Optimization Methods and Software

ISSN: 1055-6788 (Print) 1029-4937 (Online) Journal homepage: https://www.tandfonline.com/loi/goms20

Complexity of gradient descent for multiobjective optimization

J. Fliege, A. I. F. Vaz & L. N. Vicente

To cite this article: J. Fliege, A. I. F. Vaz & L. N. Vicente (2019) Complexity of gradient descent for multiobjective optimization, Optimization Methods and Software, 34:5, 949-959, DOI: 10.1080/10556788.2018.1510928

To link to this article: https://doi.org/10.1080/10556788.2018.1510928

OPTIMIZATION METHODS & SOFTWARE
2019, VOL. 34, NO. 5, 949–959
https://doi.org/10.1080/10556788.2018.1510928

Taylor & Francis
Taylor & Francis Group

Complexity of gradient descent for multiobjective optimization

J. Fliege[a], A. I. F. Vaz[b] and L. N. Vicente[c]

[a]School of Mathematical Sciences, University of Southampton, Southampton, UK; [b]ALGORITMI Research Center, University of Minho, Braga, Portugal; [c]CMUC, Department of Mathematics, University of Coimbra, Coimbra, Portugal

ABSTRACT
A number of first-order methods have been proposed for smooth multiobjective optimization for which some form of convergence to first-order criticality has been proved. Such convergence is global in the sense of being independent of the starting point. In this paper, we analyse the rate of convergence of gradient descent for smooth unconstrained multiobjective optimization, and we do it for non-convex, convex, and strongly convex vector functions. These global rates are shown to be the same as for gradient descent in single-objective optimization and correspond to appropriate worst-case complexity bounds. In the convex cases, the rates are given for implicit scalarizations of the problem vector function.

ARTICLE HISTORY
Received 17 April 2018
Accepted 2 August 2018

KEYWORDS
Multiobjective optimization; gradient descent; steepest descent; global rates; worst-case complexity

1. Introduction

Let us consider an unconstrained multiobjective optimization problem written in the form

$$\min_{x\in\mathbb{R}^n} F(x) \equiv (f_1(x), \ldots, f_m(x)),$$

where each objective function $f_i : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and with a gradient Lipschitz continuous with constant $L_i > 0$, $i = 1, \ldots, m$.

A number of descent methods have been developed and analysed for smooth multiobjective optimization (see Fukuda and Graña Drummond [9]). Steepest descent or gradient methods for multiobjective optimization (see Fliege and Svaiter [8]) converge globally (i.e. independently of the starting point) to a critical Pareto point. In gradient-based dynamic approaches (Attouch and Goudou [1]), the steepest descent method can be recovered by an appropriate time discretization of a system of differential equations whose solution converges to a Pareto point. Other first-order globally convergent algorithms include proximal algorithms (see e.g. Bonnel, Iusem, and Svaiter [4]), trust-region methods (Carrizo, Lotito, and Maciel [5]), and several conjugate gradient methods (Pérez and Prudente [14]). Newton's method for multiobjective optimization was proposed and analysed by Fliege, Graña Drummond, and Svaiter [7].

CONTACT L.N. Vicente lnv@mat.uc.pt

Perhaps the simplest gradient method for MOO takes the form $x^{k+1} = x^k + t^k d^k$, where $t^k > 0$ is the stepsize and where the search direction d^k is obtained from solving (see [8])

$$d^k \;=\; \arg\min_{d\in\mathbb{R}^n} \left\{ \max_{i\in\{1,\ldots,m\}} \nabla f_i(x_k)^\top d + \frac{1}{2}\|d\|^2 \right\}. \tag{1}$$

When $m = 1$, one retrieves the steepest descent direction $d^k = -\nabla f_1(x_k)$.

