Global convergence of a BFGS-type algorithm for nonconvex multiobjective optimization problems

L. F. Prudente[*] D. R. Souza[*]

March 18, 2024

Abstract: We propose a modified BFGS algorithm for multiobjective optimization problems with global convergence, even in the absence of convexity assumptions on the objective functions. Furthermore, we establish a local superlinear rate of convergence of the method under usual conditions. Our approach employs Wolfe step sizes and ensures that the Hessian approximations are updated and corrected at each iteration to address the lack of convexity assumption. Numerical results shows that the introduced modifications preserve the practical efficiency of the BFGS method.

Keywords: Multiobjective optimization, Pareto optimality, quasi-Newton methods, BFGS, Wolfe line search, global convergence, rate of convergence.

AMS subject classifications: 49M15, 65K05, 90C29, 90C30, 90C53

1 Introduction

Multiobjective optimization problems involve the simultaneous minimization of multiple objectives that may be conflicting. The goal is to find a set of solutions that offer different trade-offs between these objectives, helping decision makers in identifying the most satisfactory solution. *Pareto optimality* is a fundamental concept used to characterize such solutions. A solution is said to be *Pareto optimal* if none of the objectives can be improved without deterioration to at least one of the other objectives.

Over the last two decades, significant research has focused on extending iterative methods originally developed for single-criterion optimization to the domain of multiobjective optimization, providing an alternative to scalarization methods [19, 41]. This line of research was initiated by Fliege and Svaiter in 2000 with the extension of the steepest descent method [23] (see also [32]). Since then, several methods have been studied, including Newton [12, 22, 28, 31, 51], quasi-Newton [1,33,34,39,42,44,46–48], conjugate gradient [27,29,37], conditional gradient [2,10], projected gradient [3, 20, 24, 25, 30], and proximal methods [5, 8, 9, 11, 13].

Proposed independently by Broyden [6], Fletcher [21], Goldfarb [26], and Shanno [49] in 1970, the BFGS is the most widely used quasi-Newton method for solving unconstrained scalar-valued optimization problems. As a quasi-Newton method, it computes the search direction using a

[*] Instituto de Matemática e Estatística, Universidade Federal de Goiás, CEP 74001-970, Goiânia, GO, Brazil, E-mails: lfprudente@ufg.br, danilo_souza@discente.ufg.br. This work was funded by CAPES and CNPq (Grants 309628/2020-2, 405349/2021-1).

quadratic model of the objective function, where the Hessian is approximated based on first-order information. Powell [45] was the first to prove the global convergence of the BFGS method for convex functions, employing a line search that satisfies the Wolfe conditions. Some time later, Byrd and Nocedal [7] introduced additional tools that simplified the global convergence analysis, enabling the inclusion of backtracking strategies. For over three decades, the convergence of the BFGS method for nonconvex optimization remained an open question until Dai [15], in the early 2000s, provided a counterexample showing that the method can fail in such cases (see also [16,40]). Another research direction focuses on proposing suitable modifications to the BFGS algorithm that enable achieving global convergence for nonconvex general functions while preserving its desirable properties, such as efficiency and simplicity. Notable works in this area include those by Li and Fukushima [35, 36].

The BFGS method for multiobjective optimization was studied in [33, 34, 39, 42, 44, 46–48]. However, it is important to note that, except for [46, 47], the algorithms proposed in these papers are specifically designed for convex problems. The assumption of convexity is crucial to ensure that the Hessian approximations remain positive definite over the iterations, guaranteeing the well-definedness of these methods. In [47], Qu *et al.* proposed a *cautious* BFGS update scheme based on the work [36]. This approach updates the Hessian approximations only when a given safeguard criterion is satisfied, resulting in a globally convergent algorithm for nonconvex problems. In [46], Prudente and Souza proposed a BFGS method with Wolfe line searches which exactly mimics the classical BFGS method for single-criterion optimization. This variant is well defined even for general nonconvex problems, although global convergence cannot be guaranteed in this general case. Despite this, it has been shown to be globally convergent for strongly convex problems.

In the present paper, inspired by the work [35], we go a step further than [46] and introduce a modified BFGS algorithm for multiobjective optimization which possesses a global convergence property even without convexity assumption on the objective functions. Furthermore, we establish the local superlinear convergence of the method under certain conditions. Our approach employs Wolfe step sizes and ensures that the Hessian approximations are updated and corrected at each iteration to overcome the lack of convexity assumption. Numerical results comparing the proposed algorithm with the methods introduced in [46, 47] are discussed. Overall, the modifications made to the BFGS method to ensure global convergence for nonconvex problems do not compromise its practical performance.

The paper is organized as follows: Section 2 presents the concepts and preliminary results, Section 3 introduces the proposed modified BFGS algorithm and discusses its global convergence, Section 4 focuses on the local convergence analysis with superlinear convergence rate, Section 5 presents the numerical experiments, and Section 6 concludes the paper with some remarks. Throughout the main text, we have chosen to omit proofs that can be easily derived from existing literature to enhance overall readability. However, these proofs are provided in the Appendix for self-contained completeness.

Notation. $\mathbb{R}$ and $\mathbb{R}_{++}$ denote the set of real numbers and the set of positive real numbers, respectively. As usual, $\mathbb{R}^n$ and $\mathbb{R}^{n\times p}$ denote the set of n-dimensional real column vectors and the set of $n \times p$ real matrices, respectively. The identity matrix of size n is denoted by I_n. $\|\cdot\|$ is the Euclidean norm. If $u, v \in \mathbb{R}^n$, then $u \preceq v$ (or $\prec$) is to be understood in a componentwise sense, i.e., $u_i \leq v_i$ (or $<$) for all $i = 1, \ldots, n$. For $B \in \mathbb{R}^{n\times n}$, $B \succ 0$ means that B is positive definite. In this case, $\langle\cdot,\cdot\rangle_B$ and $\|\cdot\|_B$ denote the B-energy inner product and the B-energy norm, respectively, i.e., for $u, v \in \mathbb{R}^n$, $\langle u, v\rangle_B := u^\top Bv$ and $\|u\|_B := \sqrt{\langle u, u\rangle_B}$. If $K = \{k_1, k_2, \ldots\} \subseteq \mathbb{N}$, with $k_j < k_{j+1}$ for all $j \in \mathbb{N}$, then we denote $K \underset{\infty}{\subset} \mathbb{N}$.

2 Preliminaries

In this paper, we focus on the problem of finding a *Pareto optimal* point of a continuously differentiable function $F : \mathbb{R}^n \to \mathbb{R}^m$. This problem can be denoted as follows:

$$\min_{x \in \mathbb{R}^n} F(x). \tag{1}$$

A point $x^* \in \mathbb{R}^n$ is *Pareto optimal* (or *weak Pareto optimal*) of F if there is no other point $x \in \mathbb{R}^n$ such that $F(x) \preceq F(x^*)$ and $F(x) \neq F(x^*)$ (or $F(x) \prec F(x^*)$). These concepts can also be defined locally. We say that $x^* \in \mathbb{R}^n$ is a *local Pareto optimal* (or *local weak Pareto optimal*) point if there exists a neighborhood $U \subset \mathbb{R}^n$ of x^* such that x^* is Pareto optimal (or weak Pareto optimal) for F restricted to U. A necessary condition (but not always sufficient) for the local weak Pareto optimality of x^* is given by:

$$-\left(\mathbb{R}^m_{++}\right) \cap \operatorname{Image}(JF(x^*)) = \emptyset, \tag{2}$$

where $JF(x^*)$ denotes the Jacobian of F at x^*. A point x^* that satisfies (2) is referred to as a *Pareto critical* point. It should be noted that if $x \in \mathbb{R}^n$ is not Pareto critical, then there exists a direction $d \in \mathbb{R}^n$ such that $\nabla F_j(x)^\top d < 0$ for all $j = 1, \ldots, m$. This implies that d is a *descent direction* for F at x, meaning that there exists $\varepsilon > 0$ such that $F(x + \alpha d) \prec F(x)$ for all $\alpha \in (0, \varepsilon]$. Let $\mathcal{D} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be defined as follows:

$$\mathcal{D}(x, d) \coloneqq \max_{j=1,\ldots,m} \nabla F_j(x)^\top d.$$

The function $\mathcal{D}$ characterizes the descent directions for F at a given point x. Specifically, if $\mathcal{D}(x, d) < 0$, then d is a descent direction for F at x. Conversely, if $\mathcal{D}(x, d) \geq 0$ for all $d \in \mathbb{R}^n$, then x is a Pareto critical point.

We define $F : \mathbb{R}^n \to \mathbb{R}^m$ as convex (or strictly convex) if each component $F_j : \mathbb{R}^n \to \mathbb{R}$ is convex (or strictly convex) for all $j = 1, \ldots, m$, i.e., for all $x, y \in \mathbb{R}^n$ and $t \in [0, 1]$ (or $t \in (0, 1)$),

$$F((1-t)x + ty) \preceq (1-t)F(x) + tF(y) \quad (\text{or } \prec). \tag{3}$$

The following result establishes a relationship between the concepts of criticality, optimality, and convexity.

Lemma 2.1. *[22, Theorem 3.1] The following statements hold:*

(i) if x^* is local weak Pareto optimal, then x^* is a Pareto critical point for F;

(ii) if F is convex and x^* is Pareto critical for F, then x^* is weak Pareto optimal;

(iii) if F is strictly convex and x^* is Pareto critical for F, then x^* is Pareto optimal.

The class of quasi-Newton methods used to solve (1) consists of algorithms that compute the search direction $d(x)$ at a given point $x \in \mathbb{R}^n$ by solving the optimization problem:

$$\min_{d \in \mathbb{R}^n} \max_{j=1,\ldots,m} \nabla F_j(x)^\top d + \frac{1}{2} d^\top B_j d, \tag{4}$$

where $B_j \in \mathbb{R}^{n\times n}$ serves as an approximation of $\nabla^2 F_j(x)$ for all $j = 1, \dots, m$. If $B_j \succ 0$ for all $j = 1, \dots, m$, then the objective function of (4) is strongly convex, ensuring a unique solution for this problem. We denote the optimal value of (4) by $\theta(x)$, i.e.,

$$d(x) := \arg\min_{d\in\mathbb{R}^n} \max_{j=1,\dots,m} \nabla F_j(x)^\top d + \frac{1}{2} d^\top B_j d, \tag{5}$$

and

$$\theta(x) := \max_{j=1,\dots,m} \nabla F_j(x)^\top d(x) + \frac{1}{2} d(x)^\top B_j d(x). \tag{6}$$

One natural approach is to use a BFGS–type formula, which updates the approximation B_j in a way that preserves positive definiteness. In the case where $B_j = I_n$ for all $j = 1, \dots, m$, $d(x)$ represents the steepest descent direction (see [23]). Similarly, if $B_j = \nabla^2 F_j(x)$ for all $j = 1, \dots, m$, $d(x)$ corresponds to the Newton direction (see [22]).

In the following discussion, we assume that $B_j \succ 0$ for all $j = 1, \dots, m$. In this scenario, (4) is equivalent to the convex quadratic optimization problem:

$$\begin{aligned} \min_{(t,d)\in\mathbb{R}\times\mathbb{R}^n} \quad & t \\ \text{s. t.} \quad & \nabla F_j(x)^\top d + \frac{1}{2} d^\top B_j d \le t, \quad \forall j = 1, \dots, m. \end{aligned} \tag{7}$$

The unique solution to (7) is given by $(t, d) := (\theta(x), d(x))$. Since (7) is convex and has a Slater point (e.g., $(1, 0) \in \mathbb{R}\times\mathbb{R}^n$), there exists a multiplier $\lambda(x) \in \mathbb{R}^m$ such that the triple $(t, d, \lambda) := (\theta(x), d(x), \lambda(x)) \in \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^m$ satisfies its Karush-Kuhn-Tucker system given by:

$$\sum_{j=1}^m \lambda_j = 1, \quad \sum_{j=1}^m \lambda_j \left[\nabla F_j(x) + B_j d\right] = 0, \tag{8}$$

$$\lambda_j \ge 0, \ \nabla F_j(x)^\top d + \frac{1}{2} d^\top B_j d \le t, \quad \forall j = 1, \dots, m, \tag{9}$$

$$\lambda_j \left[\nabla F_j(x)^\top d + \frac{1}{2} d^\top B_j d - t\right] = 0, \quad \forall j = 1, \dots, m. \tag{10}$$

In particular, (8) and (9) imply that

$$d(x) = -\left[\sum_{j=1}^m \lambda_j(x) B_j\right]^{-1} \sum_{j=1}^m \lambda_j(x) \nabla F_j(x) \tag{11}$$

and

$$\lambda(x) \in \Delta_m, \tag{12}$$

where Δ_m represents the m-dimensional simplex defined as:

$$\Delta_m := \{\lambda \in \mathbb{R}^m \mid \sum_{j=1}^m \lambda_j = 1, \lambda \succeq 0\}.$$

Now, by summing (10) over $j = 1, \dots, m$ and using (11)–(12), we obtain

$$\begin{aligned}\theta(x) &= \left[\sum_{j=1}^{m} \lambda_j \nabla F_j(x)\right]^\top d(x) + \frac{1}{2} d(x)^\top \left[\sum_{j=1}^{m} \lambda_j(x) B_j\right] d(x) \\ &= -d(x)^\top \left[\sum_{j=1}^{m} \lambda_j(x) B_j\right] d(x) + \frac{1}{2} d(x)^\top \left[\sum_{j=1}^{m} \lambda_j(x) B_j\right] d(x) \\ &= -\frac{1}{2} d(x)^\top \left[\sum_{j=1}^{m} \lambda_j(x) B_j\right] d(x). \end{aligned} \tag{13}$$

Lemma 2.2. *Let $d : \mathbb{R}^n \to \mathbb{R}^n$ and $\theta : \mathbb{R}^n \to \mathbb{R}$ given by* (5) *and* (6)*, respectively. Assume that $B_j \succ 0$ for all $j = 1, \dots, m$. Then, we have:*

(i) x is Pareto critical if and only if $d(x) = 0$ and $\theta(x) = 0$;

(ii) if x is not Pareto critical, then $d(x) \neq 0$ and $\mathcal{D}(x, d(x)) < \theta(x) < 0$ (in particular, $d(x)$ is a descent direction for F at x).

Proof. See [22, Lemma 3.2] and [44, Lemma 2]. □

As previously mentioned, if $B_j = I_n$ for all $j = 1, \dots, m$, the solution of (4) corresponds to the steepest descent direction, denoted by $d_{SD}(x)$:

$$d_{SD}(x) \coloneqq \arg\min_{d \in \mathbb{R}^n} \max_{j=1,\dots,m} \nabla F_j(x)^\top d + \frac{1}{2}\|d\|^2. \tag{14}$$

Taking the above discussion into account, we can observe that there exists

$$\lambda^{SD}(x) \in \Delta_m \tag{15}$$

such that

$$d_{SD}(x) = -\sum_{j=1}^{m} \lambda_j^{SD}(x) \nabla F_j(x). \tag{16}$$

Next, we will review some useful properties related to $d_{SD}(\cdot)$.

