Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Multicriteria Decision Making (MCDM 2007)

On the Convergence of Multi-Objective Descent Algorithms

Martin Brown and Nicky Hutauruk
The Control Systems Centre, School of Electrical and Electronic Engineering, University of Manchester, UK
Martin.Brown@manchester.ac.uk
N.Hutauruk@student.manchester.ac.uk

Abstract – This paper investigates the convergence paths, rate of convergence and the convergence half-space associated with a class of descent multi-objective optimization algorithms. The first order descent algorithms are defined by maximizing the local objectives' reductions which can be interpreted in either the primal space (parameters) or the dual space (objectives). It is shown that the convergence paths are often aligned with a subset of the objectives gradients and that, in the limit, the convergence path is perpendicular to the local Pareto set. Similarities and differences are established for a range of p-norm descent algorithms. Bounds on the rate of convergence are established by considering the stability of first order learning rules. In addition, it is shown that the multi-objective descent algorithms implicitly generate a half-space which defines a convergence condition for family of optimization algorithms. Any procedure that generates updates that lie in this half-space will converge to the local Pareto set. This can be used to motivate the development of second order algorithms.

I. INTRODUCTION

Any multi-objective optimization process, whether it is point-wise or population-based, is uniquely defined by the design parameters, the objectives and the set of constraints. For industrial design problems, the objectives typically include cost, performance, appearance and robustness measures. Jointly minimizing each quantity is, in general, not possible as the objectives typically conflict close to optimal solutions. In fact, Pareto optimality is sometimes referred to as "zero sum optimality", as it is not possible to decrease an objective without increasing at least one other [4,5].

There are many approaches to solving multi-objective optimization problems, including evolutionary algorithms where the aim is to approximate the complete Pareto set with a population of points, and the population tries to improve its overall fitness at each iteration using operators inspired by evolutionary concepts. These techniques have been applied to difficult problems, large search space, discrete solution sets, categorical variables etc. with some success [4,5]. Other approaches have investigated how the optimal solution set can be iteratively calculated by finding (at least) one point on the local Pareto set and then spanning out from that point, calculating the exact solution [5]. This paper is mainly concerned with analyzing the convergence of point-wise optimization where the aim is to project a point onto the local Pareto set. Population-based procedures can be analysed point-wise, or by taking statistical quantities [14] to analyze the average convergence of a set of points.

The algorithms are derived from attempting to minimize the maximum change across all objectives at each iteration, for a fixed size parameter update, [2,3,8]. It has been shown that such a strategy converges to a locally Pareto optimal solution, and this paper investigates the properties of these descent trajectories. It is shown that the descent trajectories are perpendicular to the tangent to the local Pareto set at the accumulation point, and this is equivalent to requiring that all objectives should be reduced equally at each iteration. It also shows that every point on the boundary of the local Pareto set is the corresponding accumulation point of a descent trajectory, and is hence reachable. These results are true for the complete set of p-norm steepest descent algorithms, as differences in their descent trajectories only appear far from the local Pareto set. This work is then expanded to consider how orthogonal descent trajectories in the objective space and be constructed and how the theory can be expanded to handle diverse, rather than strict descent, trajectories. It is shown that the local descent trajectories implicitly defines a local convergence half-space where so long as any diverse update lies in this half-space, convergence to the local Pareto set is guaranteed. Simple, illustrative examples are used throughout the paper.

II. DIRECTED MULTI-OBJECTIVE OPTIMIZATION ALGORITHMS

In this section, the basic notation and concepts associated with directed multi-objective optimization and its convergence are reviewed.

A. Multi-Objective Optimization

The standard formulation for a multi-objective problem is of the form:

$$\begin{aligned} &\min \mathbf{f}(\mathbf{x}) \\ &st \quad \mathbf{g}(\mathbf{x}) = \mathbf{0} \\ &\quad\;\; \mathbf{h}(\mathbf{x}) \le \mathbf{0} \end{aligned} \tag{1}$$

The design parameters, $\mathbf{x}$, lie in an n-dimensional space $\mathbf{X}$, and are assumed to be continuous. The objectives, $\mathbf{f}$, are lie in an m-dimensional space $\mathbf{F}$, and may represent raw measurements or transformed quantities. In this paper, it is assumed that $\mathbf{F}$ is differentiable. The constraints are specified by two vector functionals $\mathbf{g}()$ and $\mathbf{h}()$, that represent the equality and inequality constraints, respectively. The constraints represent sections of the design space that are not feasible or parts of the objective space that the design must beat, when the aim is to evolve rather than innovate a new design. In this paper, constraints are not explicitly considered in the theory, although many of the results can be simply extended when the constraints are linear.

The basic requirement of (1) is to jointly minimize $\mathbf{f}$, which imposes a partial ordering on the set of potential solutions. One design $\mathbf{x}_1$ **dominates** a design $\mathbf{x}_2$ when:

$$\mathbf{f}(\mathbf{x}_1) \le \mathbf{f}(\mathbf{x}_2) \tag{2}$$

and the vector comparison is taken element wise. Strictly speaking, there should be at least one objective that is strictly less in value. This can be written as

$$\mathbf{x}_1 \preceq_{\mathbf{f}} \mathbf{x}_2 \tag{3}$$

When a design is globally Pareto optimal, there does not exist any other point that dominates it. The set of all such points is the solution to (1). The **Pareto set** is the set of all such optimal points in parameter space and the **Pareto front** is the set of all such points in objective space. Like conventional single objective optimization, the Pareto set is invariant to positive linear transformations of the objectives.

