Nonlinear Conjugate Gradient Methods with Guaranteed Descent for Multi-Objective Optimzation

A Preprint

Manuel Berkemeier.
Department of Computer Science
Paderborn University, Germany
manuelbb@mail.uni-paderborn.de

Konstantin Sonntag.
Department of Applied Mathematics
Paderborn University, Germany

Manuel Berkemeier.
Department of Computer Science
Paderborn University, Germany

September 28, 2023

Keywords Multi-Objective Optimization · Vector Optimization · Nonlinear Optimization · Unconstrained Optimization · Conjugate Gradient Method · Line-Search Algorithm

Abstract

In this article, we present several examples of special nonlinear conjugate gradient directions for nonlinear (non-convex) multi-objective optimization. These directions provide a descent direction for the objectives, independent of the line-search. This way, we can provide an algorithm with simple, Armijo-like backtracking and prove convergence to a first-order critical point. In contrast to other popular conjugate gradient methods, no Wolfe conditions for the step-sizes have to be satisfied. Besides investigating the theoretical properties of the algorithm, we also provide numerical examples to illustrate its efficacy.

Todo list

1 Introduction

Optimization problems with two or more competing objective functions may arise in different areas of mathematics, engineering, in the natural sciences or in economics. We call such problems multi-objective optimization problem (MOP) and multi-objective optimization (MOO) is concerned with finding acceptable trade-offs between the objectives of an MOP. In more precise terms, optimality of our vector-valued objective function $\boldsymbol{f} \colon \mathbb{R}^N \to \mathbb{R}^K$, with dimensions $N, K \in \mathbb{N}$, is determined by the partial ordering $\preceq_{\mathcal{K}}$ induced by a closed, convex, pointed cone $\mathcal{K} \subseteq \mathbb{R}^K$ with $\operatorname{int}(\mathcal{K}) \neq \emptyset$. We have $\boldsymbol{y}_1 \preceq_{\mathcal{K}} \boldsymbol{y}_2$ iff $\boldsymbol{y}_2 - \boldsymbol{y}_1 \in \mathcal{K}$ and $\boldsymbol{y}_1 \prec_{\mathcal{K}} \boldsymbol{y}_2$ iff $\boldsymbol{y}_2 - \boldsymbol{y}_1 \in \operatorname{int}(\mathcal{K})$.

Definition 1. The solutions to the unconstrained problem

$$\overset{\preceq_{\mathcal{K}}}{\min_{\boldsymbol{x}\in\mathbb{R}^N}} \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_K(\boldsymbol{x}) \end{bmatrix} = \overset{\preceq_{\mathcal{K}}}{\min_{\boldsymbol{x}\in\mathbb{R}^N}} \boldsymbol{f}(\boldsymbol{x}) \tag{MOP}$$

are minimal with respect to $\prec_{\mathcal{K}}$ and are called *Pareto-optimal*. That is, a point $\boldsymbol{x}^* \in \mathbb{R}^N$ is optimal, if there is no $\mathbb{R}^N \setminus \{\boldsymbol{x}^*\}$ with $\boldsymbol{f}(\boldsymbol{x}) \prec_{\mathcal{K}} \boldsymbol{f}(\boldsymbol{x}^*)$.

In practical applications, one often encounters $\mathcal{K} = \mathbb{R}^K_{\geq 0}$. Today, there is a multitude of methods to solve MOPs, and it would be out of the scope of this article to provide a complete overview. To see how (MOP) can be transformed into a single-objective problem via *scalarization*, there are entire books on the subjects, e.g. [3]. To obtain multiple solutions, oftentimes so-called *evolutionary algorithms* are used, NSGA-II [2] being a prominent example. If derivatives are available, local descent-based optimization allows for finding individual critical solutions [18, 8, 9], or continuing along the manifold of solutions towards more favorable points [13, 16]. Preferences can also be encoded in scalarizations, see [5, 17]. For more complete overview, consider the surveys [20] and [4].