For single-objective optimization ($m = 1$), it is well known (see the book by Nesterov [16]) that the steepest descent or gradient method decreases the gradient to zero at the rate of $1/\sqrt{k}$ regardless of the starting point. Moreover, the corresponding worst case bound in the number of iterations needed to achieve a gradient of norm smaller than $\epsilon \in (0, 1)$, which is of the order of ϵ^{-2}, was proven to be sharp or tight, in the sense that there exists an example for $n = 1$ (Cartis, Gould, and Toint [6]), dependent on an arbitrarily small parameter $\tau > 0$, for which such a number is of the order of $\epsilon^{-2+\tau}$. The *global* rate $1/\sqrt{k}$ is shared by many first-order methods which impose a sufficient decrease condition, like trust-region methods (using gradient information) [6] and direct-search methods for derivative-free optimization [17]. Such a global rate is improved to $1/k$ in the convex case (Nesterov [16]), and higher-order methods deliver a rate that tends precisely to $1/k$ when the order tends to infinity (Birgin et al. [3]). Finally, it is also well known that gradient descent exhibits a linear rate of convergence in the strongly convex case (Nesterov [16]).

The goal of this paper is to extend this theory to multiobjective optimization. We will see without too much surprise that the same rates of the single-objective setting are attainable, although the convex and strongly convex cases unveil interesting questions. The rate of $1/\sqrt{k}$ in the non-convex case does not raise any issue as it is derived for a measure of first-order criticality. In the convex cases, however, as the rates are derived for function values it is not foreseeable, before a careful analysis, what gap or error should be quantified.

It was brought to our attention by a Referee of this paper that Grapiglia in his PhD Thesis [10, Section 3.5.2] had proved the same worst-case complexity bound of the order of ϵ^{-2} for smooth unconstrained multiobjective optimization but using instead trust-region methods. This was later reported in the paper [12, Section 4.2]. Grapiglia made also a similar derivation using line-search methods in an unpublished work [11]. The results of our paper are derived in a more concise way and cover the convex and strongly convex cases.

2. Pareto criticality

Let $\nabla F(x)$ denote the transpose of the Jacobian matrix of the vector-valued objective function F. A necessary condition for a point $x \in \mathbb{R}^n$ to be a (local) weak Pareto minimizer is

$$\text{range}\left(\nabla F(x)^\top\right) \cap (-\mathbb{R}_{++})^m \;=\; \emptyset, \tag{2}$$

where $\mathbb{R}_{++}$ is the set of (strictly) positive real numbers. Points that satisfy condition (2) are then called (first-order) Pareto critical points. If a point x is not Pareto critical, then there exists a direction $d \in \mathbb{R}^n$ such that

$$\nabla F(x)^\top d \;\in\; (-\mathbb{R}_{++})^m,$$

i.e. d is a descent direction for F at the point x. This motivates the first term of the objective function in subproblem (1).

Now note that subproblem (1) can be rewritten equivalently as the following (differentiable) quadratic optimization problem

$$(d^k, \alpha^k) = \arg \min_{d \in \mathbb{R}^n, \alpha \in \mathbb{R}} \ \alpha + \frac{1}{2}\|d\|^2 \ \equiv \ q(d, \alpha)$$
$$\text{subject to } \nabla f_i(x^k)^\top d \leq \alpha, \quad i = 1, \ldots, m. \tag{3}$$

From the KKT conditions of problem (3) we have

$$d^k \ = \ -\sum_{i=1}^{m} \lambda_i^k \nabla f_i(x^k), \tag{4}$$

where $\lambda_i^k \geq 0$ are the Lagrange multipliers associated with the linear inequality constraints in (3), and

$$\sum_{i=1}^{m} \lambda_i^k \ = \ 1. \tag{5}$$

The solution of subproblem (3) is intimately related to Pareto criticality as stated in the following result from [8] (a proof is included for completeness).