A point is **locally Pareto optimal** if (2) or (3) is satisfied in a local neighbourhood about that point. The set of all such locally Pareto optimal points is a superset of the set of globally Pareto optimal points. This paper is largely concerned with calculating locally Pareto optimal points, and then the globally Pareto optimal points can be found by performing the dominance test (2) on the complete set of locally optimal points [4,5,13].

B. Primal Descent Problem

For a current candidate solution $\mathbf{x}_k$, the aim of a point-wise, iterative optimization process is to update it according to:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mu\mathbf{s} \tag{4}$$

where $\mathbf{s}$ is the n-dimensional search direction and μ is the step size. This obviously has parallels with conventional (single objective) optimization theory [7], and when the (single) objective is continuously differentiable, a variety of 1st and 2nd order methods are used to specify the search direction and step size.

So consider extending the 1st order, single objective optimization process to the multi-objective case. Assuming that the objectives are continuously differentiable, there are no constraints and it is required that the reduction in each objective at each iteration is maximal [13] (this implicitly defines a dynamic min/max scalarization on the objectives' updates). Then the aim is to calculate a multi-objective descent direction $\mathbf{s}$ which solves:

$$\begin{aligned}(\alpha^*, \mathbf{s}^*) &= \arg\min \alpha + \tfrac{1}{2}\|\mathbf{s}\|_2^2 \\ st \quad & \mathbf{J}^T\mathbf{s} \le \mathbf{1}\alpha\end{aligned} \tag{5}$$

which is a convex, quadratic programming (QP) problem with linear inequality constraints [2,3,8]. Here the inequality constraints ensure that $\mathbf{s}$ is a multi-objective descent direction as long as α is negative, as $\Delta\mathbf{f} = \mathbf{J}^T\mathbf{s} \le \mathbf{1}\alpha$ represents the local, 1st order reduction in the objectives. When α is negative, locally, the new point dominates the old one:

$$\mathbf{x}_{k+1} \preceq_{\mathbf{f}} \mathbf{x}_k \tag{6}$$

The objective in Equation 5 simply normalizes the length of the search direction, whilst minimizing the largest change in the objectives.

A number of properties about this steepest descent criterion can be shown [8]:

- If $\mathbf{x}$ is locally Pareto optimal, then $\mathbf{s}^*(\mathbf{x}) = \mathbf{0}$, and $\alpha^*(\mathbf{x}) = 0$.
- If $\mathbf{x}$ is not locally Pareto optimal, then $\alpha^*(\mathbf{x}) < 0$ and

$$\begin{aligned}\alpha^*(\mathbf{x}) &\le -\tfrac{1}{2}\left\|\mathbf{s}^*(\mathbf{x})\right\|_2^2 < 0 \\ \mathbf{J}^T(\mathbf{x})\mathbf{s}^*(\mathbf{x}) &\le \mathbf{1}\alpha^*(\mathbf{x})\end{aligned} \tag{7}$$

- The mappings $\mathbf{x} \rightarrow \mathbf{s}^*(\mathbf{x})$ and $\mathbf{x} \rightarrow \alpha^*(\mathbf{x})$ are continuous.

which show that solving this constrained optimization problem will produce a vector descent direction. In the following sections, the star superscript to denote the optimal value will be dropped, and $\mathbf{s}$ and α, will refer to their optimal values. In [8] it is shown that this steepest descent procedure causes a point to converge to an accumulation point on the local Pareto set.

Given that the parameter updates occur in the direction of the descent vector, $\mathbf{s}$, the step size still needs to be determined. As long as all the objectives are decreasing (not increasing), a line search can proceed along this direction and simply stop when one of the objectives starts to increase. Standard techniques can be used to automatically calculate an initial estimate of the step size, based on previous values [8].

Finally, it should be noted that it is possible to put preference/scaling information into the primal problem [3], by instead of specifying that all objectives should be reduced by the same amount, to require then to be reduced by $\boldsymbol{\beta}$:

$$\begin{aligned}(\alpha^*, \mathbf{s}^*) &= \arg\min \alpha + \tfrac{1}{2}\|\mathbf{s}\|_2^2 \\ st \quad & \mathbf{J}^T\mathbf{s} \le \boldsymbol{\beta}\alpha\end{aligned} \tag{8}$$

where $\boldsymbol{\beta}$ is an m-dimensional vector containing the relative importance of each objective. This work does not explicitly consider the problem of objective scaling [12], but this is important as Equation (6) implicitly assumes that the objectives are of a similar scale as the aim is to reduce them equally.

C. Dual Descent Problem

Instead of solving or analyzing the primal QP optimization sub-problem in Equation (5), consider the following Lagrangian:

$$L(\alpha, \mathbf{s}, \boldsymbol{\lambda}) = \alpha + \tfrac{1}{2}\mathbf{s}^T\mathbf{s} + \boldsymbol{\lambda}^T(\mathbf{J}^T\mathbf{s} - \mathbf{1}\alpha) \tag{9}$$

where $\boldsymbol{\lambda} > \mathbf{0}$ is the m-dimensional vector of Lagrange multipliers. This is solved when the KKT conditions are satisfied:

$$\begin{aligned}&\frac{\partial L}{\partial \alpha} = 1 - \mathbf{1}^T\boldsymbol{\lambda} = 0 \\ &\frac{\partial L}{\partial \mathbf{s}} = \mathbf{s} + \mathbf{J}\boldsymbol{\lambda} = \mathbf{0} \\ &\lambda_j((\mathbf{J}^T\mathbf{s})_j - \alpha) = 0 \qquad \forall j = 1,\ldots,m \\ &\boldsymbol{\lambda} \ge \mathbf{0}\end{aligned} \tag{10}$$

Re-arranging, the following relationships are produced:

$$\begin{aligned}&\mathbf{s} = -\mathbf{J}\boldsymbol{\lambda} \\ &\boldsymbol{\lambda} \ge \mathbf{0} \\ &\mathbf{1}^T\boldsymbol{\lambda} = 1\end{aligned} \tag{11}$$

so the multi-objective descent direction is lies in, or on the boundary of, the convex simplex generated by the individual (negative) gradients, as shown in Figure 1. In this figure, the **negative gradient simplex V** is defined as:

$$\mathbf{V} = \{-\mathbf{J}\boldsymbol{\lambda}: \boldsymbol{\lambda} \ge \mathbf{0}, \mathbf{1}^T\boldsymbol{\lambda} = 1\}$$

and, by definition, the search direction must lie on this simplex.