Lemma 2.3. *Let $d_{SD} : \mathbb{R}^n \to \mathbb{R}^n$ be given by* (14)*. Then:*

(i) x is Pareto critical if and only if $d_{SD}(x) = 0$;

(ii) if x is not Pareto critical, then we have $d_{SD}(x) \neq 0$ and $\mathcal{D}(x, d_{SD}(x)) < -(1/2)\|d_{SD}(x)\|^2 < 0$ (in particular, $d_{SD}(x)$ is a descent direction for F at x);

(iii) the mapping $d_{SD}(\cdot)$ is continuous;

(iv) for any $x \in \mathbb{R}^n$, $-d_{SD}(x)$ is the minimal norm element of the set

$$\{u \in \mathbb{R}^n \mid u = \sum_{j=1}^{m} \lambda_j \nabla F_j(x), \lambda \in \Delta_m\},$$

i.e., in the convex hull of $\{\nabla F_1(x), \dots, \nabla F_m(x)\}$;

(v) if ∇F_j, $j = 1, \ldots, m$, are L-Lipschitz continuous on a nonempty set $U \subset \mathbb{R}^n$, i.e.,

$$\|\nabla F_j(x) - \nabla F_j(y)\| \leq L\|x - y\|, \quad \forall x, y \in U, \quad \forall j = 1, \ldots, m,$$

then the mapping $x \mapsto \|d_{SD}(x)\|$ is also L-Lipschitz continuous on U.

Proof. For items *(i)*, *(ii)*, and *(iii)*, see [32, Lemma 3.3]. For items *(iv)* and *(v)*, see [50, Corollary 2.3] and [50, Theorem 3.1], respectively. □

We end this section by presenting an auxiliary result.

Lemma 2.4. *The following statements are true.*

(i) The function $h(t) := 1 - t + \ln(t)$ is nonpositive for all $t > 0$.

(ii) For any $\bar{t} < 1$, we have $\ln(1 - \bar{t}) \geq -\bar{t}/(1 - \bar{t})$.

Proof. For item *(i)*, see [43, Exercise 6.8]. For item *(ii)*, consider $\bar{t} < 1$. By applying item *(i)* with $t = 1/(1 - \bar{t})$, we obtain

$$0 \geq h\Big(\frac{1}{1-\bar{t}}\Big) = 1 - \frac{1}{1-\bar{t}} + \ln\Big(\frac{1}{1-\bar{t}}\Big) = -\frac{\bar{t}}{1-\bar{t}} - \ln(1 - \bar{t}).$$

□

3 The algorithm and its global convergence

In this section, we define the main algorithm employed in this paper and study its global convergence, with a particular focus on nonconvex multiobjective optimization problems. Let us suppose that the following usual assumptions are satisfied.

Assumption 3.1. *(i) F is continuously differentiable.*
(ii) The level set $\mathcal{L} := \{x \in \mathbb{R}^n \mid F(x) \preceq F(x^0)\}$ is bounded, where $x^0 \in \mathbb{R}^n$ is the given starting point.
(iii) There exists an open set $\mathcal{N}$ containing $\mathcal{L}$ such that ∇F_j is L-Lipschitz continuous on $\mathcal{N}$ for all $j = 1, \ldots, m$, i.e.,

$$\|\nabla F_j(x) - \nabla F_j(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathcal{N}, \quad \forall j = 1, \ldots, m.$$

The algorithm is formally described as follows.

Algorithm 1. A BFGS-type algorithm for nonconvex problems

Let $\rho \in (0, 1/2)$, $\sigma \in (\rho, 1)$, $0 < \underline{\vartheta} \leq \bar{\vartheta}$, $x^0 \in \mathbb{R}^n$, and $B_j^0 \succ 0$ for all $j = 1, \ldots, m$ be given. Initialize $k \leftarrow 0$.

Step 1. *Compute the search direction*
Compute $d^k := d(x^k)$ as in (5).

Step 2. *Stopping criterion*
If x^k is Pareto critical, then STOP.

Step 3. *Line search procedure*

Compute a step size $\alpha_k > 0$ (trying first $\alpha_k = 1$) such that

$$F_j(x^k + \alpha_k d^k) \leq F_j(x^k) + \rho\alpha_k \mathcal{D}(x^k, d^k), \quad \forall j = 1, \ldots, m, \tag{17}$$

$$\mathcal{D}(x^k + \alpha_k d^k, d^k) \geq \sigma \mathcal{D}(x^k, d^k), \tag{18}$$

and set $x^{k+1} := x^k + \alpha_k d^k$.

Step 4. *Prepare the next iteration*

Compute

$$\eta_j^k = \frac{(y_j^k)^\top s^k}{\|s^k\|^2}, \quad \forall j = 1, \ldots, m, \tag{19}$$

where $s^k := x^{k+1} - x^k$ and $y_j^k := \nabla F_j(x^{k+1}) - \nabla F_j(x^k)$. Choose $\mu^k \in \Delta_m$ and $\vartheta_k \in (\underline{\vartheta}, \bar{\vartheta})$, and define

$$r_j^k := \max\{-\eta_j^k, 0\} + \vartheta_k \|\sum_{i=1}^m \mu_i^k \nabla F_i(x^k)\|, \quad \forall j = 1, \ldots, m, \tag{20}$$

and

$$\gamma_j^k := y_j^k + r_j^k s^k, \quad \forall j = 1, \ldots, m. \tag{21}$$

Step 5. *Update the BFGS-type matrices*

Define

$$B_j^{k+1} := B_j^k - \frac{B_j^k s^k (s^k)^\top B_j^k}{(s^k)^\top B_j^k s^k} + \frac{\gamma_j^k (\gamma_j^k)^\top}{(\gamma_j^k)^\top s^k}, \quad \forall j = 1, \ldots, m. \tag{22}$$

Set $k \leftarrow k + 1$ and go to Step 1.

Some comments are in order. First, by expressing the search direction subproblem (4) as the convex quadratic optimization problem (7), we can apply well-established techniques and solvers to find its solution at Step 1. Second, some practical stopping criteria can be considered at Step 2. It is usual to use the *gap* function $\theta(x^k)$ in (6) or the norm of $d_{SD}(x^k)$ in (14) to measure criticality, see Lemmas 2.2 and 2.3, respectively. Third, at Step 3, we require that α_k satisfies (17)–(18), which corresponds to the multiobjective standard Wolfe conditions originally introduced in [37]. Under Assumption 3.1*(i)–(ii)*, given $d^k \in \mathbb{R}^n$ a descent direction for F at x^k, it is possible to show that there are intervals of positive step sizes satisfying (17)–(18), see [37, Proposition 3.2]. As we will see, under suitable assumptions, the unit step size $\alpha_k = 1$ is eventually accepted, which is essential to obtain *fast* convergence. Furthermore, an algorithm to calculate Wolfe step sizes for vector-valued problems was proposed in [38]. Fourth, the usual BFGS scheme for F_j consists of the update formula given in (22) with y_j^k in place of γ_j^k. In this case, the product $(y_j^k)^\top s^k$ in the denominator of the third term on the right-hand side of (22) can be nonpositive for some $j \in \{1, \ldots, m\}$, even when the step size satisfies the Wolfe conditions (17)–(18), see [46, Example 3.3]. This implies that update scheme (with y_j^k in place of γ_j^k) may fail to preserve positive definiteness of B_j^k. Fifth, note that $B_j^{k+1} s^k = y_j^k + r_j^k s^k$ for each $j = 1, \ldots, m$. Thus, if r_j^k is *small*, this relation can be seen as an approximation of the well-known secant equation $B_j^{k+1} s^k = y_j^k$ for F_j, see [35].

Theorem 3.1. *Algorithm 1 is well-defined.*

Proof. The proof is by induction. We start by assuming that $B_j^k \succ 0$ for all $j = 1, \dots, m$, which is trivially true for $k = 0$. This makes subproblem (5) in Step 1 solvable. If x^k is Pareto critical, Algorithm 1 stops at Step 2, thereby concluding the proof. Otherwise, Lemma 2.2*(ii)* implies that d^k is a descent direction of F at x^k. Thus, taking into account Assumption 3.1*(i)*–*(ii)*, there exist intervals of positive step sizes satisfying conditions (17)–(18), as shown in [37, Proposition 3.2]. As a result, x^{k+1} can be properly defined in Step 3. To complete the proof, let us show that B_j^{k+1} remains positive definite for all $j = 1, \dots, m$. By the definitions of γ_j^k and η_j^k in (21) and (19), respectively, we have

$$(\gamma_j^k)^\top s^k = (y_j^k)^\top s^k + r_j^k \|s^k\|^2 = \left(\frac{(y_j^k)^\top s^k}{\|s^k\|^2} + r_j^k \right) \|s^k\|^2 = (\eta_j^k + r_j^k)\|s^k\|^2, \quad \forall j = 1, \dots, m.$$

Therefore, by the definition of r_j^k in (20), Lemma 2.3*(iv)*, and Lemma 2.3*(ii)*, it follows that

$$(\gamma_j^k)^\top s^k \geq \vartheta_k \| \sum_{i=1}^m \mu_i^k \nabla F_i(x^k) \| \|s^k\|^2 \geq \underline{\vartheta} \|d_{SD}(x^k)\| \|s^k\|^2 > 0, \quad \forall j = 1, \dots, m. \tag{23}$$

Thus, the updating formulas (22) are well-defined. Now, for each $j \in \{1, \dots, m\}$ and any nonzero vector z, we have

$$z^\top B_j^{k+1} z = \|z\|^2_{B_j^k} - \frac{\langle z, s^k \rangle^2_{B_j^k}}{\|s^k\|^2_{B_j^k}} + \frac{(z^\top \gamma_j^k)^2}{(\gamma_j^k)^\top s^k} \geq \frac{(z^\top \gamma_j^k)^2}{(\gamma_j^k)^\top s^k} \geq 0, \tag{24}$$

where the first inequality is a direct consequence of the Cauchy-Schwarz inequality, which gives $\langle z, s^k \rangle^2_{B_j^k} \leq \|z\|^2_{B_j^k} \|s^k\|^2_{B_j^k}$. Finally, assume by contradiction that $z^\top B_j^{k+1} z = 0$. In this case, it follows from (24) that

$$z^\top \gamma_j^k = 0 \quad \text{and} \quad \|z\|^2_{B_j^k} - \frac{\langle z, s^k \rangle^2_{B_j^k}}{\|s^k\|^2_{B_j^k}} = 0. \tag{25}$$

The second equation in (25) implies that $|\langle z, s^k \rangle_{B_j^k}| = \|z\|_{B_j^k} \|s^k\|_{B_j^k}$, so there exists $\tau \in \mathbb{R}$ such that $z = \tau s^k$. Combining this with the first equation in (25), we obtain $\tau (\gamma_j^k)^\top s^k = 0$. Taking into account (23), we can deduce that $\tau = 0$, which contradicts the fact that z is a nonzero vector. □

Hereafter, we assume that x^k is not Pareto critical for all $k \geq 0$. Thus, Algorithm 1 generates an infinite sequence of iterates. The following result establishes that the sequence $\{x^k, d^k\}$ satisfies a Zoutendijk-type condition, which will be crucial in our analysis. Its proof is based on [37, Proposition 3.3] and will be provided in the Appendix.

Proposition 3.2. *Consider the sequence $\{x^k, d^k\}$ generated by Algorithm 1. Then,*

$$\sum_{k \geq 0} \frac{\mathcal{D}(x^k, d^k)^2}{\|d^k\|^2} < \infty. \tag{26}$$

Our analysis also exploits insights developed by Byrd and Nocedal [7] in their analysis of the classical BFGS method for single-valued optimization (i.e., for $m = 1$). They provided sufficient conditions that ensure that the angle between s^k and $B_1^k s^k$ (which coincides with the angle between d^k and $-\nabla F_1(x^k)$ in the scalar case) remains far from 0 for an arbitrary fraction of the iterates. Recently, this result was studied in the multiobjective setting in [46]. Under some mild conditions, similar to the approach taken in [46], we establish that the angles between s^k and $B_j^k s^k$ remain far from 0, *simultaneously* for all objectives, for an arbitrary fraction p of the iterates. The proof of this result can be constructed as a combination of [7, Theorem 2.1] and [46, Lemma 4.2] and will therefore be postponed to the Appendix.

Proposition 3.3. *Consider the sequence $\{x^k\}$ generated by Algorithm 1. Let β_j^k be the angle between the vectors s^k and $B_j^k s^k$, for all $k \geq 0$ and $j = 1, \dots, m$. Assume that*

$$\frac{(\gamma_j^k)^\top s^k}{\|s^k\|^2} \geq C_1 \quad \text{and} \quad \frac{\|\gamma_j^k\|^2}{(\gamma_j^k)^\top s^k} \leq C_2, \quad \forall j = 1, \dots, m, \quad \forall k \geq 0, \tag{27}$$

for some positive constants $C_1, C_2 > 0$. Then, given $p \in (0,1)$, there exists a constant $\delta > 0$ such that, for all $k \geq 1$, the relation

$$\cos \beta_j^\ell \geq \delta, \quad \forall j = 1, \dots, m,$$

holds for at least $\lceil p(k+1) \rceil$ values of $\ell \in \{0, 1, \dots, k\}$, where $\lceil \cdot \rceil$ denotes the ceiling function.

The following technical result forms the basis for applying Proposition 3.3.

Lemma 3.4. *Let $\{x^k\}$ be a sequence generated by Algorithm 1. Then, for each j = 1,...,m and all $k \geq 0$, there exist positive constants $c_1, c_2 > 0$ such that:*

(i) $\dfrac{(\gamma_j^k)^\top s^k}{\|s^k\|^2} \geq c_1 \|d_{SD}(x^k)\|$;

(ii) $\dfrac{\|\gamma_j^k\|^2}{(\gamma_j^k)^\top s^k} \leq \dfrac{c_2}{\|d_{SD}(x^k)\|}$.

Proof. Let $k \geq 0$ and $j \in \{1, \dots, m\}$ be given. As in (23), we have

$$\frac{(\gamma_j^k)^\top s^k}{\|s^k\|^2} \geq \underline{\vartheta} \|d_{SD}(x^k)\|, \quad \forall j = 1, \dots, m.$$

Thus, taking $c_1 := \underline{\vartheta}$, we conclude item *(i)*. Now consider item *(ii)*. By the Cauchy–Schwarz inequality and Assumption 3.1*(iii)*, it follows that

$$|\eta_j^k| = \frac{|(y_j^k)^\top s^k|}{\|s^k\|^2} \leq \frac{\|y_j^k\|}{\|s^k\|} = \frac{\|\nabla F_j(x^{k+1}) - \nabla F_j(x^k)\|}{\|x^{k+1} - x^k\|} \leq L.$$

On the other hand, since $\{x^k\} \subset \mathcal{L}$ and $\mu^k \in \Delta_m$, by Assumption 3.1*(i)–(ii)*, there exists a constant $\bar{c} > 0$, independent of k, such that $\| \sum_{i=1}^m \mu_i^k \nabla F_i(x^k)\| \leq \bar{c}$. The definition of r_j^k together with the last two inequalities yields

$$r_j^k \leq |\eta_j^k| + \vartheta_k \| \sum_{i=1}^m \mu_i^k \nabla F_i(x^k)\| \leq L + \bar{\vartheta}\bar{c},$$

and hence

$$\|\gamma_j^k\| \leq \|y_j^k\| + r_j^k \|s^k\| = \left(\frac{\|y_j^k\|}{\|s^k\|} + r_j^k \right) \|s^k\| \leq (2L + \bar{\vartheta}\bar{c}) \|s^k\|.$$

By squaring the latter inequality and using item *(i)*, we obtain

$$\frac{\|\gamma_j^k\|^2}{(\gamma_j^k)^\top s^k} \leq (2L + \bar{\vartheta}\bar{c})^2 \frac{\|s^k\|^2}{(\gamma_j^k)^\top s^k} \leq \frac{(2L + \bar{\vartheta}\bar{c})^2}{c_1 \|d_{SD}(x^k)\|}.$$

Thus, taking $c_2 := (2L + \bar{\vartheta}\bar{c})^2/c_1$, we conclude the proof. □

From now on, let $\lambda^k := \lambda(x^k) \in \mathbb{R}^m$ be the Lagrange multiplier associated with x^k satisfying (11)–(12). We are now able to prove the main result of this section. We show that Algorithm 1 finds a Pareto critical point of F, without imposing any convexity assumptions.

