414 (2005).
16. M. Hartikainen, K. Miettinen and M. M. Wiecek, "Constructing a Pareto front approximation for decision making". *Math. Meth. Oper. Res.*, Vol 73, No. 2, pp 209-234 (2011).
17. M. Hartikainen, K. Miettinen and M. M. Wiecek, "PAINT: Pareto front interpolation for nonlinear multiobjective optimization". *Comput. Optim. and Appl., Springer*, 52:845-867 (2012).
18. J. R. R. A. Martins and A. B. Lambe, "Multidisciplinary Design Optimization: A Survey of Architectures". *AIAA Journal: 1-27, 10.2514/1.J051895*, posted online 16 July (2013).
19. A. Messac, A. Ismail-Yahaya, C.A. Mattson, "The normalized normal constraint method for generating the Pareto frontier". *Struct. Multidisc. Optim.* 25:86-98 (2003).
20. J. Sobieszczanski and R.T. Haftka, "Multidisciplinary aerospace design optimization: survey of recent developments". *Struct. Optim. 14, 123* (1997).
21. J. Sobieszczanski-Sobieski, T. Altus, and R.R. Sandusky, "Bilevel integrated system synthesis for concurrent and distributed processing". *AIAA J. 41, 19962003* (2003)
22. A. Zerbinati, J.-A. Désidéri and R. Duvigneau, "Comparison between MGDA and PAES for Multi-Objective Optimization". *INRIA, Research Report 7667, May* (2011). `http://hal.inria.fr/inria-00605423`
23. A. Zerbinati, J.-A. Désidéri and R. Duvigneau, "Application pf the Multiple-Gradient Descent Algorithm (MGDA) and Metamodels to a Multiobjective Problem in Aerodynamics". *Proc. European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS), September 10-14, 2012 Vienna, Austria* (2012). To be referenced in Scopus.