Lemma 2.1 ([8, Lemma 1]): *Let (d^k, α^k) be the solution of problem (3).*

(1) *If x^k is Pareto critical, then $d^k = 0 \in \mathbb{R}^n$ and $\alpha^k = 0$.*
(2) *If x^k is not Pareto critical, then*

$$\alpha^k \leq -(1/2)\|d^k\|^2 \ < \ 0, \tag{6}$$

$$\nabla f_i(x^k)^\top d^k \ \leq \ \alpha^k, \ \ i = 1, \ldots, m. \tag{7}$$

Proof: If x^k is Pareto critical, then there is no d such that $\nabla f_i(x^k)^\top d < 0, \forall i \in \{1, \ldots, m\}$, otherwise condition (2) would not be satisfied, leading to the existence of an $\bar{i}$ such that $\alpha^k \geq \nabla f_{\bar{i}}(x^k)^\top d^k \geq 0$. Item 1 follows then by noting that $(d, \alpha) = (0, 0)$ is a feasible point of subproblem (3).

As for Item 2, if x^k is not Pareto critical, then there exists a d such that $\nabla f_i(x^k)^\top d < 0, \forall i$, resulting in $\alpha^k < 0$. Equation (7) follows directly from the constraints of subproblem (3). Since $(d, \alpha) = (0, 0)$ is a feasible point, one has $q(d^k, \alpha^k) \leq q(0, 0) = 0$, hence (6), where $q(\cdot, \cdot)$ has been defined in (3). ■

3. Gradient descent in the non-convex case

In this section we will analyse the gradient method described in Algorithm 1 (see [8]). At each step, the *steepest descent direction* d^k is first computed by solving (1) or equivalently (3). Then a backtracking procedure along d^k is applied which stops when a classical

Algorithm 1 MO gradient descent

1: Choose $\beta \in (0,1)$ and $x^0 \in \mathbb{R}^n$. Set $k := 0$.
2: Compute d^k by solving the subproblem (3).
3: Stop if x^k is Pareto critical.
4: Compute a stepsize $t^k \in (0,1]$ as the maximum of

$$T^k := \left\{ t = \frac{1}{2^j} | \, j \in \mathbb{N}_0, F(x^k + td^k) \leq F(x^k) + \beta t \nabla F(x^k)^\top d^k \right\}. \tag{8}$$

5: Set $x^{k+1} := x^k + t^k d^k$, $k := k+1$, and goto Step 2.

sufficient decrease condition is satisfied; see (8). Each backtracking starts at $t = 1$ and halves the stepsize until it finds one for which all functions have decreased sufficiently.

We start by showing the existence of a uniform lower bound on the stepsize t^k that will be used later in the analysis. Such a lower bound also shows that Step 4 of Algorithm 1 will always stop in a finite number of steps. The argument is a classic one in line-search methods.

Lemma 3.1: *In Algorithm* 1 *the stepsize always satisfies* $t^k \geq t_{\min} \equiv \min\{(1-\beta)/(2L_{\max}), 1\}$ *where* $L_{\max} = \max\{L_1, \ldots, L_m\}$ *(with* L_i *the Lipschitz constant of* ∇f_i, $i = 1, \ldots, m$*) and* $\beta \in (0,1)$ *is the parameter of the sufficient decrease condition* (8).

Proof: When $2t$ does not satisfy the sufficient decrease condition (8) of Algorithm 1, there exists an index $i \in \{1, \ldots, m\}$ such that

$$f_i(x^k + (2t)d^k) - f_i(x^k) \; > \; \beta(2t)\nabla f_i(x^k)^\top d^k. \tag{9}$$

Due to Lipschitz continuity we have,

$$f_i(x^k + (2t)d^k) - f_i(x^k) \; \leq \; (2t)\nabla f_i(x^k)^\top d^k + \frac{L_i}{2}\|(2t)d^k\|^2. \tag{10}$$

By combining (9) and (10) one obtains

$$0 \; < \; (2t)(1-\beta)\nabla f_i(x^k)^\top d^k + 2L_i t^2 \|d^k\|^2$$

which using (6)–(7) then implies

$$-L_i t \|d^k\|^2 \; < \; (1-\beta)\nabla f_i(x^k)^{\mathrm{T}} d^k \; < \; -\frac{(1-\beta)}{2}\|d^k\|^2,$$

establishing that

$$t \; > \; \frac{1-\beta}{2L_i}.$$

The result follows by noting that t is never larger than one and that $L_{\max} = \max\{L_1, \ldots, L_m\}$. ■

Is is then easy to prove that Algorithm 1 has a convergence rate of the order of $1/\sqrt{k}$.