Theorem 3.5. *Let $\{x^k\}$ be a sequence generated by Algorithm 1. Then*

$$\liminf_{k\to\infty} \|d_{SD}(x^k)\| = 0. \tag{28}$$

As a consequence, $\{x^k\}$ has a limit point that is Pareto critical.

Proof. Assume by contradiction that there is a constant $\varepsilon > 0$ such that

$$\|d_{SD}(x^k)\| \geq \varepsilon, \quad \forall k \geq 0. \tag{29}$$

From Lemma 3.4, taking $C_1 := c_1\varepsilon$ and $C_2 := c_2/\varepsilon$, we have

$$\frac{(\gamma_j^k)^\top s^k}{\|s^k\|^2} \geq C_1 \quad \text{and} \quad \frac{\|\gamma_j^k\|^2}{(\gamma_j^k)^\top s^k} \leq C_2, \quad \forall j = 1, \ldots, m, \quad \forall k \geq 0,$$

showing that the assumptions of Proposition 3.3 are satisfied. Thus, there exist a constant $\delta > 0$ and $\mathbb{K} \underset{\infty}{\subset} \mathbb{N}$ such that

$$\cos\beta_j^k \geq \delta, \quad \forall j = 1, \ldots, m, \quad \forall k \in \mathbb{K}.$$

Hence, by the definitions of $\cos\beta_j^k$ and s^k, we have, for all $j = 1, \ldots, m$,

$$\delta \leq \cos\beta_j^k = \frac{(s^k)^\top B_j^k s^k}{\|s^k\|\|B_j^k s^k\|} = \frac{(d^k)^\top B_j^k d^k}{\|d^k\|\|B_j^k d^k\|}, \quad \forall k \in \mathbb{K},$$

which implies

$$(d^k)^\top B_j^k d^k \geq \delta \|d^k\|\|B_j^k d^k\|, \quad \forall k \in \mathbb{K}.$$

Therefore, from Lemma 2.2*(ii)* and (13), it follows that

$$-\mathcal{D}(x^k, d^k) > -\theta(x^k) = \frac{1}{2}\sum_{j=1}^m \lambda_j^k (d^k)^\top B_j^k d^k \geq \frac{\delta}{2}\|d^k\| \sum_{j=1}^m \lambda_j^k \|B_j^k d^k\|, \quad \forall k \in \mathbb{K}.$$

Thus, from the triangle inequality, (11), (12), Lemma 2.3*(iv)*, and (29), we obtain

$$-\frac{\mathcal{D}(x^k, d^k)}{\|d^k\|} \geq \frac{\delta}{2}\|\sum_{j=1}^m \lambda_j^k B_j^k d^k\| = \frac{\delta}{2}\|\sum_{j=1}^m \lambda_j^k \nabla F_j(x^k)\| \geq \frac{\delta}{2}\|d_{SD}(x^k)\| \geq \frac{\delta\varepsilon}{2}, \quad \forall k \in \mathbb{K}.$$

Hence,

$$\sum_{k\geq 0} \frac{\mathcal{D}(x^k, d^k)^2}{\|d^k\|^2} \geq \sum_{k\in\mathbb{K}} \frac{\mathcal{D}(x^k, d^k)^2}{\|d^k\|^2} \geq \sum_{k\in\mathbb{K}} \frac{\delta^2\varepsilon^2}{4} = \infty,$$

which contradicts the Zoutendijk condition (26). Therefore, we conclude that (28) holds.

Now, (28) implies that there exists $\mathbb{K}_1 \underset{\infty}{\subset} \mathbb{N}$ such that $\lim_{k\in\mathbb{K}_1} \|d_{SD}(x^k)\| = 0$. On the other hand, given that $\{x^k\} \subset \mathcal{L}$ and $\mathcal{L}$ is compact, we can establish the existence of $\mathbb{K}_2 \subseteq \mathbb{K}_1$ and $x^* \in \mathcal{L}$ such that $\lim_{k\in\mathbb{K}_2} x^k = x^*$. Thus, from Lemma 2.3*(iii)*, we deduce that $d_{SD}(x^*) = 0$. Consequently, based on Lemma 2.3*(i)*, we conclude that x^* is Pareto critical. □

Even though the primary focus of this article is on nonconvex problems, we conclude this section by establishing full convergence of the sequence generated by Algorithm 1 in the case of strict convexity of F. Note that, under Assumption 3.1*(i)–(ii)*, the existence of at least one Pareto optimal point is assured in this particular case.

Theorem 3.6. *Let $\{x^k\}$ be a sequence generated by Algorithm 1. If F is strictly convex, then $\{x^k\}$ converges to a Pareto optimal point of F.*

Proof. According to Theorem 3.5 and Theorem 2.1*(iii)*, there exists a limit point $x^* \in \mathcal{L}$ of $\{x^k\}$ that is Pareto optimal. Let $\mathbb{K}_1 \underset{\infty}{\subset} \mathbb{N}$ be such that $\lim_{k\in\mathbb{K}_1} x^k = x^*$. To show the convergence of $\{x^k\}$ to x^*, let us suppose by contradiction that there exist $\bar{x} \in \mathcal{L}$, where $\bar{x} \neq x^*$, and $\mathbb{K}_2 \underset{\infty}{\subset} \mathbb{N}$ such that $\lim_{k\in\mathbb{K}_2} x^k = \bar{x}$. We first claim that $F(\bar{x}) \neq F(x^*)$. In fact, if $F(\bar{x}) = F(x^*)$, based on (3), for all $t \in (0,1)$, we would have

$$F((1-t)x^* + t\bar{x}) \prec (1-t)F(x^*) + tF(\bar{x}) = F(x^*),$$

which contradicts the fact that x^* is a Pareto optimal point. Hence, $F(\bar{x}) \neq F(x^*)$, as we claimed. Now, since x^* is Pareto optimal, there exists $j_* \in \{1,\ldots,m\}$ such that $F_{j_*}(x^*) < F_{j_*}(\bar{x})$. Therefore, considering that $\lim_{k\in\mathbb{K}_1} x^k = x^*$ and $\lim_{k\in\mathbb{K}_2} x^k = \bar{x}$, we can choose $k_1 \in \mathbb{K}_1$ and $k_2 \in \mathbb{K}_2$ such that $k_1 < k_2$ and $F_{j_*}(x^{k_1}) < F_{j_*}(x^{k_2})$. This contradicts (17) which implies, in particular, that $\{F_{j_*}(x^k)\}$ is decreasing. Thus, we can conclude that $\lim_{k\to\infty} x^k = x^*$, completing the proof. □

4 Local convergence analysis

In this section, we analyze the local convergence properties of Algorithm 1. The findings presented here are applicable to both convex and nonconvex problems. We will assume that the sequence $\{x^k\}$ converges to a local Pareto optimal point x^* and show, under appropriate assumptions, that the convergence rate is superlinear.

4.1 Superlinear rate of convergence

Throughout this section, we make the following assumptions.

Assumption 4.1. *(i) F is twice continuously differentiable.*
(ii) The sequence $\{x^k\}$ generated by Algorithm 1 converges to a local Pareto optimal point x^* where $\nabla^2 F_j(x^*)$ is positive definite for all $j = 1,\ldots,m$.
(iii) For each $j = 1,\ldots,m$, $\nabla^2 F_j(x)$ is Hölder continuous at x^*, i.e., there exist constants $\nu \in (0,1]$ and $M > 0$ such that

$$\|\nabla^2 F_j(x) - \nabla^2 F_j(x^*)\| \leq M\|x - x^*\|^{\nu}, \quad \forall j = 1,\ldots,m, \tag{30}$$

for all x in a neighborhood of x^*.

Under Assumption 4.1*(ii)*, there exist a neighborhood U of x^* and constants $\underline{L} > 0$ and $L > 0$ such that

$$\underline{L}\|z\|^2 \leq z^\top \nabla^2 F_j(x) z \leq L\|z\|^2, \quad \forall j = 1, \dots, m, \tag{31}$$

for all $z \in \mathbb{R}^n$ and $x \in U$. In particular, (31) implies that F_j is strongly convex and has Lipschitz continuous gradients on U. Note that constant L in (31) aligns with the L defined in Assumption 3.1*(iii)* as part of the L-Lipschitz continuity condition for ∇F_j, maintaining consistent notation for the Lipschitz constant across our analysis. Throughout this section, we assume, without loss of generality, that $\{x^k\} \subset U$ and that Assumption 4.1*(iii)* holds in U, i.e., (30) and (31) hold at x^k for all $k \geq 0$.

We also introduce the following additional assumption about $\{r_j^k\}$, which will be considered only when explicitly mentioned. In Section 4.2, we will explore practical choices for $\{r_j^k\}$ that satisfy such an assumption.

Assumption 4.2. *For each $j = 1, \dots, m$, $\{r_j^k\}$ satisfies $\sum_{k \geq 0} r_j^k < \infty$.*

The following result, which is related to the linear convergence of the sequence $\{x^k\}$ and is based on [46, Theorem 4.6], has its proof in the Appendix.

Proposition 4.1. *Suppose that Assumption 4.1(i)–(ii) holds. Let $\{x^k\}$ be a sequence generated by Algorithm 1. Then, for all $\nu > 0$, we have*

$$\sum_{k \geq 0} \|x^k - x^*\|^\nu < \infty. \tag{32}$$

As usual in quasi-Newton methods, our analysis relies on the Dennis–Moré [17] characterization of superlinear convergence. To accomplish this, we use a set of tools developed in [7] (see also [45]). For every $k \geq 0$, we define the *average Hessian* as:

$$\bar{G}_j^k := \int_0^1 \nabla^2 F_j(x^k + \tau s^k) d\tau, \quad \forall j = 1, \dots, m.$$

This leads to the relationship:

$$\bar{G}_j^k s^k = y_j^k, \quad \forall j = 1, \dots, m. \tag{33}$$

We also introduce, for each $j = 1, \dots, m$ and $k \geq 0$, the following quantities:

$$\tilde{s}_j^k := \nabla^2 F_j(x^*)^{1/2} s^k, \quad \tilde{y}_j^k := \nabla^2 F_j(x^*)^{-1/2} y_j^k, \quad \tilde{\gamma}_j^k := \nabla^2 F_j(x^*)^{-1/2} \gamma_j^k,$$

and

$$\tilde{B}_j^k := \nabla^2 F_j(x^*)^{-1/2} B_j^k \nabla^2 F_j(x^*)^{-1/2}.$$

Note that

$$\tilde{B}_j^{k+1} = \tilde{B}_j^k - \frac{\tilde{B}_j^k \tilde{s}_j^k (\tilde{s}_j^k)^\top \tilde{B}_j^k}{(\tilde{s}_j^k)^\top \tilde{B}_j^k \tilde{s}_j^k} + \frac{\tilde{\gamma}_j^k (\tilde{\gamma}_j^k)^\top}{(\tilde{\gamma}_j^k)^\top \tilde{s}_j^k}, \quad \forall j = 1, \dots, m,$$

and

$$\frac{(\tilde{\gamma}_j^k)^\top \tilde{s}_j^k}{\|s^k\|^2} = \frac{(\gamma_j^k)^\top s^k}{\|s^k\|^2} = \frac{(y_j^k)^\top s^k}{\|s^k\|^2} + r_j^k = \frac{(s^k)^\top \bar{G}_j^k s^k}{\|s^k\|^2} + r_j^k \geq \underline{L} + r_j^k \geq \underline{L}, \quad \forall j = 1, \dots, m, \tag{34}$$

where the first inequality follows from the left hand side of (31). Considering that $\tilde{B}_j^k \succ 0$ and, based on (34), it follows that $(\tilde{\gamma}_j^k)^\top \tilde{s}_j^k > 0$, we can follow the same arguments as in the proof of Theorem 3.1 to show that $\tilde{B}_j^{k+1} \succ 0$ for all $j = 1, \dots, m$ and all $k \geq 0$. In connection with Proposition 3.3 and Lemma 3.4, we additionally define the following quantities:

$$\tilde{a}_j^k := \frac{(\tilde{\gamma}_j^k)^\top \tilde{s}_j^k}{\|\tilde{s}_j^k\|^2}, \quad \tilde{b}_j^k := \frac{\|\tilde{\gamma}_j^k\|^2}{(\tilde{\gamma}_j^k)^\top \tilde{s}_j^k}, \quad \cos \tilde{\beta}_j^k := \frac{(\tilde{s}_j^k)^\top \tilde{B}_j^k \tilde{s}_j^k}{\|\tilde{s}_j^k\| \|\tilde{B}_j^k \tilde{s}_j^k\|}, \quad \text{and} \quad \tilde{q}_j^k := \frac{(\tilde{s}_j^k)^\top \tilde{B}_j^k \tilde{s}_j^k}{\|\tilde{s}_j^k\|^2}.$$

Another useful tool combines the trace and the determinant of a given positive definite matrix B, through the following function:

$$\psi(B) := \text{trace}(B) - \ln(\det(B)). \tag{35}$$

Given that, for all $j = 1, \dots, m$,

$$\text{trace}(\tilde{B}_j^{k+1}) = \text{trace}(\tilde{B}_j^k) - \frac{\|\tilde{B}_j^k \tilde{s}_j^k\|^2}{(\tilde{s}_j^k)^\top \tilde{B}_j^k \tilde{s}_j^k} + \frac{\|\tilde{\gamma}_j^k\|^2}{(\tilde{\gamma}_j^k)^\top \tilde{s}_j^k} = \text{trace}(\tilde{B}_j^k) - \frac{\tilde{q}_j^k}{\cos^2 \tilde{\beta}_j^k} + \tilde{b}_j^k$$

and

$$\det(\tilde{B}_j^{k+1}) = \det(\tilde{B}_j^k) \frac{(\tilde{\gamma}_j^k)^\top \tilde{s}_j^k}{(\tilde{s}_j^k)^\top \tilde{B}_j^k \tilde{s}_j^k} = \det(\tilde{B}_j^k) \frac{\tilde{a}_j^k}{\tilde{q}_j^k},$$

we can perform some algebraic manipulations to obtain:

$$\psi(\tilde{B}_j^{k+1}) = \psi(\tilde{B}_j^k) + \left(\tilde{b}_j^k - \ln(\tilde{a}_j^k) - 1\right) + \left[1 - \frac{\tilde{q}_j^k}{\cos^2 \tilde{\beta}_j^k} + \ln\left(\frac{\tilde{q}_j^k}{\cos^2 \tilde{\beta}_j^k}\right)\right] + \ln(\cos^2 \tilde{\beta}_j^k). \tag{36}$$

We are now ready to prove the central result of the superlinear convergence analysis: We establish that the Dennis–Moré condition holds individually for each objective function F_j. A similar result in the scalar case was given in [35, Theorem 3.8].