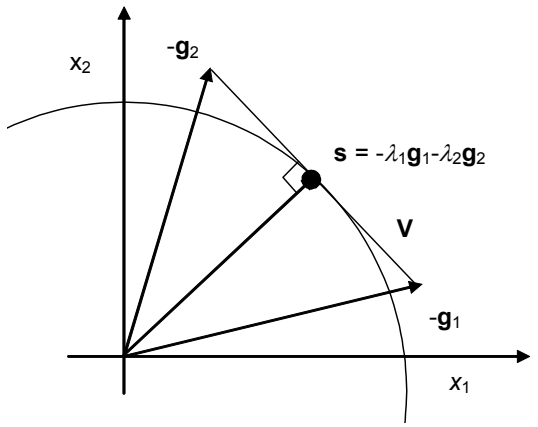

Figure 1 A geometric interpretation of calculating a vector gradient by minimising the Euclidean norm within the convex hull of the individual gradients.

Substituting these relationships into the primal sub-problem generates the dual QP problem [8]:

$$\begin{aligned} &\boldsymbol{\lambda}^* = \arg\min \tfrac{1}{2}\boldsymbol{\lambda}\mathbf{J}^T\mathbf{J}\boldsymbol{\lambda} \\ &st \quad \boldsymbol{\lambda} \geq \mathbf{0} \\ &\mathbf{1}^T\boldsymbol{\lambda} = 1 \end{aligned} \tag{12}$$

Again, as in the previous section, the star notation will be subsequently dropped. A simple interpretation of this is that the multi-objective descent **s** lies in the (negative) simplex defined by the individual columns of **J** (the individual gradients). Another point to notice is that the solving either the primal or dual QP problem is a superset of a test for local Pareto optimality, so only one QP problem actually needs to be solved if the aim is to calculate a descent direction onto the Pareto front, but checking first that the design is not already Pareto optimal. When a point is locally optimal, the solution to the dual problem will be $\mathbf{J}\boldsymbol{\lambda} = \mathbf{0}$. Also, it should be noted that the Lagrange multipliers can be thought of as representing dynamic "scalarization weights". The optimal values are a function of $\mathbf{x}, \boldsymbol{\lambda}(\mathbf{x})$, and because they multiply the functions' derivatives, the procedure converges even when the Pareto front is concave. Typically, static weights are used to explicitly set prior preferences about the desired accumulation point, and while these are generated dynamically by the algorithms, it is shown in Section III that these dynamic weights will uniquely specify a descent trajectory which implicitly specifies the accumulation point. Finally, by considering the linear inequality constraints in the primal QP problem, it can be easily seen that the negative multi-objective gradient always has a non-negative dot product with each of the individual gradient vectors, and so locally, each of the objectives will be non-increasing and the iterations will be locally dominant.

When one (or more) of the Lagrange multipliers is zero, it means that that objective does not influence the calculation of **s** and the corresponding objective reduction is strictly less than α. Typically, when a point is far from the Pareto set and the gradients are aligned but of differing magnitudes, many of the Lagrange multipliers will be zero. However, as a point approaches the Pareto set, the gradient vectors become conflicting and all of the Lagrange multipliers will be strictly positive.

Two important concepts arise from this work which will be used in evaluating the algorithms' convergence properties. The **multi-objective gradient**, **g**, (MOG) is defined as:

$$\begin{aligned} &\mathbf{g} = \mathbf{J}\boldsymbol{\lambda} \\ &\boldsymbol{\lambda} \geq \mathbf{0} \\ &\mathbf{1}^T\boldsymbol{\lambda} = 1 \end{aligned}$$

where $\boldsymbol{\lambda}$ solves Equation (12). Obviously, for the multi-objective, 2-norm, steepest descent algorithm the search direction is simply the negative gradient, although this will be relaxed later in the paper when other search directions are considered. Obviously, $\mathbf{g} \perp \mathbf{V}$ (which holds in the subspace defined by the non-zero Lagrange multipliers). Also, the **descent cone D** is defined as the set of updates that locally reduce all the gradients. It is defined in parameter space as the intersection of all the half-spaces, each of which corresponds to reducing one objective [2]. By definition, **s** must lie in **D** and when **D** is empty, the point must be locally Pareto optimal.