Theorem 3.1: *Suppose that at least one of the functions $f_1, \ldots, f_m$ is bounded from below. Let $f_i^{\min}$ be the lower bound on the function f_i when bounded from below. For those indices i, let $F^{\min}$ be the minimum of the lower bounds $f_i^{\min}$ and let $F_0^{\max}$ be the maximum of the values $f_i(x_0)$.*

The gradient method described in Algorithm 1 generates a sequence $\{x^k\}$ such that

$$\min_{0\leq \ell \leq k-1} \|d^\ell\| \;\leq\; \sqrt{\frac{F_0^{\max} - F^{\min}}{M}}\,\frac{1}{\sqrt{k}},$$

where $M = (\beta t_{\min})/2$ and $t_{\min}$ is given in Lemma 3.1.

Proof: Let i be an index of a function f_i bounded from below. From the sufficient decrease condition (8) and the properties (6)–(7) of the direction d^k,

$$f_i(x^k + t^k d^k) - f_i(x^k) \;\leq\; \beta t^k \nabla f_i(x^k)^\top d^k \;\leq\; -\beta \frac{t^k}{2}\|d^k\|^2,$$

and then from Lemma 3.1,

$$f_i(x^k) - f_i(x^k + t^k d^k) \;\geq\; \frac{\beta t^k}{2}\|d^k\|^2 \;\geq\; \frac{\beta t_{\min}}{2}\|d^k\|^2 \;\equiv\; M\|d^k\|^2.$$

Summing up all decreases until iteration $k-1$, yields

$$\begin{aligned} f_i(x^0) - f_i(x^{k-1} + t^{k-1}d^{k-1}) &= \sum_{\ell=0}^{k-1} f_i(x^\ell) - f_i(x^\ell + t^\ell d^\ell) \\ &\geq M \sum_{\ell=0}^{k-1} \|d^\ell\|^2 \;\geq M(k) \left(\min_{0\leq \ell \leq k-1} \|d^\ell\|\right)^2, \end{aligned}$$

and the proof is concluded from the definitions of $F_0^{\max}$ and $F^{\min}$. ■

4. The convex and strongly convex cases

In single-objective optimization when the function is convex, the analysis of the gradient method is typically carried out for a fixed stepsize, inversely proportional to the Lipschitz constant of the gradient of the objective function. It is also known that it can be alternatively imposed a sufficient decrease condition, different from the traditional one used in the non-convex case. We restate in Algorithm 2 the gradient method for multiobjective optimization using such an alternative sufficient decrease condition (11).

It is also known that a lower bound on the stepsize can be obtained when imposing this alternative sufficient decrease condition.

Lemma 4.1: *In Algorithm 2 the stepsize always satisfies $t^k \geq t_{\min} \equiv \min\{\gamma/(2L_{\max}), 1\}$ where $L_{\max} = \max\{L_1, \ldots, L_m\}$.*

Algorithm 2 MO gradient descent (convex case)

1: Choose $\gamma \in (0,1)$ and $x^0 \in \mathbb{R}^n$. Set $k := 0$.
2: Compute d^k by solving the subproblem (3).
3: Stop if x^k is Pareto critical.
4: Compute a stepsize $t^k \in (0,1]$ as the maximum of

$$T^k := \left\{ t = \frac{1}{2^j} | \, j \in \mathbb{N}_0, F(x^k + td^k) \leq F(x^k) + t\nabla F(x^k)^\top d^k + \frac{\gamma\, t}{2}\|d^k\|^2 e \right\}, \quad (11)$$

where e is the vector of ones in $\mathbb{R}^m$.
5: Set $x^{k+1} := x^k + t^k d^k$, $k := k+1$, and goto Step 2.