Conjugate Gradient Methods

The main motivation for our work stems from recent improvements of the convergence rate of some first-order methods in MOO. Multi-objective steepest descent suffers from the same sublinear convergence rates of its single-objective counterpart [7]. For convex objectives, it can be shown that faster first-order methods exist [23, 21]. In the non-convex case, nonlinear conjugate gradient (CG) algorithms have empirically proven themselves good alternatives.

Originally, the CG method is an iterative method used for the numerical solution of particular systems of linear equations, specifically those whose matrix is symmetric and positive-definite. The method is best suited to large-scale problems where direct methods are not feasible [19]. The desirable convergence properties of the linear conjugate gradient method has motivated the use of similar directions in iterative schemes for large-scale *nonlinear* optimization problems. Similarly to the linear case, the descent direction is a linear combination of the negative gradient and the previous direction, but the multipliers are different. Today, there is a multitude of different nonlinear conjugate methods which tend to be faster than the steepest descent method [19].

Recently, Lucambio Pérez and Prudente [14] have adapted many of the popular nonlinear CG methods to the multi-objective setting. Their directions [10, 14] rely on strong Wolfe conditions being fulfilled. To this end, a suitable step-size algorithm is provided [15]. These multi-objective nonlinear CG methods work well in experiments, but the line-search algorithm might require step-sizes that are undesirably large, and its implementation is more involved than using simple Armijo-like backtracking. In [11], a nonlinear CG method for vector optimization is proposed that works with a simple backtracking algorithm, but it requires estimates of the Lipschitz constants of the objectives.

In contrast, the directions in this work satisfy a *sufficient decrease* condition by construction – independent of the line-search. The directions are adapted (or "translated") from the single-objective setting, albeit there already is a scheme for bi-objective optimization [6].

2 Criticality

Given smooth objective functions, there is necessary condition for Pareto-optimality in (MOP) similar to Fermat's theorem in single objective optimization. Let $\boldsymbol{\nabla f}(\boldsymbol{x}) \in \mathbb{R}^{K\times N}$ denote the Jacobian of $\boldsymbol{f}$ at $\boldsymbol{x}$. If $\boldsymbol{x}^*$ is Pareto-optimal, then it is also critical, according to the following definition:

Definition 2. The point $\boldsymbol{x}^*$ is Pareto-critical iff

$$-\operatorname{int}(\mathcal{K}) \cap \operatorname{img}(\boldsymbol{\nabla f}(\boldsymbol{x}^*)) = \emptyset.$$

Vice versa, if $\boldsymbol{x}$ is not critical, then there is a *descent direction* $\boldsymbol{v} \in \mathbb{R}^N$, with the defining property $\boldsymbol{\nabla f}(\boldsymbol{x})\boldsymbol{v} \in -\operatorname{int}(\mathcal{K})$. For such a direction, there is some step-size bound $\bar{\sigma} > 0$ with

$$\boldsymbol{f}(\boldsymbol{x} + \sigma\boldsymbol{v}) \preceq_{\mathcal{K}} \boldsymbol{f}(\boldsymbol{x}) \qquad \forall \sigma \in (0, \bar{\sigma}) \quad \text{(see [12]).}$$

We will proceed to introduce the maps $\varphi(\bullet)$ and $\mathfrak{f}(\bullet, \bullet)$ to facilitate working with the definitions of Pareto-optimality and -criticality. Adopting the notation from [12, 14], let $\langle \bullet, \bullet \rangle$ be the usual inner product on $\mathbb{R}^K$ and

$$\mathcal{K}^* = \{\boldsymbol{w} \in \mathbb{R}^K : \langle \boldsymbol{w}, \boldsymbol{y} \rangle \geq 0 \ \forall \boldsymbol{y} \in \mathcal{K}\},$$

the dual cone of $\mathcal{K}$. Further, let $C \subset \mathcal{K}^* \setminus \{\mathbf{0}\}$ be a compact set generating $\mathcal{K}^*$ as its conical hull:

$$\mathcal{K}^* = \operatorname{coni}(C) = \left\{ \sum_{i=1}^{P} \lambda_i \boldsymbol{y}_i : \lambda_i \geq 0, \boldsymbol{y}_i \in C, P \in \mathbb{N}_0 \right\}.$$