D. Alternative p-norm updates

As noted in [8], other p-norms could be used to bound the size of the parameter update when the min/max objective reduction is being calculated in (6), with two obvious candidates being the 1-norm and ∞-norm. The (scaled) primal descent problems would be expressed as:

$$\begin{aligned} &\min \quad \alpha && \min \quad \alpha \\ &\mathbf{J}^T\mathbf{s} \leq \mathbf{1}\alpha && \mathbf{J}^T\mathbf{s} \leq \mathbf{1}\alpha \\ &\|\mathbf{s}\|_1 = 1 && \|\mathbf{s}\|_\infty = 1 \end{aligned} \tag{13}$$

where it has been assumed that the parameter update **s** is normalised to length 1, which is slightly different from the 2-norm steepest descent algorithm described in Sections II.B and II.C. Both can be expressed as more familiar Linear Programming (LP) problems:

$$\begin{aligned} &\min \quad \alpha && \min \quad \alpha \\ &[\mathbf{J}^T - \mathbf{J}^T]\mathbf{z} \leq \mathbf{1}\alpha && \mathbf{J}^T\mathbf{s} \leq \mathbf{1}\alpha \\ &-\mathbf{z} \leq \mathbf{0} && \mathbf{s} \leq \mathbf{1} \\ &\mathbf{1}^T\mathbf{z} = 1 && -\mathbf{s} \leq \mathbf{1} \end{aligned} \tag{14}$$

where **z** is a vector of length $2n$ defined by $z_i = s_i$ if $s_i > 0$, $z_{n+i} = -s_i$ if $s_i < 0$ and $z_i = 0$ otherwise. Many solvers are available for calculating these steepest descent directions and the convergence properties noted in Section II.B also hold for these alternative formulations, as well as the more general *p*-norm measure.

As with any LP algorithm, it is possible to express the 1-norm and ∞-norm problems in their dual form. However, in both of these cases, the number of linear inequality constraints is approximately $m+2n$, which is substantially larger than the 2-norm case and less insight is gained about the structure of the problem in dual space. An analysis of the relative performance of these alternative steepest descent algorithms is performed in Section III.C. It is also worthwhile noting that for these alternative *p*-norms, **s** will

not, in general, lie in the negative gradient simplex because its size is constrained to be 1, although the direction will lie in the descent cone **D**.

III. CONVERGENCE OF 1ST ORDER DESCENT METHODS

In [8] it was shown that with a suitable step size calculation, the 2-norm steepest descent method converged to an accumulation point on the local Pareto set. In this section, the convergence of the 2-norm steepest descent algorithms are initially considered and it is shown that the descent trajectories are orthogonal to the local tangent to the Pareto set at the accumulation point. This is used to prove results about the reachability of points on the local Pareto sets and is then extended to analysing the convergence behaviour/trajectories of more general *p*-norms and considering the behaviour of descent trajectories in objective space.

A. Descent Trajectories and Point-Wise Convergence

Using the 2-norm steepest descent algorithm developed in Sections II.B and II.C, it is possible to consider the discrete parameter updates as an approximation to a continuous **descent trajectory**. The descent trajectory for a point shows the continuous path taken to reach the accumulation point in the local Pareto set. By (7), each descent trajectory is unique and they do not intersect. The descent trajectories are formed from solving either the primal or dual optimization problems (5) or (10) and taking infinitely small steps. The set of descent trajectories for a simple 2 parameter, 2 objective optimization problem is shown in Figure 2. The red arrows indicate points involving only two non-zero Lagrange multipliers, whereas the blue and green arrows involve only one.

The problem illustrated in Figure 2, is described by:

$$f_1(\mathbf{x}) = -\exp\left(-\left((x_1-0.3)^2+(x_2-0.1)^2\right)^{0.5}\right)$$

$$f_2(\mathbf{x}) = -\exp\left(-\left((x_1-0.7)^2+(x_2-0.9)^2\right)^{0.5}\right)$$

where the two parameters are assumed to lie in the interval [0,1]. The Pareto set is a straight line and the Pareto front is concave.

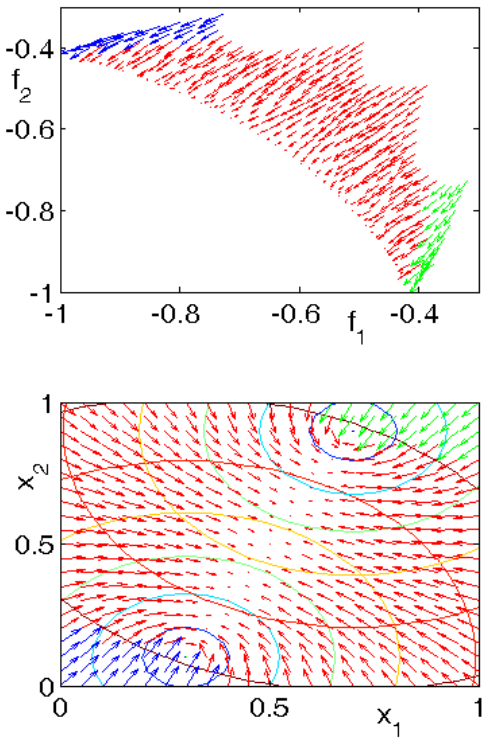

Figure 2 The objective (top) and parameter (bottom) convergence for a non-convex optimization problem. The plots show the negative multi-objective gradients. Red arrows involve both objectives, whereas blue and green arrows only involve a single objective in the calculation.

B. Descent Trajectory Properties

Theorem 1. As a descent trajectory approaches the local Pareto set, it is perpendicular to the tangent of the local Pareto set at the accumulation point.

Proof. It is assumed that the steepest descent calculation has two or more non-zero Lagrange multipliers, otherwise the local Pareto set is simply a point and the concept of a tangent is not defined. In this case, the descent trajectory is the same as for single objective optimization.

By (12), the negative gradient simplex (involves two or more non-zero Lagrange multipliers), is orthogonal to multi-objective gradient at that point.

$$\mathbf{g} \perp \mathbf{V} \qquad (15)$$

In the limit, as the point approaches the local Pareto set, the negative gradient simplex is tangentional to the local Pareto set. The definition of local Pareto optimality is;

$$\mathbf{J}\boldsymbol{\lambda} = \mathbf{0}$$

which also corresponds to the negative gradient simplex subspace. Combining these two results proves the theorem. QED.