Proof: By using the Lipschitz continuity of ∇f_i, one can easily see that, for all $t \in (0, (\gamma/L_i)]$,

$$\begin{aligned} f_i(x^k + td^k) &\leq f_i(x^k) + t\nabla f_i(x^k)^\top d^k + \frac{L_i}{2}\|td^k\|^2 \\ &\leq f_i(x^k) + t\nabla f_i(x^k)^\top d^k + \frac{\gamma}{2t}\|td^k\|^2. \end{aligned}$$

Hence the sufficient decrease condition (11) is satisfied for $t \in (0, (\gamma/L_{\max})]$ and the result comes then from the fact that in the backtracking scheme the stepsize starts at one and is halved each time. ■

Notice that when imposing (11) one obtains from (6)–(7), for all $i \in \{1, \ldots, m\}$,

$$\begin{aligned} f_i(x^k + t^k d^k) &\leq f_i(x^k) - \frac{t^k}{2}\|d^k\|^2 + \frac{\gamma t^k}{2}\|d^k\|^2 \\ &= f_i(x^k) - \frac{1-\gamma}{2} t^k \|d^k\|^2. \end{aligned}$$

Using the lower bound on t^k from Lemma 4.1 we thus obtain a decrease that leads to a global rate of $1/\sqrt{k}$ for $\min_{0 \leq l \leq k-1} \|d^\ell\|$ as in Theorem 3.1. This then proves that $\liminf_{k\to\infty} \|d^k\| = 0$. Then, if $L(x^0) = \{x \in \mathbb{R}^n : F(x) \leq F(x^0)\}$ is bounded, the sequence $\{x^k\}$ has a limit point x^* that is Pareto critical. As the multipliers λ^k lie in a bounded set, one can also say, without loss of generality, that the subsequence K for which x^k converges to x^* is such that λ^k converges to a λ^* such that

$$\sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) = 0, \quad \sum_{i=1}^m \lambda_i^* = 1, \quad \lambda_i^* \geq 0, \; i = 1, \ldots, m. \quad (12)$$

In the derivation of the global rates for the convex cases we will make the slightly stronger assumption that the whole sequence (x^k, λ^k) converges to (x^*, λ^*). Under the convexity assumption on the objectives f_i, the point x^* is then a weak Pareto point, and if in addition x^* is the unique minimum of the scalar function $\sum_i \lambda_i^* f_i$, then x^* is a Pareto point (see Theorem 5.13 and Lemma 5.14 in the book of Jahn [13]).

We will now assume convexity of all components of F. As so, we will make use of the following known inequality

$$f_i(x) \;\leq\; f_i(y) + \nabla f_i(x)^\top (x-y) - \frac{\mu_i}{2}\|x-y\|^2, \tag{13}$$

valid for all x,y either when f_i is convex and μ_i is set to zero or when f_i is strongly convex with modulus $\mu_i > 0$. We start by an intermediate lemma establishing an upper bound in the same vein of the known case $m=1$. At this point there is no need to make assumptions about the point x^*.

Lemma 4.2: *Assume that $\{x^k\}$ converges to x^*. If f_i is convex or strongly convex of modulus $\mu_i > 0$, $i = 1,\ldots,m$, then*

$$\sum_{i=1}^{m} \lambda_i^k \left(f_i(x^{k+1}) - f_i(x^*)\right) \;\leq\; \frac{1}{2t_{\min}} \left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right) - \frac{\mu}{2}\|x^k - x^*\|^2, \tag{14}$$

where $\mu = \min_{1\leq i\leq m} \mu_i$ and $t_{\min}$ is given in Lemma 4.1.