Then, the map

$$\varphi \colon \mathbb{R}^K \to \mathbb{R}, \boldsymbol{y} \mapsto \sup_{\boldsymbol{w} \in C} \langle \boldsymbol{y}, \boldsymbol{w} \rangle = \max_{\boldsymbol{w} \in C} \langle \boldsymbol{y}, \boldsymbol{w} \rangle \tag{1}$$

allows for a characterization of $-\mathcal{K}$ and $-\operatorname{int}(\mathcal{K})$ as its (strict) sublevel sets at 0:

$$\boldsymbol{y} \in -\mathcal{K} \Leftrightarrow \varphi(\boldsymbol{y}) \leq 0 \quad \text{and} \quad \boldsymbol{y} \in -\operatorname{int}(\mathcal{K}) \Leftrightarrow \varphi(\boldsymbol{y}) < 0.$$

This map, the support function of the dual cone, has the following properties:

Proposition 3 (Lemma 3.1 in [12])**.** *Let $\boldsymbol{y}, \boldsymbol{y\prime} \in \mathbb{R}^K$. Then*

1. *$\varphi(\boldsymbol{y} + \boldsymbol{y\prime}) \leq \varphi(\boldsymbol{y}) + \varphi(\boldsymbol{y\prime})$ and $\varphi(\boldsymbol{y}) - \varphi(\boldsymbol{y\prime}) \leq \varphi(\boldsymbol{y} - \boldsymbol{y\prime})$.*
2. *If $\boldsymbol{y} \preceq_{\mathcal{K}} \boldsymbol{y\prime}$, then $\varphi(\boldsymbol{y}) \leq \varphi(\boldsymbol{y\prime})$. If $\boldsymbol{y} \prec_{\mathcal{K}} \boldsymbol{y\prime}$, then $\varphi(\boldsymbol{y}) < \varphi(\boldsymbol{y\prime})$.*
3. *φ is Lipschitz-continuous.*

Furthermore, if we define $\mathfrak{f} \colon \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ by

$$\mathfrak{f}(\boldsymbol{x}, \boldsymbol{d}) = \mathfrak{f}_{\boldsymbol{x}}(\boldsymbol{d}) = \varphi(\boldsymbol{\nabla f}(\boldsymbol{x})\boldsymbol{d}) = \max_{\boldsymbol{w} \in C} \langle \boldsymbol{\nabla f}(\boldsymbol{x})\boldsymbol{d}, \boldsymbol{w} \rangle,$$

then we can infer criticality from the function $\mathfrak{f}$ as follows:

Proposition 4 (Lemma 3.3 in [12])**.** *Suppose $\boldsymbol{f}$ is continuously differentiable. Consider the following optimization problem:*

$$\min_{\boldsymbol{d} \in \mathbb{R}^N} \mathfrak{f}_{\boldsymbol{x}}(\boldsymbol{d}) + \frac{1}{2} \|\boldsymbol{d}\|_2^2. \tag{2}$$

Denote the minimizer by $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{x}) \in \mathbb{R}^N$ and the optimal value by $\alpha = \alpha(\boldsymbol{x}) \in \mathbb{R}$.

1. *If $\boldsymbol{x}$ is critical, then $\boldsymbol{\delta} = \mathbf{0}$ and $\alpha = 0$.*
2. *If $\boldsymbol{x}$ is* not *critical, then $\boldsymbol{\delta} \neq \mathbf{0}$, $\alpha < 0$ and $\mathfrak{f}_{\boldsymbol{x}}(\boldsymbol{\delta}) < -\frac{1}{2} \|\boldsymbol{\delta}\|^2 < 0$, and $\boldsymbol{\delta}$ is a descent direction.*
3. *The mappings $\boldsymbol{x} \mapsto \boldsymbol{\delta}(\boldsymbol{x}), \boldsymbol{x} \mapsto \alpha(\boldsymbol{x})$ are continuous.*