Theorem 4.2. *Suppose that Assumptions 4.1 and 4.2 hold. Let $\{x^k\}$ be a sequence generated by Algorithm 1. Then, for each $j = 1, \dots, m$, we have*

$$\lim_{k \to \infty} \frac{(s^k)^\top B_j^k s^k}{(s^k)^\top \nabla^2 F_j(x^*) s^k} = 1, \tag{37}$$

and

$$\lim_{k \to \infty} \frac{\|(B_j^k - \nabla^2 F_j(x^*)) d^k\|}{\|d^k\|} = 0. \tag{38}$$

Proof. Let $j \in \{1, \dots, m\}$ be an arbitrary index. From (33), we obtain

$$y_j^k - \nabla^2 F_j(x^*) s^k = [\bar{G}_j^k - \nabla^2 F_j(x^*)] s^k,$$

and hence

$$\tilde{y}_j^k - \tilde{s}_j^k = \nabla^2 F_j(x^*)^{-1/2} [\bar{G}_j^k - \nabla^2 F_j(x^*)] \nabla^2 F_j(x^*)^{-1/2} \tilde{s}_j^k,$$

for all $k \geq 0$. Therefore, by the definition of $\bar{G}_j^k$ and (30), we obtain

$$\begin{aligned}
\|\tilde{y}_j^k - \tilde{s}_j^k\| &\leq \|\nabla^2 F_j(x^*)^{-1/2}\|^2 \|\tilde{s}_j^k\| \|\bar{G}_j^k - \nabla^2 F_j(x^*)\| \\
&\leq M \|\nabla^2 F_j(x^*)^{-1/2}\|^2 \|\tilde{s}_j^k\| \int_0^1 \|x^k + \tau s^k - x^*\|^\nu d\tau \leq \bar{c}_j \varepsilon_k \|\tilde{s}_j^k\|, \quad \forall k \geq 0,
\end{aligned} \tag{39}$$

where $\bar{c}_j := M\|\nabla^2 F_j(x^*)^{-1/2}\|^2$ and $\varepsilon_k := \max\{\|x^{k+1} - x^*\|^\nu, \|x^k - x^*\|^\nu\}$. Now, since $|\|\tilde{y}_j^k\| - \|\tilde{s}_j^k\|| \leq \|\tilde{y}_j^k - \tilde{s}_j^k\|$, it follows from (39) that

$$(1 - \bar{c}_j \varepsilon_k)\|\tilde{s}_j^k\| \leq \|\tilde{y}_j^k\| \leq (1 + \bar{c}_j \varepsilon_k)\|\tilde{s}_j^k\|. \tag{40}$$

Without loss of generality, let us assume that $1 - \bar{c}_j \varepsilon_k > 0$, for all $k \geq 0$. Therefore, the left hand side of (40) together with (39) yields

$$(1 - \bar{c}_j \varepsilon_k)^2 \|\tilde{s}_j^k\|^2 - 2(\tilde{y}_j^k)^\top \tilde{s}_j^k + \|\tilde{s}_j^k\|^2 \leq \|\tilde{y}_j^k\|^2 - 2(\tilde{y}_j^k)^\top \tilde{s}_j^k + \|\tilde{s}_j^k\|^2 \leq \bar{c}_j^2 \varepsilon_k^2 \|\tilde{s}_j^k\|^2,$$

so that

$$2(\tilde{y}_j^k)^\top \tilde{s}_j^k \geq (1 - \bar{c}_j \varepsilon_k)^2 \|\tilde{s}_j^k\|^2 + \|\tilde{s}_j^k\|^2 - \bar{c}_j^2 \varepsilon_k^2 \|\tilde{s}_j^k\|^2 = 2(1 - \bar{c}_j \varepsilon_k)\|\tilde{s}_j^k\|^2. \tag{41}$$

By the definition of $\tilde{a}_j^k$, we have

$$\tilde{a}_j^k = \frac{(\tilde{y}_j^k + r_j^k \nabla^2 F_j(x^*)^{-1} \tilde{s}_j^k)^\top \tilde{s}_j^k}{\|\tilde{s}_j^k\|^2} = \frac{(\tilde{y}_j^k)^\top \tilde{s}_j^k}{\|\tilde{s}_j^k\|^2} + r_j^k \frac{(\tilde{s}_j^k)^\top \nabla^2 F_j(x^*)^{-1} \tilde{s}_j^k}{\|\tilde{s}_j^k\|^2}.$$

Thus, by (31) and (41), we obtain

$$\tilde{a}_j^k \geq 1 - \bar{c}_j \varepsilon_k + \frac{r_j^k}{L} \geq 1 - \bar{c}_j \varepsilon_k. \tag{42}$$

From the definition of $\tilde{b}_j^k$, (34), the right hand side of (40), and (42), by performing some manipulations, we also obtain

$$\begin{aligned}
\tilde{b}_j^k &= \frac{\|\tilde{y}_j^k\|^2}{\tilde{a}_j^k \|\tilde{s}_j^k\|^2} + 2r_j^k \frac{(\tilde{y}_j^k)^\top \nabla^2 F_j(x^*)^{-1} \tilde{s}_j^k}{(\tilde{\gamma}_j^k)^\top \tilde{s}_j^k} + (r_j^k)^2 \frac{(\tilde{s}_j^k)^\top \nabla^2 F_j(x^*)^{-2} \tilde{s}_j^k}{(\tilde{\gamma}_j^k)^\top \tilde{s}_j^k} \\
&\leq \frac{(1 + \bar{c}_j \varepsilon_k)^2}{1 - \bar{c}_j \varepsilon_k} + 2r_j^k \frac{\|\nabla^2 F_j(x^*)^{-1/2}\| \|y_j^k\| \|\nabla^2 F_j(x^*)^{-1}\| \|\nabla^2 F_j(x^*)^{1/2}\| \|s^k\|}{\underline{L}\|s^k\|^2} \\
&\quad + (r_j^k)^2 \frac{\|\nabla^2 F_j(x^*)^{-2}\| \|\tilde{s}_j^k\|^2}{\underline{L}\|s^k\|^2} \\
&\leq 1 + \frac{3\bar{c}_j + \bar{c}_j^2 \varepsilon_k}{1 - \bar{c}_j \varepsilon_k} \varepsilon_k + \frac{2LC_1}{\underline{L}} r_j^k + \frac{C_2}{\underline{L}} (r_j^k)^2,
\end{aligned} \tag{43}$$

where $C_1 := \|\nabla^2 F_j(x^*)^{-1/2}\| \|\nabla^2 F_j(x^*)^{-1}\| \|\nabla^2 F_j(x^*)^{1/2}\|$ and $C_2 := \|\nabla^2 F_j(x^*)^{-2}\| \|\nabla^2 F_j(x^*)\|$. Assumptions 4.1*(ii)* and 4.2 imply that $\varepsilon_k \to 0$ and $r_j^k \to 0$, respectively. Thus, by using (42) and (43), for all sufficiently large k, we have $\bar{c}_j \varepsilon_k < 1/2$, $(r_j^k)^2 \leq r_j^k$, and there exists a constant $C > \max\{3\bar{c}_j, (2LC_1 + C_2)/\underline{L}\}$ such that

$$\ln(\tilde{a}_j^k) \geq \ln(1 - \bar{c}_j \varepsilon_k) \geq -\frac{\bar{c}_j \varepsilon_k}{1 - \bar{c}_j \varepsilon_k} \geq -2\bar{c}_j \varepsilon_k > -2C\varepsilon_k,$$

where the second inequality follows from Lemma 2.4 *(ii)* (with $\bar{t} = \bar{c}_j \varepsilon_k$), and

$$\tilde{b}_j^k \leq 1 + C\varepsilon_k + Cr_j^k.$$

Let $k_0 \in \mathbb{N}$ be such that the latter two inequalities hold for all $k \geq k_0$. Therefore, by (36), we obtain

$$\ln\Big(\frac{1}{\cos^2 \tilde{\beta}_j^k}\Big) - \left[1 - \frac{\tilde{q}_j^k}{\cos^2 \tilde{\beta}_j^k} + \ln\Big(\frac{\tilde{q}_j^k}{\cos^2 \tilde{\beta}_j^k}\Big)\right] < \psi(\tilde{B}_j^k) - \psi(\tilde{B}_j^{k+1}) + 3C\varepsilon_k + Cr_j^k,$$

for all $k \geq k_0$. By summing this expression and making use of (32) and Assumption 4.2, we have

$$\sum_{\ell \geq k_0} \left\{ \ln\Big(\frac{1}{\cos^2 \tilde{\beta}_j^\ell}\Big) - \left[1 - \frac{\tilde{q}_j^\ell}{\cos^2 \tilde{\beta}_j^\ell} + \ln\Big(\frac{\tilde{q}_j^\ell}{\cos^2 \tilde{\beta}_j^\ell}\Big)\right] \right\} \leq \psi(\tilde{B}_j^{k_0}) + 3C \sum_{\ell \geq k_0} \varepsilon_k + C \sum_{\ell \geq k_0} r_j^\ell < \infty.$$

Since $\ln(1/\cos^2 \tilde{\beta}_j^\ell) > 0$ for all $\ell \geq k_0$ and, by Lemma 2.4 *(i)*, the term in the square brackets is nonpositive, we have

$$\lim_{\ell \to \infty} \ln\Big(\frac{1}{\cos^2 \tilde{\beta}_j^\ell}\Big) = 0 \quad \text{and} \quad \lim_{\ell \to \infty} \left[1 - \frac{\tilde{q}_j^\ell}{\cos^2 \tilde{\beta}_j^\ell} + \ln\Big(\frac{\tilde{q}_j^\ell}{\cos^2 \tilde{\beta}_j^\ell}\Big)\right] = 0,$$

and hence

$$\lim_{\ell \to \infty} \cos^2 \tilde{\beta}_j^\ell = 1 \quad \text{and} \quad \lim_{\ell \to \infty} \tilde{q}_j^\ell = 1. \tag{44}$$

Note that the second limit in (44) is equivalent to (37). Now, it follows that

$$\begin{aligned}
\lim_{k \to \infty} \frac{\|\nabla^2 F_j(x^*)^{-1/2}(B_j^k - \nabla^2 F_j(x^*))s^k\|^2}{\|\nabla^2 F_j(x^*)^{1/2} s^k\|^2} &= \lim_{k \to \infty} \frac{\|(\tilde{B}_j^k - I_n)\tilde{s}_j^k\|^2}{\|\tilde{s}_j^k\|^2} \\
&= \lim_{k \to \infty} \frac{\|\tilde{B}_j^k \tilde{s}_j^k\|^2 - 2(\tilde{s}_j^k)^\top \tilde{B}_j^k \tilde{s}_j^k + \|\tilde{s}_j^k\|^2}{\|\tilde{s}_j^k\|^2} \\
&= \lim_{k \to \infty} \left[\frac{(\tilde{q}_j^k)^2}{\cos^2 \tilde{\beta}_j^k} - 2\tilde{q}_j^k + 1\right] = 0,
\end{aligned}$$

where the last equality follows from (44). The above limit trivially implies (38), concluding the proof. □

Based on the Dennis–Moré characterization established in Theorem 4.2, we can easily replicate the proofs presented in [46, Theorem 5.5] and [46, Theorem 5.7] to show that unit step size eventually satisfies the Wolfe conditions (17)–(18) and the rate of convergence is superlinear, as detailed in the Appendix. We formally state the results as follows.

Theorem 4.3. *Suppose that Assumptions 4.1 and 4.2 hold. Let $\{x^k\}$ be a sequence generated by Algorithm 1. Then, the step size $\alpha_k = 1$ is admissible for all sufficiently large k and $\{x^k\}$ converges to x^* at a superlinear rate.*

4.2 Suitable choices for r_j^k

As we have seen, Algorithm 1 is globally convergent regardless of the particular choice of r_j^k. On the other hand, the superlinear convergence rate depends on whether r_j^k satisfies Assumption 4.2. Next, we explore suitable choices for the multiplier $\mu^k \in \Delta_m$ in (20) to ensure that r_j^k satisfies the aforementioned assumption. In what follows, we will assume that Assumption 4.1 holds. First note that, as in (34), we have $\eta_j^k = (y_j^k)^\top s^k/\|s^k\|^2 \geq \underline{L} > 0$ and hence

$$r_j^k = \max\{-\eta_j^k, 0\} + \vartheta_k \| \sum_{i=1}^m \mu_i^k \nabla F_i(x^k)\| = \vartheta_k \| \sum_{i=1}^m \mu_i^k \nabla F_i(x^k)\|, \quad \forall j = 1, \ldots, m, \quad \forall k \geq 0.$$

Choice 1: One natural choice is to set $\mu^k := \lambda^{SD}(x^k) \in \Delta_m$ for all $k \geq 0$, where $\lambda^{SD}(x^k)$ is the steepest descent Lagrange multiplier associated with x^k as in (16). In this case, by Lemma 2.3*(i)* and Lemma 2.3*(v)*, we have

$$r_j^k = \vartheta_k \|d_{SD}(x^k)\| = \vartheta_k(\|d_{SD}(x^k)\| - \|d_{SD}(x^*)\|) \leq \bar{\vartheta} L \|x^k - x^*\|, \quad \forall j = 1, \ldots, m, \quad \forall k \geq 0.$$

By summing this expression and making use of (32), we conclude that r_j^k satisfies Assumption 4.2 for all $j = 1, \ldots, m$. One potential drawback of this approach is the need to compute the multipliers $\lambda^{SD}(x^k)$, which involves solving the subproblem in (14).

Choice 2: Another natural choice is to set $\mu^k := \lambda^k \in \Delta_m$ for all $k \geq 0$, where λ^k is the Lagrange multiplier corresponding to the search direction d^k, see (11). Since the subproblem in Step 2 is typically solved in the form of (7) using a primal-dual algorithm, this approach does not require any additional computational cost. Let us assume that the sequences $\{B_j^k\}$ and $\{(B_j^k)^{-1}\}$ are bounded for all $j = 1, \ldots, m$. In this case, using [28, Lemma 6], there exists a constant $\delta > 0$ such that $\| \sum_{i=1}^m \lambda_i^k \nabla F_i(x^k)\| \leq \delta \|d_{SD}(x^k)\|$ for all $k \geq 0$. Therefore, for all $j = 1, \ldots, m$ and $k \geq 0$, similarly to the previous choice, we have

$$r_j^k = \vartheta_k \| \sum_{i=1}^m \lambda_i^k \nabla F_i(x^k)\| \leq \delta \bar{\vartheta} \|d_{SD}(x^k)\| = \delta \bar{\vartheta}(\|d_{SD}(x^k)\| - \|d_{SD}(x^*)\|) \leq \delta \bar{\vartheta} L \|x^k - x^*\|,$$

and hence r_j^k satisfies Assumption 4.2 for all $j = 1, \ldots, m$.