Proof: Since $\gamma < 1$, one has for all $i = 1,\ldots,m$

$$f_i(x^{k+1}) \;\leq\; f_i(x^k) + \nabla f_i(x^k)^\top (t^k d^k) + \frac{1}{2t^k}\|t^k d^k\|^2. \tag{15}$$

One can now use the convexity ($\mu_i = 0$) / strong convexity ($\mu_i > 0$) of f_i, see (13), to bound $f_i(x^k)$ in (15), obtaining

$$\begin{aligned} f_i(x^{k+1}) \leq f_i(x^*) &+ \nabla f_i(x^k)^\top (x^k - x^*) - \frac{\mu}{2}\|x^k - x^*\|^2 \\ &+ \nabla f_i(x^k)^\top (t^k d^k) + \frac{1}{2t^k}\|t^k d^k\|^2, \quad i = 1,\ldots,m. \end{aligned}$$

Rearranging terms, multiplying by λ_i^k and summing for all $i = 1,\ldots,m$,

$$\begin{aligned} \sum_{i=1}^{m} \lambda_i^k \left(f_i(x^{k+1}) - f_i(x^*)\right) \leq &\left(\sum_{i=1}^{m} \lambda_i^k \nabla f_i(x^k)\right)(x^k - x^* + t^k d^k) \\ &+ \left(\frac{t^k}{2}\|d^k\|^2 - \frac{\mu}{2}\|x^k - x^*\|^2\right) \sum_{i=1}^{m} \lambda_i^k. \end{aligned}$$

From (4) and (5),

$$\begin{aligned}\sum_{i=1}^{m} \lambda_i^k \left(f_i(x^{k+1}) - f_i(x^*)\right) &\leq -(d^k)^\top (x^k - x^* + t^k d^k) + \frac{t^k}{2}\|d^k\|^2 - \frac{\mu}{2}\|x^k - x^*\|^2 \\ &= -\frac{1}{2t^k}\left(2(t^k d^k)^\top (x^k - x^*) + \|t^k d^k\|^2\right) - \frac{\mu}{2}\|x^k - x^*\|^2 \\ &= -\frac{1}{2t^k}\left(\|x^k - x^* + t^k d^k\|^2 - \|x^k - x^*\|^2\right) - \frac{\mu}{2}\|x^k - x^*\|^2 \\ &= \frac{1}{2t^k}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right) - \frac{\mu}{2}\|x^k - x^*\|^2 \\ &\leq \frac{1}{2t_{\min}}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right) - \frac{\mu}{2}\|x^k - x^*\|^2.\end{aligned}$$

The last inequality results from $t^k \geq t_{\min}$ and from the fact that the nonnegativity of the terms $\lambda_i^k[(f_i(x^{k+1}) - f_i(x^*)]$ necessarily implies the nonnegativity of $\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2$. ■

Remark 4.1: If a fixed stepsize with $t^k = \bar{t}$ constant would be used (instead of imposing sufficient decrease), then as long as $0 < \bar{t} \leq 1/L_{\max}$ the result of Lemma 4.2 would still be true without assuming that $\{x_k\}$ converges to x^*. In such a case we would not know whether the left-hand-side in (14) is nonnegative or not, and thus if such a result could be used later to prove an effective rate.

We proceed separating the convex case from the strongly convex one. Next we address the convex case establishing the desired $1/k$ rate for a certain sequence of weights $\{\bar{\lambda}^k\}$ that converges to the optimal weights λ^* in (12) when the multipliers $\{\lambda^k\}$ do so.