Moreover, in [12] it is shown that

$$\alpha(\boldsymbol{x}) = -\frac{1}{2} \|\boldsymbol{\delta}(\boldsymbol{x})\|^2, \quad \text{and thus} \quad \mathfrak{f}_{\boldsymbol{x}}(\boldsymbol{\delta}) = -\|\boldsymbol{\delta}\|^2.$$

In the single-objective case, with $\mathcal{K} = \mathbb{R}_{\geq 0}$, the solution is $\boldsymbol{\delta} = -\nabla f(\boldsymbol{x})$. The problem in (2) thus generalizes the concept of the steepest descent direction, and we obtain a recipe for "translating" nonlinear CG directions for multiple objectives. Obviously, the choice of C influences the solution set of (2), which is why we might wish to assume $\|\boldsymbol{y}\| = 1$ for all $\boldsymbol{y} \in C$.

Just as in single-objective optimization a sequence of directions $\boldsymbol{d}^{(k)} \in \mathbb{R}^N$ is said to fulfill the sufficient decrease condition if there is a constant $\kappa_{\text{sd}} > 0$ such that $\langle -\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}), \boldsymbol{d}^{(k)} \rangle \geq \kappa_{\text{sd}} \left\| \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}) \right\|^2$, we qualify them accordingly in the multi-objective case:

Definition 5. The directions $\{\boldsymbol{d}^{(k)}\}$ are said to have the *sufficient decrease* property iff

$$-\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k)}) = -\varphi(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k)}) \geq -\kappa_{\text{sd}} \varphi(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{\delta}^{(k)}) = -\kappa_{\text{sd}} \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)}) = \kappa_{\text{sd}} \left\| \boldsymbol{\delta}^{(k)} \right\|^2. \tag{3}$$

Should this hold independent of the line-search used to determine a step-size in an algorithm, we say that the directions $\{\boldsymbol{d}^{(k)}\}$ provide *guaranteed* descent.

3 Algorithm

The algorithm will be stated in a very generic manner. That is, we do not yet give specific formulas to compute the directions $\{\boldsymbol{d}^{(k)}\}$, but only assume them to have the sufficient decrease property (3). Additionally, we have to determine step-sizes. In the next subsection, we justify a simple backtracking procedure.

3.1 (Modified) Armijo Stepsize

Let $\boldsymbol{d} \in \mathbb{R}^N$ be a descent direction for $\boldsymbol{f}$ at $\boldsymbol{x}$ and let $\boldsymbol{e} \in \mathcal{K}$ be a vector such that

$$0 < \mathtt{c}_{\boldsymbol{e}} \leq \langle \boldsymbol{w}, \boldsymbol{e} \rangle \leq 1 \qquad \forall \boldsymbol{w} \in C. \tag{4}$$

We can find a suitable vector because $\mathcal{K}$ is pointed and C spans its dual cone. In case that $\mathcal{K}$ is $\mathbb{R}^K_{\geq 0}$ and C contains the canonical basis of $\mathbb{R}^K$, simply choose $\boldsymbol{e} = [1, \dots, 1]^{\mathsf{T}}$.

Our step-size should satisfy an Armijo-like condition, where the right-hand side (RHS) is modified to shrink quadratically. Modifying the condition found in [14], we get the strict modified Armijo condition:

Definition 6. Suppose that $\boldsymbol{f}$ is differentiable in an open set containing $\boldsymbol{x} \in \mathbb{R}^N$ and that $\boldsymbol{d}$ is a descent-direction. Let $\mathtt{a} \in (0, 1)$ be constant. The step-size $\sigma > 0$ satisfies the (strict) modified Armijo condition if

$$\boldsymbol{f}(\boldsymbol{x} + \sigma \boldsymbol{d}) - \boldsymbol{f}(\boldsymbol{x}) \preceq_{\mathcal{K}} -\mathtt{a}\sigma^2 \left\|\boldsymbol{d}\right\|^2 \boldsymbol{e}. \tag{5}$$

The next proposition shows that a suitable step-size actually exists.