Corollary 2. At any point on the boundary of a local Pareto set, the tangent's normal specifies a local change in objectives $\Delta\mathbf{f}\ \alpha$ **-1.**

Proof. Simply by combining (15) with Theorem 1. QED

These results demonstrate the 2-norm steepest descent algorithm locally updates a point towards to the "closest" point in the Pareto set, even though any point in the complete Pareto set could be seen as a potential accumulation point. This "perpendicular closeness" in parameter space ensures that all objectives are, in general, reduced by the same amount.

Theorem 3. Every point on the boundary of the local Pareto set is reachable.

Proof The dimension of the Pareto set is min(n,m-1), therefore the dimension of the subspace normal to the Pareto set is n-min(n,m-1). In the proof below it is assumed that $n<m$-1. Let **s** be a direction in the null space, then locally any point $\mathbf{x}^*+\mu\mathbf{s}$ will converge to $\mathbf{x}^*$. Therefore, the point is reachable by the trajectories that converge to that point.

Consider now when $n\geq m$-1, so the local Pareto set is an n-dimensional subspace. As the descent trajectories are continuous, points in the interior of the Pareto set will not be reachable. However, for any point lying on the Pareto set's boundary and by considering the one-sided normal **s** to that tangent at that point, it is locally reachable by the $\mathbf{x}^*+\mu\mathbf{s}$. QED

These results can be combined to provide some obvious results such as the fact that for each local Pareto set, there is a unique basin of attraction from which every initial point will descent onto that local Pareto set, and the basins of attraction form a mutually exclusive coverage of the parameter space **X**.

As with conventional single objective steepest descent methods, the condition of the underlying optimization problem determines the overall rate of convergence. Some simple results for the convergence of multi-objective steepest descent algorithms are now given.

Theorem 4. When each objective is quadratic and the Hessian matrices have the same set of eigenvectors, then the rate of convergence is dependent on:

$$C(<\mathbf{H}_i>)=\frac{\max_{ij}\sigma_j^i}{\min_{ij}\sigma_j^i}$$

where σ_j^i is the j^{th} eigenvalue of the i^{th} Hessian matrix $\mathbf{H}_i$.

Proof The dynamic scalarization that occurs along the descent trajectory ensures that the (static) learning rate, μ, should be selected to stabilize the fastest mode (largest eigenvalue) of the set of single objective optimization problems. However, the trajectory path may include sections that are dominated by the objective with the smallest eigenvalue (slowest convergence). The worst case rate of parameter convergence will therefore be determined by the ratio of these two eigenvalues and will be determined by the condition of the multi-objective problem (max/min eigenvalue). QED

In practice, the convergence properties along a particular trajectory may be significantly better than this worst case performance. This could be due to several factors, such as the dynamic scalarization weights not changing significantly along the trajectory, or the local Pareto set lying along the eigenvector with the smallest eigenvalue. However, the worst case performance is at least as bad as any of the individual Hessians.

C. Convergence of the 1 and ∞–norm Descent Trajectories

The descent trajectories for the 1 and ∞-norm steepest descent algorithms for the 2 parameter/objective problem described in Section III.A are illustrated in FFigure 3. It can be seen that close to the Pareto set, the descent trajectories of all three steepest descent algorithms are similar, however, further away from the Pareto set, the descent trajectories in FFigure 3 have a very distinct structure. For the 1-norm descent trajectories, the updates often parallel to the coordinate axes while for the ∞-norm trajectories, the updates are often at 45° to the coordinate axes. These observations are proved in this section, in particular it is shown that, close to the Pareto set, all three steepest descent algorithms have the same descent trajectories, and hence the convergence results shown in Section III.B apply to all these methods and more generally to any p-norm where $1\leq p\leq\infty$.

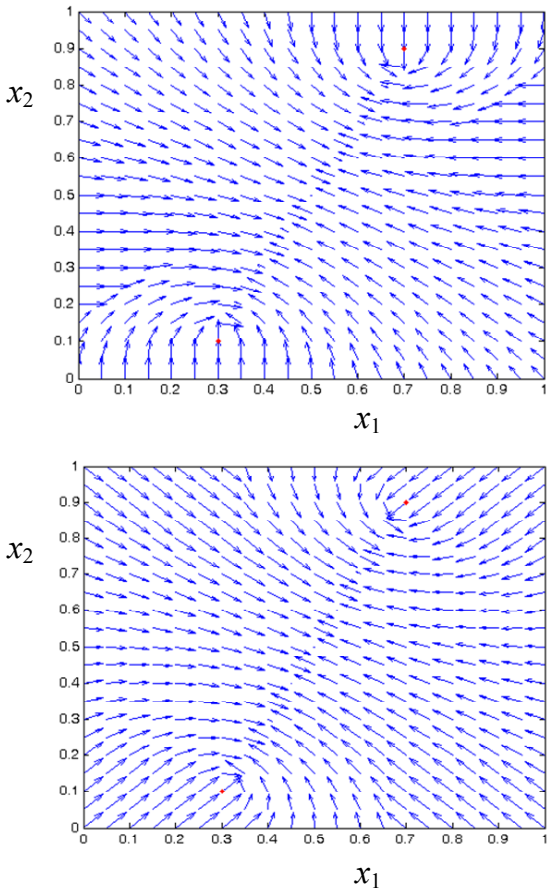

Figure 3 Convergence in parameter space for the 1-norm (top) and ∞-norm (bottom) steepest descent algorithms.