Theorem 4.1: *Suppose that the functions $f_1, \ldots, f_m$ are convex. Assume that $\{x^k\}$ converges to x^*.*

The gradient method described in Algorithm 2 generates a sequence $\{x^k\}$ such that

$$\sum_{i=1}^{m} \bar{\lambda}_i^{k-1} f_i(x^k) - \sum_{i=1}^{m} \bar{\lambda}_i^{k-1} f_i(x^*) \;\leq\; \frac{\|x^0 - x^*\|^2}{2t_{\min} k},$$

where the weights $\bar{\lambda}_i^{k-1} \equiv (1/k)\sum_{\ell=0}^{k-1} \lambda_i^\ell$ satisfy

$$\sum_{i=1}^{m} \bar{\lambda}_i^{k-1} \;=\; 1 \quad \text{and} \quad \bar{\lambda}_i^{k-1} \;\geq\; 0, \quad i = 1, \ldots, m.$$

Finally, if $\{\lambda^k\}$ converges to λ^*, then so does $\{\bar{\lambda}^k\}$.

Proof: By summing (14) from $\ell = 0, \ldots, k-1$, and since $f_i(x^\ell) \leq f_i(x^{\ell-1})$ for all i, ℓ, one derives

$$\sum_{\ell=0}^{k-1} \sum_{i=1}^{m} \lambda_i^\ell \left(f_i(x^k) - f_i(x^*) \right) \leq \frac{1}{2t_{\min}} \left(\|x^0 - x^*\|^2 - \|x^k - x^*\|^2 \right)$$
$$\leq \frac{1}{2t_{\min}} \|x^0 - x^*\|^2.$$

Hence

$$\sum_{i=1}^{m} \left(\sum_{\ell=0}^{k-1} \lambda_i^\ell \right) \left(f_i(x^k) - f_i(x^*) \right) \leq \frac{1}{2t_{\min}} \|x^0 - x^*\|^2.$$

The proof is completed dividing both sides of this last inequality by k. ■

Now we show that gradient descent also attains for multiobjective optimization a linear convergence rate in the strongly convex case.

Theorem 4.2: *Suppose that the functions f_i are strongly convex with modulus $\mu_i > 0$, $i = 1, \ldots, m$. Assume that $\{x^k\}$ converges to x^*.*

The gradient method described in Algorithm 2 generates a sequence $\{x^k\}$ such that

$$\|x^k - x^*\| \leq \left(\sqrt{1 - t_{\min}\mu} \right)^k \|x^0 - x^*\|. \tag{16}$$

Proof: We go back to (14) and write

$$\sum_{i=1}^{m} \lambda_i^k \left(f_i(x^{k+1}) - f_i(x^*) \right) \leq \frac{1}{2t_{\min}} \left((1 - t_{\min}\mu) \|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 \right).$$

By noting that the left-hand side is nonnegative,

$$\|x^{k+1} - x^*\|^2 \leq (1 - t_{\min}\mu) \|x^k - x^*\|^2,$$

and the proof is completed applying this last inequality recursively. ■

If the pair (x^*, λ^*) is Pareto critical and the f_i's are convex, $f^* \equiv \sum_{i=1}^m \lambda_i^* f_i$ is a convex function with minimizer at x^*. When all functions f_i are strongly convex, so is f^* with modulus $\mu^* = \min_{\lambda_i^* > 0} \mu_i$. Moreover the function f^* has a gradient that is Lipschitz continuous with constant $L^* = \max_{\lambda_i^* > 0} L_i$. Hence, $f^*(x^k) - f^*(x^*) \leq (L^*/2)\|x^k - x^*\|^2$ (see, e.g. [2, Theorem 5.8]) and from (16) we also derive a linear rate for the optimality gap in f^*,

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^k) - \sum_{i=1}^{m} \lambda_i^* f_i(x^*) \leq \frac{L^*}{2} \|x^k - x^*\|^2 \leq \frac{L^*}{2} (1 - t_{\min}\mu)^k \|x^0 - x^*\|^2.$$

5. Concluding remarks

We derived global rates for gradient descent for smooth multiobjective optimization matching what is known in single-objective optimization, for non-convex ($1/\sqrt{k}$), for convex ($1/k$), and for strongly convex (r^k for some $r \in (0,1)$) vector-valued objective functions. Such global rates translate into worst-case complexity bounds of the order of $1/\epsilon^2$, $1/\epsilon$, and $\log(1/\epsilon)$ iterations, respectively, to reach an approximate optimality criterion of the form $\|d^k\| \leq \epsilon$ for some $\epsilon \in (0,1)$, where d^k is the steepest descent direction (1).