Proposition 7. *Suppose the conditions of Definition 6 hold. Then there is a suitable step-size $\sigma > 0$ satisfying* (5)*.*

Proof. Suppose there was not:

$$\boldsymbol{f}(\boldsymbol{x} + \sigma \boldsymbol{d}) - \boldsymbol{f}(\boldsymbol{x}) + \mathtt{a}\sigma^2 \left\|\boldsymbol{d}\right\|^2 \boldsymbol{e} \notin -\mathcal{K} \qquad \forall \sigma > 0.$$

Then there is some $\boldsymbol{w} \in C$ such that for all $\sigma > 0$:

$$\left\langle \boldsymbol{w}, \boldsymbol{f}(\boldsymbol{x} + \sigma \boldsymbol{d}) - \boldsymbol{f}(\boldsymbol{x}) + \mathtt{a}\sigma^2 \left\|\boldsymbol{d}\right\|^2 \boldsymbol{e} \right\rangle > 0$$

$$\left\langle \boldsymbol{w}, \sigma \boldsymbol{\nabla f}(\boldsymbol{x})\boldsymbol{d} + \boldsymbol{R}(\sigma) + \mathtt{a}\sigma^2 \left\|\boldsymbol{d}\right\|^2 \boldsymbol{e} \right\rangle > 0$$

Rearranging and dividing by $\sigma > 0$ gives

$$\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x})\boldsymbol{d} \rangle > -\mathtt{a}\sigma \left\|\boldsymbol{d}\right\|^2 \langle \boldsymbol{w}, \boldsymbol{e} \rangle - \left\langle \boldsymbol{w}, \frac{\boldsymbol{R}(\sigma)}{\sigma} \right\rangle$$

$$\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x})\boldsymbol{d} \rangle > -\mathtt{a}\sigma \left\|\boldsymbol{d}\right\|^2 - \left\langle \boldsymbol{w}, \frac{\boldsymbol{R}(\sigma)}{\sigma} \right\rangle \tag{6}$$

where, by definition of the total differential, $\dfrac{\boldsymbol{R}(\sigma)}{\sigma} \to \boldsymbol{0}$, as $\sigma \to 0$ from above. Because $\boldsymbol{d}$ is a descent direction, the value on the left-hand side (LHS) in (6) is constant and strictly negative, while the RHS goes to zero. A contradiction! □

There is also a less strict variant of the modified Armijo condition:

Definition 8. Under the same conditions as in Definition 6, the step-size $\sigma > 0$ satisfies the weak modified Armijo condition if

$$\varphi\left(\boldsymbol{f}(\boldsymbol{x} + \sigma \boldsymbol{d})\right) - \varphi\left(\boldsymbol{f}(\boldsymbol{x})\right) \leq -\mathtt{a}\sigma^2 \left\|\boldsymbol{d}\right\|^2 \mathtt{c}_{\boldsymbol{e}}. \tag{7}$$

For $\mathcal{K} = \mathbb{R}^K_{\geq 0}$, the weak condition only guarantees descent in one objective, so the algorithm will produce value vectors that are not monotonic with respect to $\preceq_{\mathcal{K}}$. Likely, larger steps are taken and the sequence $\left\{\varphi(\boldsymbol{f}(\boldsymbol{x}^{(k)}))\right\}_k$ will be monotonic.

Proposition 9. *The strict modified Armijo condition* (5) *implies the weak condition* (7)*.*

Proof. Suppose the strict Armijo condition (5) is fulfilled. With Proposition 3 it follows that

$$\varphi\left(\boldsymbol{f}(\boldsymbol{x} + \sigma \boldsymbol{d})\right) - \varphi\left(\boldsymbol{f}(\boldsymbol{x})\right) \leq \varphi\left(\boldsymbol{f}(\boldsymbol{x} + \sigma \boldsymbol{d}) - \boldsymbol{f}(\boldsymbol{x})\right) \leq \varphi\left(-\mathtt{a}\sigma^2 \left\|\boldsymbol{d}\right\|^2 \boldsymbol{e}\right)$$