5 Numerical experiments

In this section, we present some numerical experiments to evaluate the effectiveness of the proposed scheme. We are particularly interested in verifying how the introduced modifications affect the numerical performance of the method. Toward this goal, we considered the following methods in our tests.

- Algorithm 1 (Global BFGS): our globally convergent algorithm with r_j^k chosen according to Choice 2 (see Section 4.2) and $\vartheta_k = 0.1$ for all $k \geq 0$. It is worth noting that preliminary numerical tests demonstrated the superior efficiency of Choice 2 over Choice 1.

- BFGS-Wolfe [46]: a BFGS algorithm in which the Hessian approximations are updated, for each $j = 1, \dots, m$, by

$$B_j^{k+1} := B_j^k - \frac{(\rho_j^k)^{-1} B_j^k s^k (s^k)^\top B_j^k + [(s^k)^\top B_j^k s^k] y_j^k (y_j^k)^\top}{\left[(\rho_j^k)^{-1} - (y_j^k)^\top s^k\right]^2 + (\rho_j^k)^{-1} (s^k)^\top B_j^k s^k}$$
$$+ \left[(\rho_j^k)^{-1} - (y_j^k)^\top s^k\right] \frac{y_j^k (s^k)^\top B_j^k + B_j^k s^k (y_j^k)^\top}{\left[(\rho_j^k)^{-1} - (y_j^k)^\top s^k\right]^2 + (\rho_j^k)^{-1} (s^k)^\top B_j^k s^k},$$

where

$$\rho_j^k := \begin{cases} 1/\left((y_j^k)^\top s^k\right), & \text{if } (y_j^k)^\top s^k > 0 \\ 1/\left(\mathcal{D}(x^{k+1}, s^k) - \nabla F_j(x^k)^\top s^k\right), & \text{otherwise.} \end{cases}$$

and the step sizes are calculated satisfying the Wolfe conditions (17)–(18). We point out that this algorithm is well-defined for nonconvex problems, although it is not possible to establish global convergence in this general case. Additionally, in the case of scalar optimization ($m = 1$), it retrieves the classical scalar BFGS algorithm.

- Cautious BFGS-Armijo [47]: a BFGS algorithm in which the Hessian approximations are updated, for each $j = 1, \dots, m$, by

$$B_j^{k+1} := \begin{cases} B_j^k - \dfrac{B_j^k s^k (s^k)^\top B_j^k}{(s^k)^\top B_j^k s^k} + \dfrac{y_j^k (y_j^k)^\top}{(y_j^k)^\top s^k}, & \text{if } (y_j^k)^\top s^k \geq \varepsilon \min\{1, |\theta(x^k)|\}, \\ B_j^k, & \text{otherwise,} \end{cases}$$

where $\varepsilon > 0$ is an algorithmic parameter and the step sizes are calculated satisfying the Armijo-type condition given in (17). In our experiments, we set $\varepsilon = 10^{-6}$. This combination also leads to a globally convergent scheme, see [47].

We implemented the algorithms using Fortran 90. The search directions $d(x^k)$ (see (5)) and optimal values $\theta(x^k)$ (see (6)) were obtained by solving subproblem (7) using the software Algencan [4]. To compute step sizes satisfying the Wolfe conditions (17)–(18), we employed the algorithm proposed in [38]. This algorithm utilizes quadratic/cubic polynomial interpolations of the objective functions, combining backtracking and extrapolation strategies, and is capable of finding step sizes in a finite number of iterations. Interpolation techniques were also used to calculate step sizes satisfying only the Armijo-type condition. We set $\rho = 10^{-4}$, $\sigma = 0.1$, and initialized B_j^0 as the identity matrix for all $j = 1, \dots, m$. Convergence was reported when $|\theta(x^k)| \leq 5 \times \texttt{eps}^{1/2}$, where $\texttt{eps} = 2^{-52} \approx 2.22 \times 10^{-16}$ represents the machine precision. When this criterion is met, we consider the problem successfully solved. The maximum number of allowed iterations was set to 2000. If this limit is reached, it means an unsuccessful termination. Our codes are freely available at `https://github.com/lfprudente/GlobalBFGS`.

The chosen set of test problems consists of both convex and nonconvex multiobjective problems commonly found in the literature and coincides with the one used in [46]. Table 1 presents their main characteristics: The first column contains the problem name, while the "n" and "m" columns provide the number of variables and objectives, respectively. The column "Conv." indicates whether the problem is convex or not. For each problem, the starting points were chosen within a box defined as $\{x \in \mathbb{R}^n \mid \ell \leq x \leq u\}$, where the lower and upper bounds, denoted by ℓ and $u \in \mathbb{R}^n$, are presented in the last two columns of Table 1. It is important to note that the boxes specified in the table were used solely for defining starting points and were not employed as constraints during the algorithmic processes. For detailed information regarding the references and corresponding formulations of each problem, we refer the reader to [46].

Problem	n	m	Conv.	ℓ	u
AP1	2	3	Y	$(-10, -10)$	$(10, 10)$
AP2	1	2	Y	-100	100
AP3	2	2	N	$(-100, -100)$	$(100, 100)$
AP4	3	3	Y	$(-10, -10, -10)$	$(10, 10, 10)$
BK1	2	2	Y	$(-5, -5)$	$(10, 10)$
DD1	5	2	N	$(-20, \ldots, -20)$	$(20, \ldots, 20)$
DGO1	1	2	N	-10	13
DGO2	1	2	Y	-9	9
DTLZ1	7	3	N	$(0, \ldots, 0)$	$(1, \ldots, 1)$
DTLZ2	7	3	N	$(0, \ldots, 0)$	$(1, \ldots, 1)$
DTLZ3	7	3	N	$(0, \ldots, 0)$	$(1, \ldots, 1)$
DTLZ4	7	3	N	$(0, \ldots, 0)$	$(1, \ldots, 1)$
FA1	3	3	N	$(0.01, 0.01, 0.01)$	$(1, 1, 1)$
Far1	2	2	N	$(-1, -1)$	$(1, 1)$
FDS	5	3	Y	$(-2, \ldots, -2)$	$(2, \ldots, 2)$
FF1	2	2	N	$(-1, -1)$	$(1, 1)$
Hil1	2	2	N	$(0, 0)$	$(1, 1)$
IKK1	2	3	Y	$(-50, -50)$	$(50, 50)$
IM1	2	2	N	$(1, 1)$	$(4, 2)$
JOS1	2	2	Y	$(-100, \ldots, -100)$	$(100, \ldots, 100)$
JOS4	20	2	N	$(-100, \ldots, -100)$	$(100, \ldots, 100)$
KW2	2	2	N	$(-3, -3)$	$(3, 3)$
LE1	2	2	N	$(1, 1)$	$(10, 10)$
Lov1	2	2	Y	$(-10, -10)$	$(10, 10)$
Lov2	2	2	N	$(-0.75, -0.75)$	$(0.75, 0.75)$
Lov3	2	2	N	$(-20, -20)$	$(20, 20)$
Lov4	2	2	N	$(-20, -20)$	$(20, 20)$
Lov5	3	2	N	$(-2, -2, -2)$	$(2, 2, 2)$
Lov6	6	2	N	$(0.1, -0.16, \ldots, -0.16)$	$(0.425, 0.16, \ldots, 0.16)$
LTDZ	3	3	N	$(0, 0, 0)$	$(1, 1, 1)$
MGH9	3	15	N	$(-2, -2, -2)$	$(2, 2, 2)$
MGH16	4	5	N	$(-25, -5, -5, -1)$	$(25, 5, 5, 1)$
MGH26	4	4	N	$(-1, -1, -1 - 1)$	$(1, 1, 1, 1)$
MGH33	10	10	Y	$(-1, \ldots, -1)$	$(1, \ldots, 1)$
MHHM2	2	3	Y	$(0, 0)$	$(1, 1)$
MLF1	1	2	N	0	20
MLF2	2	2	N	$(-100, -100)$	$(100, 100)$
MMR1	2	2	N	$(0.1, 0)$	$(1, 1)$
MMR2	2	2	N	$(0, 0)$	$(1, 1)$
MMR3	2	2	N	$(-\pi, -\pi)$	(π, π)
MMR4	3	2	N	$(0, 0, 0)$	$(4, 4, 4)$
MOP2	2	2	N	$(-4, -4)$	$(4, 4)$
MOP3	2	2	N	$(-\pi, -\pi)$	(π, π)
MOP5	2	3	N	$(-30, -30)$	$(30, 30)$
MOP6	2	2	N	$(0, 0)$	$(1, 1)$
MOP7	2	3	Y	$(-400, -400)$	$(400, 400)$
PNR	2	2	Y	$(-2, -2)$	$(2, 2)$
QV1	10	2	N	$(0.01, \ldots, 0.01)$	$(5, \ldots, 5)$
SD	4	2	Y	$(1, \sqrt{2}, \sqrt{2}, 1)$	$(3, 3, 3, 3)$
SK1	1	2	N	-100	100
SK2	4	2	N	$(-10, -10, -10, -10)$	$(10, 10, 10, 10)$
SLCDT1	2	2	N	$(-1.5, -1.5)$	$(1.5, 1.5)$
SLCDT2	10	3	Y	$(-1, \ldots, -1)$	$(1, \ldots, 1)$
SP1	2	2	Y	$(-100, -100)$	$(100, 100)$
SSFYY2	1	2	N	-100	100
TKLY1	4	2	N	$(0.1, 0, 0, 0)$	$(1, 1, 1, 1)$
Toi4	4	2	Y	$(-2, -2, -2, -2)$	$(5, 5, 5, 5)$
Toi8	3	3	Y	$(-1, -1, -1, -1)$	$(1, 1, 1, 1)$
Toi9	4	4	N	$(-1, -1, -1, -1)$	$(1, 1, 1, 1)$
Toi10	4	3	N	$(-2, -2, -2, -2)$	$(2, 2, 2, 2)$
VU1	2	2	N	$(-3, -3)$	$(3, 3)$
VU2	2	2	Y	$(-3, -3)$	$(3, 3)$
ZDT1	30	2	Y	$(0, \ldots, 0)$	$(1, \ldots, 1)$
ZDT2	30	2	N	$(0.01, \ldots, 0.01)$	$(1, \ldots, 1)$
ZDT3	30	2	N	$(0.01, \ldots, 0.01)$	$(1, \ldots, 1)$
ZDT4	30	2	N	$(0.01, -5, \ldots, -5)$	$(1, 5, \ldots, 5)$
ZDT6	10	2	N	$(0.01, \ldots, 0.01)$	$(1, \ldots, 1)$
ZLT1	10	5	Y	$(-1000, \ldots, -1000)$	$(1000, \ldots, 1000)$

Table 1: List of test problems.

In multiobjective optimization, the primary objective is to estimate the Pareto frontier of a given problem. A commonly used strategy is to execute the algorithm from multiple distinct starting points and collect the Pareto optimal points found. Thus, each problem listed in Table 1 was addressed by running all algorithms from 300 randomly generated starting points within their respective boxes. In this first stage, each problem/starting point was considered an independent instance and a run was considered successful if an approximate critical point was found, regardless of the objective functions values. Figure 1 presents the comparison of the algorithms in terms of CPU time using a performance profile [18]. As can be seen, Algorithm 1 and the BFGS-Wolfe algorithm exhibited virtually identical performance, outperforming the Cautious BFGS-Armijo algorithm. All methods proved to be robust, successfully solving more than 98% of the problem instances. It is worth noting that although the BFGS-Wolfe algorithm enjoys (theoretical) global convergence only under convexity assumptions, it also performs exceptionally well for nonconvex problems, which is consistent with observations in the scalar case.

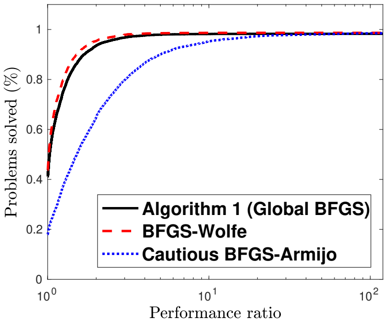
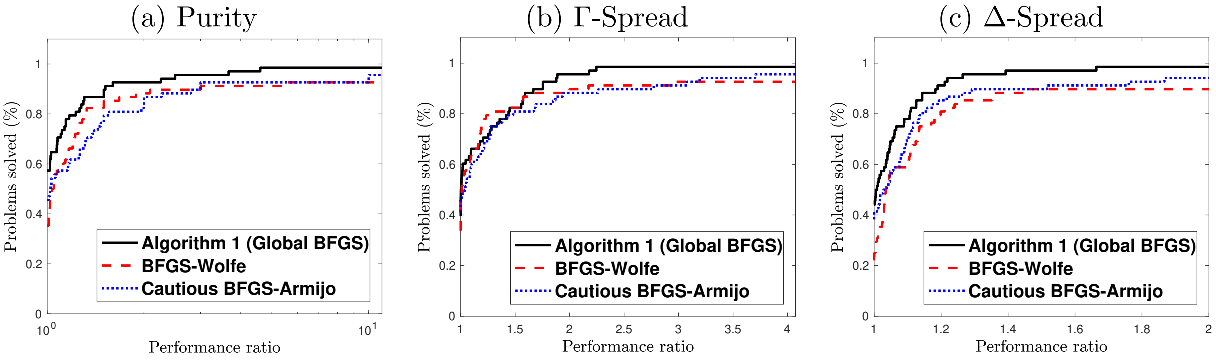


Figure 1: Performance profiles considering 300 starting points for each test problem using the CPU time as performance measurement.

In the following, we evaluate the algorithms based on their ability to properly generate Pareto frontiers. To assess this, we employ the widely recognized *Purity* and (Γ and Δ) *Spread* metrics. In summary, the Purity metric measures the solver's ability to identify points on the Pareto frontier, while the Spread metric evaluates the distribution quality of the obtained Pareto frontier. For a detailed explanation of these metrics and their application together with performance profiles, we refer the reader to [14]. It is important to note that, at this stage, data referring to all starting points are combined for each problem, taking into account the objective function values found. The results in Figure 2 indicate that Algorithm 1 performed slightly better in terms of the Purity and Δ-Spread metrics, with no significant difference observed for the Γ-Spread metric among the three algorithms.


(a) Purity

(b) Γ-Spread

(c) Δ-Spread

Figure 2: Metric performance profiles: (a) Purity; (b) Γ-Spread; (c) Δ-Spread.

The numerical results allow us to conclude that the modifications made to the BFGS method to ensure global convergence for nonconvex problems do not compromise its practical performance.

6 Final remarks

Based on the work of Li and Fukushima [35], we presented a modified BFGS scheme that achieves global convergence without relying on convexity assumptions for the objective function F. The global convergence analysis depends only on the requirement of F having continuous Lipschitz gradients. Furthermore, we showed that by appropriately selecting r_j^k to satisfy Assumption 4.2 and under suitable conditions, the rate of convergence becomes superlinear. We also discussed some practical choices for r_j^k. The introduced modifications preserve the simplicity and practical efficiency of the BFGS method. It is worth emphasizing that the assumptions considered in our approach are natural extensions of those commonly employed in the context of scalar-valued optimization.

Data availability statement
The codes supporting the numerical experiments are freely available in the Github repository, `https://github.com/lfprudente/GlobalBFGS`.

Conflicts of interest
The authors declare that they have no conflict of interest.