Theorem 5 As a descent trajectory approaches the local Pareto set, it is perpendicular to the tangent to the local Pareto set for any *p*-norm steepest descent method.

Proof Let **x** be a point in parameter space which is close to the local Pareto set, $\boldsymbol{\lambda}$ be the corresponding set of Lagrange multipliers, and **s** be the corresponding optimal update direction for the steepest descent 2-norm optimization criterion. As **x** is close to the local Pareto set, $\boldsymbol{\lambda}>\mathbf{0}$, all the objectives are reduced by the same amount

$$\mathbf{J}^T\mathbf{s} = \mathbf{1}\alpha$$

and $\mathbf{s} \perp \mathbf{V}$, so **V** represents the local tangent to the 2-norm ball at that point. QED

Consider using any *p*-norm steepest descent algorithm, where $p \in [1,\infty)$, and consider the solution at the point **s**. The set of all the *p*-norm local contours at that point generates two cones $\mathbf{P}_1$ and $\mathbf{P}_2$ which have as extrema the 1 and ∞-nom contours, as illustrated in Figure 4. Now consider a local update $\Delta\mathbf{s}$ which lies in either cone along one p-norm tangent. To show that this point is optimal for all *p*-norms, it is necessary to show that for each potential update. $\Delta\mathbf{s}$ along the corresponding contour, at least one objective will increase and hence α is larger.

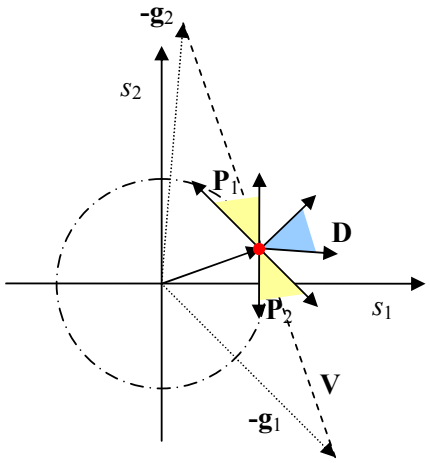

Figure 4 An illustration of the *p*-norm cones for a point close to the local Pareto set. The intersection with the descent cone, D, is zero.

As **x** is close to the Pareto set, the descent cone D is narrow and contains **s**. Therefore, in the limit, $\mathbf{D} \perp \mathbf{V}$. In addition, $\mathbf{V} \subset \mathbf{P}_1 \cup \mathbf{P}_2$ and the maximum angle between **V** and an edge of either of the cones is 45^o. Therefore $\mathbf{D} \cap (\mathbf{P}_1 \cup \mathbf{P}_2) = \varnothing$ and there does not exist an update to **s** which simultaneously reduces both objectives, α, while keeping the step size constant, for any *p*-norm.

Based on this observation, Corollary 2 and Theorem 3 also apply to any p-norm steepest descent algorithm.

Theorem 6 When the descent trajectory is far from the local Pareto set, the descent trajectory moves parallel or 45^o to the coordinate axes for the 1-norm and ∞-norm steepest descent algorithms, respectively.

Proof When a point is far from the Pareto set it can be assumed that the objectives' gradient are aligned, in general, and the descent cone **D**, approaches a half-space. By definition, any update in this cone will cause the objectives to be reduced and if the *p*-norm's tangent/contour passing through **s** lies in **D**, then the update is not optimal because the objectives can be simultaneously reduced without changing the norm's value. It therefore remains to show that, in general, the tangent to a *p*-norm's contour at **s** will lie in **D**. The orientation of **D** depends on the orientation of the individual gradients, although as they are aligned, **D** is approximately a half-space. When **D** is a half-space, the boundary corresponds to the tangent to the 2-norm contour. Therefore, one side of the tangent to any other *p*-norm contour must lie in **D**. QED

This is illustrated in Figure 5 where the descent cone approaches a half-space and it has a non-zero intersection with both $\mathbf{P}_1$ and $\mathbf{P}_2$. In the former, the update **s** is not optimal with respect to an ∞-norm steepest descent algorithm as s_2 can be made larger which does not change the norm's value but does reduce all objectives simultaneously. Similarly, in $\mathbf{P}_2$ the update **s** is not optimal with respect to a 1-norm steepest descent algorithm as s_1 can be increased while s_2 is reduced, not changing the norm's value, which simultaneously reduces all the objectives.

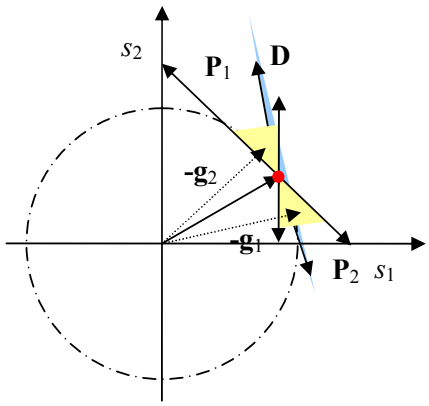

Figure 5 When a point is far from the local Pareto set, the descent cone, D, has a non-zero intersection with either $\mathbf{P}_1$ and/or $\mathbf{P}_2$, in general.

D. Orthogonal convergence in the objective space

All of the min/max, steepest descent learning rules have the property that close to the Pareto set, the descent trajectory is perpendicular to the Pareto set's local tangent. The objectives are all decreased by an equal amount. While this is desirable in some cases, it amounts to a strong prior on the type of solution that is preferred. Another prior would be to desire the convergence in objective space to be normal to the tangent to local Pareto front. This would embody the idea of converging to the "closest point" in objective space.