There are a number of aspects to be further investigated, among which are the use of momentum and/or proximal operators (see the recent book [2] by Beck). In particular, proving a global rate of $1/k^2$ for an accelerated gradient method, as Nesterov [15] did for single-objective optimization, is more intricate than it seems at least using the steepest descent direction (1) and the proof technology of our paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Support for A.I.F. Vaz was partially provided by FCT [grant number COMPETE:POCI-01-0145-FEDER-007043], [grant number UID/CEC/00319/2013], and support for L.N. Vicente was partially provided by FCT [grant number UID/MAT/00324/2013], [grant number P2020 SAICTPAC/0011/2015.]

References

[1] H. Attouch and X. Goudou, *A continuous gradient-like dynamical approach to pareto-optimization in Hilbert spaces*, Set-Valued. Var. Anal. 22 (2014), pp. 189–219.
[2] A. Beck, *First-order Methods in Optimization*, MPS-SIAM Series on Optimization; SIAM, Philadelphia, 2017.
[3] E.G. Birgin, J.L. Gardenghi, J.M. Martínez, S.A. Santos and Ph.L. Toint, *Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models*, Math. Program. 163 (2017), pp. 359–368.
[4] H. Bonnel, A.N. Iusem and B.F. Svaiter, *Proximal methods in vector optimization*, SIAM J. Optim. 15 (2005), pp. 953–970.
[5] G.A. Carrizo, P.A. Lotito and M.C. Maciel, *Trust region globalization strategy for the nonconvex unconstrained multiobjective optimization problem*, Math. Program. 159 (2016), pp. 339–369.
[6] C. Cartis, N.I.M. Gould and Ph.L. Toint, *On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization*, SIAM J. Optim. 20 (2010), pp. 2833–2852.
[7] J. Fliege, L.M. Graña Drummond and B.F. Svaiter, *Newton's method for multiobjective optimization*, SIAM J. Optim. 20 (2009), pp. 602–626.
[8] J. Fliege and B.F. Svaiter, *Steepest descent methods for multicriteria optimization*, Math. Methods Oper. Res. 51 (2000), pp. 479–494.
[9] E.H. Fukuda and L.M. Graña Drummond, *A survey on multiobjective descent methods*, Pesquisa Operacional 34 (2014), pp. 585–620.
[10] G.N. Grapiglia, *Três Contribuições em Otimização Não Linear e Não Convexa*, Ph.D. thesis, Universidade Federal do Paraná, 2014. In Portuguese
[11] G.N. Grapiglia, On the worst-case complexity of projected gradient methods for convex constrained multiobjective optimization. 2016.

[12] G.N. Grapiglia, J. Yuan and Y.-X. Yuan, *On the convergence and worst-case complexity of trust-region and regularization methods for unconstrained optimization*, Math. Program. 152 (2015), pp. 491–520.
[13] J. Jahn, *Vector Optimization*, Springer, Berlin, 2009.
[14] L.R. Lucambio Pérez and L.F. Prudente, Non-linear conjugate gradient methods for vector optimization, preprint (2017), Federal University of Goias.
[15] Y. Nesterov, *A method of solving a convex programming problem with convergence rate* $O(1/k^2)$, Soviet Math. Doklady 27 (1983), pp. 372–376.
[16] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Kluwer Academic Publishers, Dordrecht, 2004.
[17] L.N. Vicente, *Worst case complexity of direct search*, EURO J. Comput. Optim. 1 (2013), pp. 143–153.