A Appendix

In the main body of the article, we have chosen to exclude proofs that can be readily derived from existing sources in order to enhance the overall readability of the text. However, in this appendix, we provide these proofs to ensure self-contained completeness.

Notation. The cardinality of a set C is denoted by $|C|$. The ceiling and floor functions are denoted by $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$, respectively; i.e., if $x \in \mathbb{R}$, then $\lceil x\rceil$ is the least integer greater than or equal to x and $\lfloor x\rfloor$ is the greatest integer less than or equal to x. The notation $\varphi(t) := o(t)$ for $t > 0$ means that $\lim_{t\to 0}\varphi(t)/t = 0$.

A.1 Proofs of Section 3

Throughout this section, we assume that Assumption 3.1 holds.

Proof of Proposition 3.2. It follows from (18), the definition of $\mathcal{D}(\cdot,\cdot)$, the Cauchy-Schwarz inequality, and Assumption 3.1*(iii)* that

$$\begin{aligned}-(1-\sigma)\mathcal{D}(x^k,d^k) &\leq \mathcal{D}(x^{k+1},d^k) - \mathcal{D}(x^k,d^k)\\ &\leq \max_{j=1,\ldots,m}\Big(\nabla F_j(x^{k+1}) - \nabla F_j(x^k)\Big)^\top d^k\\ &\leq \max_{j=1,\ldots,m}\|\nabla F_j(x^{k+1}) - \nabla F_j(x^k)\|\|d^k\| \leq L\alpha_k\|d^k\|^2,\end{aligned}$$

where the second inequality follows from the fact that, for any $u, v \in \mathbb{R}^m$, we have $\max_j(u_j - v_j) \geq \max_j u_j - \max_j v_j$. Hence,

$$\frac{\mathcal{D}(x^k,d^k)^2}{\|d^k\|^2} \leq -\frac{L}{(1-\sigma)}\alpha_k\mathcal{D}(x^k,d^k). \tag{45}$$

Now, since $\{x^k\} \subset \mathcal{L}$, Assumption 3.1*(i)–(ii)* implies the existence of $\mathcal{F} \in \mathbb{R}$ such that $F_j(x^k) \geq \mathcal{F}$ for all $k \geq 0$ and $j = 1, \ldots, m$. Therefore, by (17), we have

$$\mathcal{F} \leq F_j(x^{k+1}) \leq F_j(x^0) + \rho \sum_{\ell=0}^{k} \alpha_\ell \mathcal{D}(x^\ell, d^\ell), \quad \forall j = 1, \ldots, m.$$

Some algebraic manipulations yields

$$-\frac{L}{\rho(1-\sigma)} \min_{j=1,\ldots,m} \left\{ \mathcal{F} - F_j(x^0) \right\} \geq -\frac{L}{(1-\sigma)} \sum_{\ell=0}^{k} \alpha_\ell \mathcal{D}(x^\ell, d^\ell) > 0.$$

Therefore,

$$-\frac{L}{(1-\sigma)} \sum_{k \geq 0} \alpha_k \mathcal{D}(x^k, d^k) < \infty,$$

which together with (45) gives (26). □

To prove Proposition 3.3, we will make use of function (35). Let us define

$$q_j^k := \frac{(s^k)^\top B_j^k s^k}{(s^k)^\top s^k}, \quad \forall j = 1, \ldots, m.$$

Thus, from the same arguments that led to (36), we obtain

$$\psi(B_j^{k+1}) = \psi(B_j^k) + \left[\frac{\|\gamma_j^k\|^2}{(\gamma_j^k)^\top s^k} - \ln\left(\frac{(\gamma_j^k)^\top s^k}{(s^k)^\top s^k} \right) - 1 \right] - \xi_j^k, \tag{46}$$

where

$$\xi_j^k := -\ln(\cos^2 \beta_j^k) - \left[1 - \frac{q_j^k}{\cos^2 \beta_j^k} + \ln\left(\frac{q_j^k}{\cos^2 \beta_j^k} \right) \right].$$

Note from Lemma 2.4*(i)* that $\xi_j^k \geq 0$.

Proof of Proposition 3.3. Let $k \geq 1$ and $p \in (0,1)$ be given and set $\varepsilon := 1-p$ and $\bar{p} := 1-\varepsilon/m$. Let $j \in \{1, \ldots, m\}$ be an arbitrary index. From (46) and (27), we have

$$\psi(B_j^{k+1}) \leq \psi(B_j^0) + \left[C_2 - \ln(C_1) - 1 \right](k+1) - \sum_{\ell=0}^{k} \xi_j^\ell.$$

Therefore, since $\psi(B_j^{k+1}) > 0$, we obtain

$$\frac{1}{k+1} \sum_{\ell=0}^{k} \xi_j^\ell \leq \frac{\psi(B_j^0)}{k+1} + \left[C_2 - \ln(C_1) - 1 \right].$$

Let $\mathcal{J}_j^k$ be the set consisting of the $\lceil \bar{p}(k+1) \rceil$ indices corresponding to the $\lceil \bar{p}(k+1) \rceil$ smallest values of ξ_j^ℓ, for $\ell \leq k$, and define $\bar{\xi}_j^k := \max_{\ell \in \mathcal{J}_j^k} \xi_j^\ell$. Then,

$$\frac{1}{k+1} \sum_{\ell=0}^{k} \xi_j^\ell \geq \frac{1}{k+1} \left[\bar{\xi}_j^k + \sum_{\ell=0, \ell \notin \mathcal{J}_j^k}^{k} \xi_j^\ell \right] \geq \frac{1}{k+1} \left[\bar{\xi}_j^k + \bar{\xi}_j^k (k+1 - \lceil \bar{p}(k+1) \rceil) \right] \geq \bar{\xi}_j^k (1 - \bar{p}),$$

where the last inequality is due to $\lceil \bar{p}(k+1) \rceil \leq \bar{p}(k+1) + 1$. By combining the above two inequalities, we get, for all $\ell \in \mathcal{J}_j^k$,

$$\xi_j^\ell \leq \bar{\xi}_j^k \leq \frac{1}{1-\bar{p}} \Big[\psi(B_j^0) + C_2 - \ln(C_1) - 1 \Big] =: \zeta_j.$$

Therefore, by the definition of ξ_j^ℓ, we obtain, for all $\ell \in \mathcal{J}_j^k$,

$$-\ln(\cos^2 \beta_j^\ell) \leq \xi_j^\ell \leq \zeta_j,$$

and hence

$$\cos \beta_j^\ell \geq e^{-\zeta_j/2} =: \delta_j.$$

This means that $\cos \beta_j^\ell \geq \delta_j$ for at least $\lceil \bar{p}(k+1) \rceil$ values of $\ell \in \{0, 1, \ldots, k\}$.

Now, let us define $\delta := \min_{j=1,\ldots,m} \delta_j$ and, for all $j = 1, \ldots, m$,

$$\mathcal{G}_j^k := \{\ell \in \{0, 1, \ldots, k\} \mid \cos \beta_j^\ell \geq \delta\} \quad \text{and} \quad \mathcal{B}_j^k := \{\ell \in \{0, 1, \ldots, k\} \mid \cos \beta_j^\ell < \delta\}.$$

It is easy to see that $\mathcal{J}_j^k \subset \mathcal{G}_j^k$, $\mathcal{G}_j^k \cap \mathcal{B}_j^k = \emptyset$ and $|\mathcal{G}_j^k| + |\mathcal{B}_j^k| = k+1$. Therefore, by the definition of $\bar{p}$ and using some properties of the ceiling and floor functions, we have, for all $j = 1, \ldots, m$,

$$|\mathcal{G}_j^k| \geq |\mathcal{J}_j^k| = \lceil \bar{p}(k+1) \rceil = (k+1) + \lceil -\frac{\varepsilon}{m}(k+1) \rceil = (k+1) - \lfloor \frac{\varepsilon}{m}(k+1) \rfloor,$$

and hence $|\mathcal{B}_j^k| \leq \lfloor \frac{\varepsilon}{m}(k+1) \rfloor$. Thus,

$$|\bigcup_{j=1}^m \mathcal{B}_j^k| \leq m \lfloor \frac{\varepsilon}{m}(k+1) \rfloor \leq \varepsilon(k+1).$$

As consequence, since we also have $|\cap_{j=1}^m \mathcal{G}_j^k| + |\cup_{j=1}^m \mathcal{B}_j^k| = k+1$, by using the definition of ε, it follows that

$$|\bigcap_{j=1}^m \mathcal{G}_j^k| \geq (k+1) - \varepsilon(k+1) = (1-\varepsilon)(k+1) = p(k+1),$$

which concludes the proof. □

A.2 Proofs of Section 4

In this section, we make use of Assumption 4.1. In particular, and without loss of generality, we assume that $\{x^k\} \subset U$, where U is a neighborhood of x^* such that (30) and (31) hold.

A.2.1 Proof of Proposition 4.1

We start with some auxiliary technical results.

Lemma A.1. *Suppose that Assumption 4.1 holds. Let β_j^k be the angle between the vectors s^k and $B_j^k s^k$, for all $k \geq 0$ and $j = 1, \ldots, m$. Then, for all $k \geq 0$,*

$$\mathcal{D}(x^k, d^k) \leq -\frac{\delta_k}{2} \|d^k\| \|d_{SD}(x^k)\|,$$

where $\delta_k := \min_{j=1,\ldots,m} \cos \beta_j^k$.

Proof. For a given $k \geq 0$, by using the definitions of δ_k, $\cos\beta_j^k$, and s^k, we obtain

$$\delta_k \leq \cos\beta_j^k = \frac{(s^k)^\top B_j^k s^k}{\|s^k\|\|B_j^k s^k\|} = \frac{(d^k)^\top B_j^k d^k}{\|d^k\|\|B_j^k d^k\|}, \quad \forall j = 1, \ldots, m.$$

Therefore, from Lemma 2.2*(ii)* and (13), we have

$$-\mathcal{D}(x^k, d^k) > -\theta(x^k) = \frac{1}{2}\sum_{j=1}^{m} \lambda_j^k (d^k)^\top B_j^k d^k \geq \frac{\delta_k}{2}\|d^k\| \sum_{j=1}^{m} \lambda_j^k \|B_j^k d^k\|.$$

Applying the triangle inequality, together with (11), (12), and Lemma 2.3*(iv)*, we obtain:

$$-\mathcal{D}(x^k, d^k) \geq \frac{\delta_k}{2}\|d^k\|\|\sum_{j=1}^{m} \lambda_j^k B_j^k d^k\| = \frac{\delta_k}{2}\|d^k\|\|\sum_{j=1}^{m} \lambda_j^k \nabla F_j(x^k)\| \geq \frac{\delta_k}{2}\|d^k\|\|d_{SD}(x^k)\|.$$

□

Lemma A.2. *Suppose that Assumption 4.1 holds. Then, for all $k \geq 0$, we have:*

(i) $\|x^k - x^*\| \leq \dfrac{2}{\underline{L}}\|d_{SD}(x^k)\|$*;*

(ii) $\|s^k\| \geq \dfrac{(1-\sigma)}{2L}\delta_k\|d_{SD}(x^k)\|$*, where δ_k is given as in Lemma A.1;*

(iii) $\dfrac{(\gamma_j^k)^\top s^k}{\|s^k\|^2} \geq \underline{L}$*, for all $j = 1, \ldots, m$;*

(iv) $\dfrac{\|\gamma_j^k\|^2}{(\gamma_j^k)^\top s^k} \leq \dfrac{(2L + \bar{\vartheta}\bar{c})^2}{\underline{L}}$*, for all $j = 1, \ldots, m$ and some constant $\bar{c} > 0$.*

Proof. Consider part *(i)*. For a given value of $k \geq 0$, consider $\lambda^{SD}(x^k) \in \mathbb{R}^m$ as in (15)–(16), and define the scalar-valued function $F_{SD}\colon \mathbb{R}^n \to \mathbb{R}$ as follows:

$$F_{SD}(x) \coloneqq \sum_{j=1}^{m} \lambda_j^{SD}(x^k) F_j(x).$$

Therefore, by taking $z \coloneqq x^* - x^k$, it follows from (15) and (31) that

$$\int_0^1 (1-\tau) z^\top \nabla^2 F_{SD}(x^k + \tau z) z d\tau \geq \frac{\underline{L}}{2}\|z\|^2.$$

Evaluating this integral (which can be done by integration by parts), and considering that $d_{SD}(x^k) = -\nabla F_{SD}(x^k)$, we obtain

$$F_{SD}(x^*) - F_{SD}(x^k) + d_{SD}(x^k)^\top (x^* - x^k) \geq \frac{\underline{L}}{2}\|x^* - x^k\|^2.$$

Given that $F_j(x^*) \leq F_j(x^k)$ for all $j = 1, \ldots, m$, we have $F_{SD}(x^*) - F_{SD}(x^k) \leq 0$ and thus

$$\frac{\underline{L}}{2}\|x^* - x^k\|^2 \leq d_{SD}(x^k)^\top (x^* - x^k) \leq \|d_{SD}(x^k)\|\|x^* - x^k\|,$$

which proves part *(i)*.

Consider part *(ii)*. By using (18) and the definitions of $\mathcal{D}(\cdot,\cdot)$ and y_j^k, we obtain

$$-(1-\sigma)\mathcal{D}(x^k,d^k) \le \mathcal{D}(x^{k+1},d^k) - \mathcal{D}(x^k,d^k) \le \max_{j=1,\dots,m}(y_j^k)^\top d^k,$$

which, together with (33), yields

$$-(1-\sigma)\mathcal{D}(x^k,d^k) \le \max_{j=1,\dots,m}(s^k)^\top \bar{G}_j^k d^k = \alpha_k \max_{j=1,\dots,m}(d^k)^\top \bar{G}_j^k d^k \le L\alpha_k\|d^k\|^2 = L\|s^k\|\|d^k\|,$$

where the latter inequality comes from (31). Therefore, taking into account that $\sigma < 1$, by Lemma A.1, we obtain

$$(1-\sigma)\frac{\delta_k}{2}\|d^k\|\|d_{SD}(x^k)\| \le L\|s^k\|\|d^k\|,$$

which gives the desired inequality.