Theorem 7 When a point is updated according to

$$\mathbf{s} = (\mathbf{J}\mathbf{J}^T)^{-1}\mathbf{J}\boldsymbol{\lambda}$$

where $\boldsymbol{\lambda}$ solves the dual steepest descent problem (12), the descent trajectory in objective space will be locally orthogonal to the local Pareto front.

Proof At a point on the local Pareto front, $\boldsymbol{\lambda}$ represents the normal to the local tangent. As all change in objectives is equal to

$$\Delta\mathbf{f} = \mathbf{J}^T\mathbf{s} = \boldsymbol{\lambda}$$

and this will occur when

$$\mathbf{s} = (\mathbf{J}\mathbf{J}^T)^{-1}\mathbf{J}\boldsymbol{\lambda}$$

assuming $\mathbf{J}\mathbf{J}^T$ is full rank. QED

Other researchers have proposed studying convergence and distribution concepts in the objective space [5], and explicitly specifying the objectives associated with the descent trajectories is an important aspect for these steepest descent methods as they implicitly specify the final accumulation point.

IV. DIVERSITY AND CONVERGENCE

This paper has mainly considered multi-objective descent, however a fundamental part of multi-objective optimization theory is concerned with how diversity can be represented and interpreted, where diversity refers to spreading points along the Pareto set or front, rather than descending towards Pareto front. In practice, the diversity can be represented within a more general multi-objective convergence theory, and this section considers how such generalised measures can be created.

A. Local Pareto Set Reachability

Theorem 8 Any point on the local Pareto set is reachable from any initial point within the local neighbourhood of attraction by updating the point when the step direction lies in the negative gradient span.

Proof There exists a unique descent trajectory that will map the initial point onto a point in the Pareto set. The diversity sub-space (negative gradient span) then spans the local Pareto set, so any other point on the local Pareto set is reachable. QED

While this proof may seem a little strange it should be noted that rather than taking the steepest descent update to reach the Pareto set, it is possible to combine descent and diverse updates by selecting an update that lies within the negative gradient simplex. Minimising along any of the vertices will cause the point to minimise the individual objectives whereas moving in the interior of the subspace will jointly minimize the equivalent combination of objectives. This is now considered further.

B. Diversity Convergence

The steepest descent algorithms converge because they maximally reduce each objective at each iteration. However, when the parameter updates are chosen simply to lie in the negative gradient simplex **V**, or even outside it, the idea of parameter convergence must be relaxed, because some objective may increase. To motivate a measure of convergence, consider the half-spaces which are defined by **V** passing through the point $\mathbf{x}_k$. Then the negative multi-objective gradient, $-\mathbf{g}$, and all of the individual negative gradients $-\mathbf{g}_i$ lie in the same half-space as illustrated in Figure 6. Minimizing along $-\mathbf{g}_i$ will mean that a single objective is minimized, and when the aim is to consider any potential update, in the negative gradient simplex for instance, it is not possible simply to measure distance to an accumulation point, as it is not unique, or to ensure that a single/all objective is reduced, as some of the objectives may increase. This section uses the size of the multi-objective gradient as a measure of convergence. It is zero for any point on the local Pareto set and (locally) decreases if the update direction lies in the negative half-space. It should be noted that, in a first order sense, an update that lies in the negative gradient cone will locally dominate any other update that lies outside it and these updates span the space of converging to the local Pareto set. However, when a more general set of methods for constructing the search direction are considered (second order, …), the descent half-space may still provide a useful tool for analysing convergence.

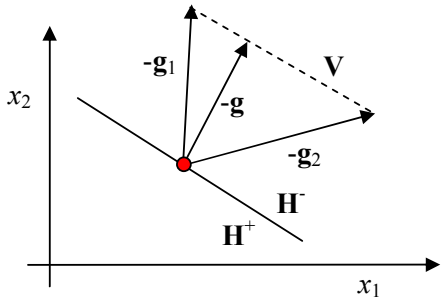

Figure 6 The descent, $\mathbf{H}^-$, and ascent half-spaces, $\mathbf{H}^+$, defined by the multi-objective gradient, g, which is normal to the separating plane.

Theorem 9. When the parameter update lies in the negative gradient half-space, the iterative optimization algorithm converges to the local Pareto set as the size of the multi-objective gradient tends to zero.

Proof Using a first order Taylor series for the i^{th} objective, the new gradient is related to the original gradient via:

$$\mathbf{g}_i(\mathbf{x}_{k+1}) = \mathbf{g}_i(\mathbf{x}_k) + \mu\mathbf{H}_i(\mathbf{x}_k)\mathbf{s}(\mathbf{x}_k)$$

when the step size is small, so second order terms can be neglected. Now consider the weighted size of the multi-objective gradient, given by:

$$\|\mathbf{g}\|^2_{2,\mathbf{H}^{-1}} = \mathbf{g}^T\mathbf{H}^{-1}\mathbf{g}$$

where $\mathbf{H} = \sum_i \lambda_i \mathbf{H}_i$ and is assumed to be positive definite. The use of the inverse scalarized Hessian to measure the size of the multi-objective gradient normalizes the locally quadratic performance function. After updating, the size of the new multi-objective gradient is given by:

$$\begin{aligned}\left\|\mathbf{g}(\mathbf{x}_{k+1})\right\|^2_{2,\mathbf{H}^{-1}} &\le \sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\mathbf{g}_i^T(\mathbf{x}_{k+1})\mathbf{H}^{-1}(\mathbf{x}_k)\mathbf{g}_j(\mathbf{x}_{k+1})\\ &\le \sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j(\mathbf{g}_i^T(\mathbf{x}_k)\mathbf{H}^{-1}(\mathbf{x}_k)\mathbf{g}_j(\mathbf{x}_k)+\\ &\quad 2\mu\mathbf{g}_i^T\mathbf{H}^{-1}(\mathbf{x}_k)\mathbf{H}_j(\mathbf{x}_k)\mathbf{s}(\mathbf{x}_k))\\ &= \left(\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j\mathbf{g}_i^T(\mathbf{x}_k)\mathbf{H}^{-1}\mathbf{g}_j(\mathbf{x}_k)\right)+\\ &\quad 2\mu\left(\sum\nolimits_{i=1}^{m}\lambda_i\mathbf{g}_i^T\mathbf{H}^{-1}(\mathbf{x}_k)\sum\nolimits_{j=1}^{m}\lambda_j\mathbf{H}_j(\mathbf{x}_k)\right)\mathbf{s}(\mathbf{x}_k)\\ &= \left\|\mathbf{g}(\mathbf{x}_k)\right\|^2_{2,\mathbf{H}^{-1}}+2\mu\mathbf{g}(\mathbf{x}_k)^T\mathbf{s}(\mathbf{x}_k)\end{aligned}$$

Therefore, when **s** is orthogonal to $\mathbf{g}(\mathbf{x}_k)$, the size of the MOG is locally unchanged. When the dot product is negative, it decreases and when it is positive, it increases, so when the search direction lies in the negative gradient half-space the MOG's decreases and convergence to the local Pareto set will occur as long as the angle between **s** and the negative gradient half-space boundary does not tend to zero. It has been assumed that the search direction is oriented away from the separating hyperplane by a small, but fixed amount during the complete trajectory. QED

The proof has also assumed that **H** is positive definite. This is similar to the assumption made when many first and second order single objective algorithms are analysed. In practice, this can be relaxed.

The importance of using the size of the multi-objective gradient for analysing the convergence is an important concept from this work. As long as an update lies in the negative gradient half-space, convergence to the local Pareto set should result. This allows a wide range of diverse learning rules and second order learning algorithms to be derived, and this will be addressed in future work.

Updating points in this manner moves the analysis of MOO convergence from one based on descent/ascent/diversity cones (or negative efficiency preserving regions/rules [9]) to one based around local convergence half-spaces. The problem of a MOO algorithm producing not negative efficiency preserving sequences where after two or more diverse updates, the new point has increased all the objectives relative to the starting point [10] cannot occur.

As a final point, as the negative gradient simplex always lies in the negative gradient half-space, any first-order learning algorithms (descent or diversity-based) should always ensure that the updates lie in the negative gradient simplex, as such updates dominate any other updates in a first-order sense.

V. CONCLUSIONS

This paper has considered the update trajectories associated with a class of first order, multi-objective descent algorithms. In general, they aim to maximally reduce the objectives equally (dynamic min/max scalarization) and this causes the trajectory to approach the local Pareto set perpendicularly. This is true for any of the p-norm descent algorithms. The rate of convergence depends on the eigenvalue span of the complete set of eigenvalues for all the objectives which is always larger than any single objective. However, the most general result is using the concept of the multi-objective gradient to define a convergence half-space which as long as the update direction lies in this half-space, convergence to some point on the local Pareto set is assured. This forms the basis for current work as well as investigating how second order gradient methods, which lie in the convergence half-space, can be developed to improve the rate of convergence. In addition, issues associated with sampling a local population of points to obtain relevant gradient information and introducing diversity into the problem statement [3] are being addressed.

VI. REFERENCES

1. Bosman, P.A.N, De Jong, E.D. (2005) Exploiting Gradient Information in Numerical Multi-Objective Evolutionary Optimization, GECCO '05, 755-762
2. Brown M., Smith, R.E, (2003) Effective Use of Directional Information in Multi-Objective Evolutionary Computation, GECCO '03, 778-789.
3. Brown M., Smith R.E. (2005) Directed Multi-Objective Optimization, Int J. Computers, Signals and Systems, Vol. 6(1):3-17.
4. Coello, C.A. Lamont, G.B. (2005) Applications of Multi-Objective Evolutionary Algorithms, World Scientific Publishing
5. Das, I., Dennis J. (1998) Normal-Boundary Intersection: An Alterative Method for Generating Pareto Optimal Points in Multicriteria Optimization Problems, SIAM Journal on Optimization 8:631-657.
6. Deb K., (2001) Mutli-Objective Optimization using Evolutionary Algorithms, Wiley Interscience Series in Systems and Optimization
7. Fletcher R. (2000) "Practical Methods of Optimization", John Wiley and Sons,
8. Fliege, J., Svaiter, B.F. (2000) Steepest Descent Methods for Multicriteria Optimization. Mathematical Methods of Operations Research. Vol. 51(3):479-494
9. Hanne T., (1999) On the convergence of Multiobjective Evolutionary Algorithms, European Journal of Operational Research, 117(3):553-564
10. Knowles J. (2002) Local Search and Hybrid Evolutionary Algorithms for Pareto Optimization, PhD Thesis, Department of Computer Science, University of Reading, UK.
11. Laumanns, M., Thiele, L., Deb, K. and Zitzler E. (2002) Combining Convergence and Diversity in Evolutionary Multiobjective Optimization, Evolutionary Computation, Vol. 10(3):263-282.
12. Marler, R.T. Arora, J.S. (2005) Function-transformation methods for multi-objective optimization, Engineering Optimization, Vol. 37(6):551-570.
13. Mittanen, K. (1999) Nonlinear Multicriterion Optimization, Kluwer Academic Publishers, Boston.
14. Rudolph G. and Agapie A. Convergence properties of some multi-objective evolutionary algorithms. Proc of the 2000 Congress on Evolutionary Computation, 2:1010-1016