Part *(iii)* is a direct consequence of (34). Finally, consider part *(iv)*. From (19) and (31), we have

$$|\eta_j^k| \le \frac{\|y_j^k\|}{\|s^k\|} = \frac{\|\nabla F_j(x^{k+1}) - \nabla F_j(x^k)\|}{\|x^{k+1}-x^k\|} \le L.$$

Furthermore, since $\{x^k\} \subset U$, by (20) and using continuity arguments, there exists a constant $\bar{c} > 0$ such that

$$0 \le r_j^k \le |\eta_j^k| + \vartheta_k\|\sum_{i=1}^m \mu_i^k \nabla F_i(x^k)\| \le L + \bar{\vartheta}\bar{c},$$

and hence, by (21),

$$\|\gamma_j^k\| \le \|y_j^k\| + r_j^k\|s^k\| = \left(\frac{\|y_j^k\|}{\|s^k\|} + r_j^k\right)\|s^k\| \le (2L+\bar{\vartheta}\bar{c})\|s^k\|.$$

Therefore, using the inequality in part *(iii)*, we obtain

$$\frac{\|\gamma_j^k\|^2}{(\gamma_j^k)^\top s^k} = \frac{\|\gamma_j^k\|^2}{\|s^k\|^2}\frac{\|s^k\|^2}{(\gamma_j^k)^\top s^k} \le \frac{(2L+\bar{\vartheta}\bar{c})^2}{\underline{L}}, \quad \forall j = 1,\dots,m,$$

concluding the proof. □

We are now able to prove Proposition 4.1

Proof of Proposition 4.1. Let $\lambda^{SD}(x^*) \in \mathbb{R}^m$ be a steepest descent multiplier associated with x^* as in (15)–(16), and define the scalar-valued function $F_*\colon \mathbb{R}^n \to \mathbb{R}$ as follows:

$$F_*(x) := \sum_{j=1}^m \lambda_j^{SD}(x^*)F_j(x).$$

Note that

$$\nabla F_*(x^*) = \sum_{j=1}^m \lambda_j^{SD}(x^*)\nabla F_j(x^*) = -d_{SD}(x^*) = 0, \tag{47}$$

where the last equality comes from Lemma 2.3*(i)*. Now, by using (31), we obtain

$$\nabla F_j(x^*)^\top(x^k-x^*) + \frac{\underline{L}}{2}\|x^k-x^*\|^2 \le F_j(x^k) - F_j(x^*) \le \nabla F_j(x^*)^\top(x^k-x^*) + \frac{L}{2}\|x^k-x^*\|^2,$$

for all $j = 1, \ldots, m$ and for all $k \geq 0$. By multiplying this expression by $\lambda_j^{SD}(x^*)$, summing over all indices $j = 1, \ldots, m$, and taking into account (15) and (47), we obtain

$$\frac{\underline{L}}{2}\|x^k - x^*\|^2 \leq F_*(x^k) - F_*(x^*) \leq \frac{L}{2}\|x^k - x^*\|^2, \quad \forall k \geq 0. \tag{48}$$

From the right hand side of (48) and Lemma A.2*(i)*, we obtain

$$F_*(x^k) - F_*(x^*) \leq \frac{2L}{\underline{L}^2}\|d_{SD}(x^k)\|^2, \quad \forall k \geq 0. \tag{49}$$

On the other hand, (17) gives

$$F_*(x^{k+1}) \leq F_*(x^k) + \rho\alpha_k\mathcal{D}(x^k, d^k), \quad \forall k \geq 0.$$

Therefore, from Lemma A.1 and Lemma A.2*(ii)*, we have

$$F_*(x^{k+1}) \leq F_*(x^k) - \frac{\rho}{2}\delta_k\|s^k\|\|d_{SD}(x^k)\| \leq F_*(x^k) - \frac{\rho(1-\sigma)}{4L}\delta_k^2\|d_{SD}(x^k)\|^2, \quad \forall k \geq 0.$$

Hence, by subtracting the term $F_*(x^*)$ in both sides of the latter inequality, and using (49), we obtain

$$F_*(x^{k+1}) - F_*(x^*) \leq \left(1 - \frac{\rho(1-\sigma)\underline{L}^2}{8L^2}\delta_k^2\right)\Big(F_*(x^k) - F_*(x^*)\Big), \quad \forall k \geq 0. \tag{50}$$

For each $k \geq 0$, define $\bar{r}_k := 1 - \rho(1-\sigma)\underline{L}^2\delta_k^2/(8L^2)$. It is easy to see that $\bar{r}_k \in (0, 1]$, for all $k \geq 0$.

Now, given $p \in (0, 1)$, we can invoke Lemma A.2*(iii)–(iv)* to apply Proposition 3.3. This implies that there exists a constant $\delta > 0$ such that, for any $k \geq 1$, the number of elements $\ell \in \{0, 1, \ldots, k\}$ for which $\delta_\ell \geq \delta$ is at least $\lceil p(k+1) \rceil$. By defining $\mathcal{G}_k := \{\ell \in \{0, 1, \ldots, k\} \mid \delta_\ell \geq \delta\}$, we have $|\mathcal{G}_k| \geq \lceil p(k+1) \rceil$ and

$$\bar{r}_\ell \leq 1 - \frac{\rho(1-\sigma)\underline{L}^2\delta^2}{8L^2} := \bar{r} < 1, \quad \forall \ell \in \mathcal{G}_k.$$

Thus, from (50) and considering that $F_*(x^0) - F_*(x^*) > 0$, we obtain, for all $k \geq 1$,

$$\begin{aligned} F_*(x^{k+1}) - F_*(x^*) &\leq \left[\prod_{\ell=0}^{k}\bar{r}_\ell\right]\Big(F_*(x^0) - F_*(x^*)\Big) \leq \left[\prod_{\ell\in\mathcal{G}_k}\bar{r}_\ell\right]\Big(F_*(x^0) - F_*(x^*)\Big) \\ &\leq \left[\prod_{\ell\in\mathcal{G}_k}\bar{r}\right]\Big(F_*(x^0) - F_*(x^*)\Big) \leq \bar{r}^{\lceil p(k+1)\rceil}\Big(F_*(x^0) - F_*(x^*)\Big), \end{aligned}$$

where the second inequality follows from the fact that $\bar{r}_\ell \leq 1$ for all $\ell \notin \mathcal{G}_k$. Therefore, by taking $r := \bar{r}^p$, we obtain

$$F_*(x^{k+1}) - F_*(x^*) \leq r^{k+1}\Big(F_*(x^0) - F_*(x^*)\Big), \quad \forall k \geq 1.$$

Combining this with the left hand side of (48), we find

$$\|x^{k+1} - x^*\|^\nu \leq \left[\frac{2}{\underline{L}}\Big(F_*(x^0) - F_*(x^*)\Big)\right]^{\nu/2}(r^{\nu/2})^{k+1}.$$

Finally, by summing this expression and taking into account that $r < 1$, we conclude that (32) holds. □

A.2.2 Proof of Theorem 4.3

We start by introducing an auxiliary result.

Lemma A.3. *Suppose that Assumptions 4.1 and 4.2 hold. Then, there exists $\bar{a} > 0$ such that*

$$|\theta(x^k)| \geq \bar{a}\|d^k\|^2, \tag{51}$$

for all k sufficiently large. Moreover,

$$\lim_{k\to\infty} \|d^k\| = 0. \tag{52}$$

Proof. By choosing $\gamma \in (0,1)$ and recalling that $s^k = \alpha_k d^k$, it follows from (37) that

$$\frac{(d^k)^\top B_j^k d^k}{(d^k)^\top \nabla^2 F_j(x^*) d^k} \geq 1 - \gamma, \quad \forall j = 1, \ldots, m,$$

for all k sufficiently large. Thus, by (31), we obtain

$$(d^k)^\top B_j^k d^k \geq \underline{L}(1-\gamma)\|d^k\|^2, \quad \forall j = 1, \ldots, m,$$

for all k sufficiently large. Therefore, using (12) and (13), we have

$$|\theta(x^k)| = \frac{1}{2}\sum_{j=1}^m \lambda_j^k (d^k)^\top B_j^k d^k \geq \frac{\underline{L}(1-\gamma)}{2}\|d^k\|^2,$$

for all k sufficiently large. Defining $\bar{a} := \underline{L}(1-\gamma)/2$, we establish (51). Finally, by combining (51), Lemma 2.2*(ii)*, and Proposition 3.2, we obtain

$$0 \leq \lim_{k\to\infty} \bar{a}\|d^k\| \leq \lim_{k\to\infty} \frac{|\theta(x^k)|}{\|d^k\|} \leq \lim_{k\to\infty} \frac{|\mathcal{D}(x^k, d^k)|}{\|d^k\|} = 0,$$

which concludes the proof. □

Recalling that $\lambda^k \in \mathbb{R}^m$ is the Lagrange multiplier associated to x^k of problem (7) fulfilling (11)–(12), let us define

$$F_\lambda^k(x) := \sum_{j=1}^m \lambda_j^k F_j(x) \quad \text{and} \quad B_\lambda^k := \sum_{j=1}^m \lambda_j^k B_j^k, \quad \forall k \geq 0. \tag{53}$$

Next, we show that the sequence of functions $\{F_\lambda^k(x)\}_{k\geq 0}$ fulfills a Dennis–Moré-type condition.

Theorem A.4. *Suppose that Assumptions 4.1 and 4.2 hold. For each $k \geq 0$, consider $F_\lambda^k \colon \mathbb{R}^n \to \mathbb{R}$ and B_λ^k as in* (53)*.Then,*

$$\lim_{k\to\infty} \frac{\|(B_\lambda^k - \nabla^2 F_\lambda^k(x^*))d^k\|}{\|d^k\|} = 0 \tag{54}$$

or, equivalently,

$$\lim_{k\to\infty} \frac{\|\nabla F_\lambda^k(x^k) + \nabla^2 F_\lambda^k(x^k) d^k\|}{\|d^k\|} = 0. \tag{55}$$

Proof. By (53) and taking into account (12), we have

$$\begin{aligned}\lim_{k\to\infty}\frac{\|(B_\lambda^k-\nabla^2F_\lambda^k(x^*))d^k\|}{\|d^k\|}&\le\lim_{k\to\infty}\sum_{j=1}^m\lambda_j^k\frac{\|(B_j^k-\nabla^2F_j(x^*))d^k\|}{\|d^k\|}\\&\le\lim_{k\to\infty}\max_{j=1,\dots,m}\frac{\|(B_j^k-\nabla^2F_j(x^*))d^k\|}{\|d^k\|},\end{aligned}$$

which, combined with (38), yields (54). We proceed to show that (54) implies (55). Firstly, considering (11), since $B_\lambda^kd^k=-\nabla F_\lambda^k(x^k)$, it follows that (55) is equivalent to

$$\lim_{k\to\infty}\frac{\|(B_\lambda^k-\nabla^2F_\lambda^k(x^k))d^k\|}{\|d^k\|}=0. \tag{56}$$

Note that

$$\lim_{k\to\infty}\frac{\|(B_\lambda^k-\nabla^2F_\lambda^k(x^k))d^k\|}{\|d^k\|}\le\lim_{k\to\infty}\frac{\|(B_\lambda^k-\nabla^2F_\lambda^k(x^*))d^k\|}{\|d^k\|}+\lim_{k\to\infty}\|\nabla^2F_\lambda^k(x^*)-\nabla^2F_\lambda^k(x^k)\|$$

and, by using continuity arguments,

$$\begin{aligned}\lim_{k\to\infty}\|\nabla^2F_\lambda^k(x^*)-\nabla^2F_\lambda^k(x^k)\|&\le\lim_{k\to\infty}\sum_{j=1}^m\lambda_j^k\|\nabla^2F_j(x^*)-\nabla^2F_j(x^k)\|\\&\le\lim_{k\to\infty}\max_{j=1,\dots,m}\|\nabla^2F_j(x^*)-\nabla^2F_j(x^k)\|=0.\end{aligned}$$

Therefore, combining the two latter inequalities, we obtain (56). The proof that (55) implies (54) can be obtained similarly. □

The following result shows that the unit step size eventually satisfies the Wolfe conditions (17)–(18).

Theorem A.5. *Suppose that Assumptions 4.1 and 4.2 hold. Then, the step size $\alpha_k=1$ is admissible for all k sufficiently large.*

Proof. Let $j\in\{1,\dots,m\}$ be an arbitrary index. It is easy to see that (38) is equivalent to

$$\lim_{k\to\infty}\frac{\|(B_j^k-\nabla^2F_j(x^k))d^k\|}{\|d^k\|}=0.$$

Thus, by Taylor's theorem, it follows that

$$\begin{aligned}F_j(x^k+d^k)=&F_j(x^k)+\nabla F_j(x^k)^\top d^k+\frac{1}{2}(d^k)^\top B_j^kd^k+\frac{1}{2}(d^k)^\top\left(\nabla^2F_j(x^k)-B_j^k\right)d^k+o(\|d^k\|^2)\\=&F_j(x^k)+\nabla F_j(x^k)^\top d^k+\frac{1}{2}(d^k)^\top B_j^kd^k+o(\|d^k\|^2),\end{aligned}$$

Therefore, by using (6) and setting $t:=2\rho<1$, we have

$$F_j(x^k+d^k)\le F_j(x^k)+t\theta(x^k)+(1-t)\theta(x^k)+o(\|d^k\|^2).$$

Consequently, according to (51), for sufficiently large k,

$$F_j(x^k + d^k) \leq F_j(x^k) + t\theta(x^k) + \left[-\bar{a}(1-t) + \frac{o(\|d^k\|^2)}{\|d^k\|^2} \right] \|d^k\|^2.$$

As the term in square brackets is negative for k large enough, we conclude that

$$F_j(x^k + d^k) \leq F_j(x^k) + t\theta(x^k).$$

On the other hand, combining (11)–(13), we find

$$\theta(x^k) = \frac{1}{2} \sum_{j=1}^{m} \lambda_j^k \nabla F_j(x^k)^\top d^k \leq \frac{1}{2} \mathcal{D}(x^k, d^k). \tag{57}$$

Hence, from the last two inequalities and the definition of t, we obtain

$$F_j(x^k + d^k) \leq F_j(x^k) + \rho \mathcal{D}(x^k, d^k),$$

for all k sufficiently large. Given the arbitrary choice of $j \in \{1, \dots, m\}$, we conclude that the step size $\alpha_k = 1$ satisfies (17) for all sufficiently large k.

Consider the curvature condition (18). From the definition of F_λ^k in (53), we have

$$\begin{aligned} -\sum_{j=1}^{m} \lambda_j^k \nabla F_j(x^k)^\top d^k &= \sum_{j=1}^{m} \lambda_j^k (d^k)^\top \nabla^2 F_j(x^k) d^k - \sum_{j=1}^{m} \lambda_j^k \left[\nabla^2 F_j(x^k) d^k + \nabla F_j(x^k) \right]^\top d^k \\ &= \sum_{j=1}^{m} \lambda_j^k (d^k)^\top \nabla^2 F_j(x^k) d^k - \left[\nabla F_\lambda^k(x^k) + \nabla^2 F_\lambda^k(x^k) d^k \right]^\top d^k. \end{aligned}$$

Thus, by (12), (31), and (55), we obtain

$$-\sum_{j=1}^{m} \lambda_j^k \nabla F_j(x^k)^\top d^k \geq \underline{L} \|d^k\|^2 + o(\|d^k\|^2) = \|d^k\|^2 \left[\underline{L} + \frac{o(\|d^k\|^2)}{\|d^k\|^2} \right].$$

Hence, taking into account (52) and (57), for k sufficiently large, it follows that

$$-2\theta(x^k) = -\sum_{j=1}^{m} \lambda_j^k \nabla F_j(x^k)^\top d^k \geq \frac{\underline{L}}{2} \|d^k\|^2. \tag{58}$$

On the other hand, applying the Mean Value Theorem to the scalar function $\nabla F_\lambda^k(\cdot)^\top d^k$, there exists $v^k := x^k + t_k d^k$ for some $t_k \in (0,1)$ such that

$$\nabla F_\lambda^k(x^k + d^k)^\top d^k = \nabla F_\lambda^k(x^k)^\top d^k + (d^k)^\top \nabla^2 F_\lambda^k(v^k) d^k.$$

Therefore,

$$\frac{|\nabla F_\lambda^k(x^k + d^k)^\top d^k|}{\|d^k\|^2} \leq \frac{\|\nabla F_\lambda^k(x^k) + \nabla^2 F_\lambda^k(x^k) d^k\|}{\|d^k\|} + \|\nabla^2 F_\lambda^k(v^k) - \nabla^2 F_\lambda^k(x^k)\|.$$

Now, by the definitions of F_λ^k and v^k, and considering (12) and (52), we obtain

$$\begin{aligned}\lim_{k\to\infty}\|\nabla^2 F_\lambda^k(v^k)-\nabla^2 F_\lambda^k(x^k)\| &\le \lim_{k\to\infty}\sum_{j=1}^m \lambda_j^k\|\nabla^2 F_j(x^k+t_k d^k)-\nabla^2 F_j(x^k)\|\\ &\le \lim_{k\to\infty}\max_{j=1,\dots,m}\|\nabla^2 F_j(x^k+t_k d^k)-\nabla^2 F_j(x^k)\| = 0.\end{aligned}$$

Thus, combining the latter two inequalities with (55), we have

$$\lim_{k\to\infty}\frac{|\nabla F_\lambda^k(x^k+d^k)^\top d^k|}{\|d^k\|^2}=0.$$

Hence, for k large enough, we have

$$|\nabla F_\lambda^k(x^k+d^k)^\top d^k| \le \sigma\frac{L}{4}\|d^k\|^2,$$

which, together with (58), yields

$$\sum_{j=1}^m \lambda_j^k \nabla F_j(x^k+d^k)^\top d^k = \nabla F_\lambda^k(x^k+d^k)^\top d^k \ge -\sigma\frac{L}{4}\|d^k\|^2 \ge \sigma\theta(x^k).$$

Therefore, by the definition of $\mathcal{D}(\cdot,\cdot)$, (12), and Lemma 2.2*(ii)*, we obtain

$$\mathcal{D}(x^k+d^k,d^k) \ge \sum_{j=1}^m \lambda_j^k \nabla F_j(x^k+d^k)^\top d^k \ge \sigma\theta(x^k) \ge \sigma\mathcal{D}(x^k,d^k),$$

for all k sufficiently large, concluding the proof. □

We require an additional auxiliary result.

Lemma A.6. *Suppose that Assumption 4.1 holds. Then,*

$$\|\nabla F_\lambda^k(x^{k+1})-\nabla F_\lambda^k(x^k)-\nabla^2 F_\lambda^k(x^*)(x^{k+1}-x^k)\| \le M\|x^{k+1}-x^k\|\varepsilon_k,$$

where $\varepsilon_k := \max\{\|x^{k+1}-x^*\|^\nu, \|x^k-x^*\|^\nu\}$.

Proof. By the definition of F_λ^k in (53) and taking into account (12), we obtain

$$\begin{aligned}&\|\nabla F_\lambda^k(x^{k+1})-\nabla F_\lambda^k(x^k)-\nabla^2 F_\lambda^k(x^*)(x^{k+1}-x^k)\|\\ &\qquad\le \max_{j=1,\dots,m}\|\nabla F_j(x^{k+1})-\nabla F_j(x^k)-\nabla^2 F_j(x^*)(x^{k+1}-x^k)\|.\end{aligned}$$

On the other hand, for each $j\in\{1,\dots,m\}$, using (33) and (30), we have

$$\|\nabla F_j(x^{k+1})-\nabla F_j(x^k)-\nabla^2 F_j(x^*)(x^{k+1}-x^k)\| \le \int_0^1 \|\left(\nabla^2 F_j(x^k+\tau s^k)-\nabla^2 F_j(x^*)\right)s^k\|d\tau$$

$$\le M\|s^k\|\int_0^1 \|x^k+\tau s^k-x^*\|^\nu d\tau \le M\|s^k\|\max\{\|x^{k+1}-x^*\|^\nu, \|x^k-x^*\|^\nu\}.$$

By combining the last two inequalities, we obtain the desired result. □

Now, we can establish the superlinear convergence of Algorithm 1.

Theorem A.7. *Suppose that Assumptions 4.1 and 4.2 hold. Then, $\{x^k\}$ converges to x^* superlinearly.*

Proof. According to Theorem A.5, $d^k = x^{k+1} - x^k$ for all k sufficiently large. Consequently, $B_\lambda^k(x^{k+1} - x^k) = -\nabla F_\lambda^k(x^k)$ (see (11)), and hence

$$(B_\lambda^k - \nabla^2 F_\lambda^k(x^*))(x^{k+1} - x^k) = \nabla F_\lambda^k(x^{k+1}) - \nabla F_\lambda^k(x^k) - \nabla^2 F_\lambda^k(x^*)(x^{k+1} - x^k) - \nabla F_\lambda^k(x^{k+1}),$$

for all k sufficiently large. Therefore,

$$\begin{aligned}\frac{\|\nabla F_\lambda^k(x^{k+1})\|}{\|x^{k+1} - x^k\|} \leq \frac{\|(B_\lambda^k - \nabla^2 F_\lambda^k(x^*))(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|}& \\ + \frac{\|\nabla F_\lambda^k(x^{k+1}) - \nabla F_\lambda^k(x^k) - \nabla^2 F_\lambda^k(x^*)(x^{k+1} - x^k)\|}{\|x^{k+1} - x^k\|}&,\end{aligned}$$

for all k sufficiently large. Taking limits on both sides of the latter inequality, using (54) and Lemma A.6, we get

$$\lim_{k\to\infty} \frac{\|\nabla F_\lambda^k(x^{k+1})\|}{\|x^{k+1} - x^k\|} = 0. \tag{59}$$

On the other hand, considering the definition of F_λ^k in (53), Lemma 2.3*(iv)*, and Lemma A.2*(i)*, we find that

$$\begin{aligned}\frac{\|\nabla F_\lambda^k(x^{k+1})\|}{\|x^{k+1} - x^k\|} &\geq \frac{\|\sum_{j=1}^m \lambda_j^k \nabla F_j(x^{k+1})\|}{\|x^{k+1} - x^*\| + \|x^k - x^*\|} \geq \frac{\|d_{SD}(x^{k+1})\|}{\|x^{k+1} - x^*\| + \|x^k - x^*\|} \\ &\geq \frac{L}{2}\frac{\|x^{k+1} - x^*\|}{\|x^{k+1} - x^*\| + \|x^k - x^*\|} = \frac{L}{2}\frac{1}{1 + \frac{\|x^k - x^*\|}{\|x^{k+1} - x^*\|}}.\end{aligned}$$

Therefore, by using (59), we conclude that

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0,$$

which completes the proof. □

Proof of Theorem 4.3. The proof follows straightforwardly from Theorems A.5 and A.7. □

References

[1] M. A. Ansary and G. Panda. A modified quasi-Newton method for vector optimization problem. *Optimization*, 64(11):2289–2306, 2015.

[2] P. B. Assunção, O. P. Ferreira, and L. F. Prudente. Conditional gradient method for multiobjective optimization. *Comput. Optim. Appl.*, 78(3):741–768, 2021.

[3] J. Y. Bello Cruz, L. R. Lucambio Pérez, and J. G. Melo. Convergence of the projected gradient method for quasiconvex multiobjective optimization. *Nonlinear Anal.*, 74(16):5268–5273, 2011.

[4] E. Birgin and J. Martínez. *Practical Augmented Lagrangian Methods for Constrained Optimization.* SIAM, Philadelphia, 2014.

[5] H. Bonnel, A. N. Iusem, and B. F. Svaiter. Proximal methods in vector optimization. *SIAM J. Optim.*, 15(4):953–970, 2005.

[6] C. G. Broyden. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA J. Appl. Math.*, 6(1):76–90, 1970.

[7] R. H. Byrd and J. Nocedal. A tool for the analysis of quasi-Newton methods with application to unconstrained minimization. *SIAM J. Numer. Anal.*, 26(3):727–739, 1989.

[8] L. C. Ceng, B. S. Mordukhovich, and J. C. Yao. Hybrid approximate proximal method with auxiliary variational inequality for vector optimization. *J. Optimiz. Theory App.*, 146(2):267–303, 2010.

[9] L. C. Ceng and J. C. Yao. Approximate proximal methods in vector optimization. *Eur. J. Oper. Res.*, 183(1):1–19, 2007.

[10] W. Chen, X. Yang, and Y. Zhao. Conditional gradient method for vector optimization. *Comput. Optim. Appl.*, 85(3):857–896, 2023.

[11] T. D. Chuong. Generalized proximal method for efficient solutions in vector optimization. *Numer. Funct. Anal. Optim.*, 32(8):843–857, 2011.

[12] T. D. Chuong. Newton-like methods for efficient solutions in vector optimization. *Comput. Optim. Appl.*, 54(3):495–516, 2013.

[13] T. D. Chuong, B. S. Mordukhovich, and J. C. Yao. Hybrid approximate proximal algorithms for efficient solutions in vector optimization. *J. Nonlinear Convex Anal.*, 12(2):257–285, 2011.

[14] A. L. Custódio, J. F. A. Madeira, A. I. F. Vaz, and L. N. Vicente. Direct Multisearch for Multiobjective Optimization. *SIAM J. Optim.*, 21(3):1109–1140, 2011.

[15] Y.-H. Dai. Convergence properties of the BFGS algorithm. *SIAM J. Optim.*, 13(3):693–701, 2002.

[16] Y.-H. Dai. A perfect example for the BFGS method. *Math. Program.*, 138(1-2):501–530, 2013.

[17] J. E. Dennis and J. J. Moré. A characterization of superlinear convergence and its application to quasi-Newton methods. *Math. Comp.*, 28(126):549–560, 1974.

[18] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Math. Program.*, 91(2):201–213, 2002.

[19] G. Eichfelder. *Adaptive Scalarization Methods in Multiobjective Optimization.* Springer Berlin Heidelberg, 2008.

[20] N. S. Fazzio and M. L. Schuverdt. Convergence analysis of a nonmonotone projected gradient method for multiobjective optimization problems. *Optim. Lett.*, 13(6):1365–1379, 2019.

[21] R. Fletcher. A new approach to variable metric algorithms. *Comput. J.*, 13(3):317–322, 1970.

[22] J. Fliege, L. M. Graña Drummond, and B. F. Svaiter. Newton's method for multiobjective optimization. *SIAM J. Optim.*, 20(2):602–626, 2009.

[23] J. Fliege and B. F. Svaiter. Steepest descent methods for multicriteria optimization. *Math. Methods of Oper. Res.*, 51(3):479–494, 2000.

[24] E. H. Fukuda and L. M. Graña Drummond. On the convergence of the projected gradient method for vector optimization. *Optimization*, 60(8-9):1009–1021, 2011.

[25] E. H. Fukuda and L. M. Graña Drummond. Inexact projected gradient method for vector optimization. *Comput. Optim. Appl.*, 54(3):473–493, 2013.

[26] D. Goldfarb. A family of variable-metric methods derived by variational means. *Math. Comput.*, 24:23–26, 1970.

[27] M. Gonçalves, F. Lima, and L. Prudente. A study of Liu-Storey conjugate gradient methods for vector optimization. *Appl. Math. Comput.*, 425:127099, 2022.

[28] M. L. N. Gonçalves, F. S. Lima, and L. F. Prudente. Globally convergent Newton-type methods for multiobjective optimization. *Comput. Optim. Appl.*, 83(2):403–434, 2022.

[29] M. L. N. Gonçalves and L. F. Prudente. On the extension of the Hager–Zhang conjugate gradient method for vector optimization. *Comput. Optim. Appl.*, 76(3):889–916, 2020.

[30] L. M. Graña Drummond and A. N. Iusem. A Projected Gradient Method for Vector Optimization Problems. *Comput. Optim. Appl.*, 28(1):5–29, 2004.

[31] L. M. Graña Drummond, F. M. P. Raupp, and B. F. Svaiter. A quadratically convergent Newton method for vector optimization. *Optimization*, 63(5):661–677, 2014.

[32] L. M. Graña Drummond and B. F. Svaiter. A steepest descent method for vector optimization. *J. Comput. Appl. Math.*, 175(2):395–414, 2005.

[33] K. K. Lai, S. K. Mishra, and B. Ram. On q-Quasi-Newton's Method for Unconstrained Multiobjective Optimization Problems. *Mathematics*, 8(4), 2020.

[34] M. Lapucci and P. Mansueto. A limited memory quasi-Newton approach for multi-objective optimization. *Comput. Optim. Appl.*, 85(1):33–73, 2023.

[35] D.-H. Li and M. Fukushima. A modified BFGS method and its global convergence in nonconvex minimization. *J. Comput. Appl. Math.*, 129(1):15–35, 2001.

[36] D.-H. Li and M. Fukushima. On the global convergence of the BFGS method for nonconvex unconstrained optimization problems. *SIAM J. Optim.*, 11(4):1054–1064, 2001.

[37] L. R. Lucambio Pérez and L. F. Prudente. Nonlinear conjugate gradient methods for vector optimization. *SIAM J. Optim.*, 28(3):2690–2720, 2018.

[38] L. R. Lucambio Pérez and L. F. Prudente. A Wolfe line search algorithm for vector optimization. *ACM Trans. Math. Softw.*, 45(4):23, 2019.

[39] N. Mahdavi-Amiri and F. S. Sadaghiani. A superlinearly convergent nonmonotone quasi-newton method for unconstrained multiobjective optimization. *Optim. Methods Softw.*, 35(6):1223–1247, 2020.

[40] W. F. Mascarenhas. The BFGS method with exact line searches fails for non-convex objective functions. *Math. Program.*, 99(1):49–61, 2004.

[41] K. Miettinen. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 1999.

[42] V. Morovati, H. Basirzadeh, and L. Pourkarimi. Quasi-Newton methods for multiobjective optimization problems. *4OR-Q J Oper Res*, 16(3):261–294, 2017.

[43] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer, New York, NY, USA, second edition, 2006.

[44] Z. Povalej. Quasi-Newton's method for multiobjective optimization. *J. Comput. Appl. Math.*, 255:765–777, 2014.

[45] M. J. D. Powell. Some global convergence properties of a variable metric algorithm for minimization without exact line searches. *Nonlinear Programming, SIAM-AMS Proceedings*, 4:53–72, 1976.

[46] L. F. Prudente and D. R. Souza. A quasi-Newton method with Wolfe line searches for multiobjective optimization. *J. Optim. Theory Appl.*, 194:1107–1140, 2022.

[47] S. Qu, M. Goh, and F. T. Chan. Quasi-Newton methods for solving multiobjective optimization. *Oper. Res. Lett.*, 39(5):397–399, 2011.

[48] S. Qu, C. Liu, M. Goh, Y. Li, and Y. Ji. Nonsmooth multiobjective programming with quasi-Newton methods. *Eur. J. Oper. Res.*, 235(3):503–510, 2014.

[49] D. F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Math. Comput.*, 24:647–656, 1970.

[50] B. F. Svaiter. The multiobjective steepest descent direction is not Lipschitz continuous, but is Hölder continuous. *Oper. Res. Lett.*, 46(4):430–433, 2018.

[51] J. Wang, Y. Hu, C. K. Wai Yu, C. Li, and X. Yang. Extended Newton methods for multiobjective optimization: majorizing function technique and convergence analysis. *SIAM J. Optim.*, 29(3):2388–2421, 2019.