For the RHS we get from the definition (1) that

$$\varphi\left(-\mathtt{a}\sigma^2 \left\|\boldsymbol{d}\right\|^2 \boldsymbol{e}\right) = \max_{\boldsymbol{w} \in C} \left(-\mathtt{a}\sigma^2 \left\|\boldsymbol{d}\right\|^2\right) \langle \boldsymbol{w}, \boldsymbol{e} \rangle = -\mathtt{a}\sigma^2 \left\|\boldsymbol{d}\right\|^2 \min_{\boldsymbol{w} \in C} \langle \boldsymbol{w}, \boldsymbol{e} \rangle = -\mathtt{a}\sigma^2 \left\|\boldsymbol{d}\right\|^2 \mathtt{c}_{\boldsymbol{e}}.$$

□

A step-size satisfying (5) or (7) can be found by backtracking: Let $k \in \mathbb{N}$, $\boldsymbol{x}^{(k)} \in \mathbb{R}^N$ and let $\boldsymbol{d}^{(k)} \in \mathbb{R}^N$ be a descent direction of $\boldsymbol{f}$ at $\boldsymbol{x}^{(k)}$. Further, let $\mathtt{b} \in (0, 1)$ and $\mathtt{a} \in (0, 1)$ be constants and $\sigma_0^{(k)}$ an initial step-size bounded below by the constant $\mathtt{M} > 0$.

$$\sigma_{(k)} = \max_{j \in \mathbb{N}_0} \mathtt{b}^j \sigma_0^{(k)} \quad \text{such that (5) (or (7)) holds.} \tag{8}$$

3.2 Algorithm

```
Algorithm 1: Algorithm with Generic Descent Direction
Data: $N \in \mathbb{N}, K \in \mathbb{N}, \boldsymbol{f} \colon \mathbb{R}^N \to \mathbb{R}^K, \boldsymbol{x}^{(0)} \in \mathbb{R}^N, \mathtt{a} \in (0,1), \mathtt{b} \in (0,1), \kappa_{\mathrm{sd}} > 0, \sigma_0^{(k)} \geq \mathtt{M} > 0$. A step-size
      condition $(\star)$: either (5) or (7).
Result: A critical point $\boldsymbol{x}^{(k)}$ or a critical sequence $\{\boldsymbol{x}^{(k)}\}_{k \in \mathbb{N}_0}$.
for $k \in \mathbb{N}_0$ do
    if $\boldsymbol{x}^{(k)}$ is critical then STOP;
    Compute a direction $\boldsymbol{d}^{(k)}$ satisfying (3);
    Compute a step-size $\sigma_{(k)}$ satisfying $(\star)$ by backtracking like in (8);
    Set $\boldsymbol{x}^{(k+1)} \leftarrow \boldsymbol{x}^{(k)} + \sigma_{(k)} \boldsymbol{d}^{(k)}$;
end
```

We are now in a position to state the complete algorithm in Algorithm 1. In the following we continue to establish results meant to prove converge for specific directions $\left\{\boldsymbol{d}^{(k)}\right\}$ in subsequent sections. Of course, there is nothing to show if we stop at a critical point with finite termination. We hence implicitly assume infinite sequences from now on. For the analysis, we introduce a set of assumptions:

Assumption 1. For a given initial point $\boldsymbol{x}^{(0)}$, the function $\boldsymbol{f} \colon \mathbb{R}^N \to \mathbb{R}^K$ is defined on the set

$$\mathcal{F} = \begin{cases} \left\{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{f}(\boldsymbol{x}) \preceq_{\mathcal{K}} \boldsymbol{f}(\boldsymbol{x}^{(0)})\right\}, & \text{if (5) is used,} \\ \left\{\boldsymbol{x} \in \mathbb{R}^N : \varphi(\boldsymbol{f}(\boldsymbol{x})) \leq \varphi(\boldsymbol{f}(\boldsymbol{x}^{(0)}))\right\}, & \text{if (7) is used.} \end{cases}$$

Furthermore, $\boldsymbol{f}$ is continuously differentiable in an open set containing $\mathcal{F}$.

Assumption 2. Assumption 1 holds and the Jacobian of $\boldsymbol{f}$ is Lipschitz-continuous with constant $\mathtt{L}_f > 0$.

Assumption 3. Depending on the step-size condition, the following holds:

1. If the strict condition (5) is used: Every non-increasing sequence in $\boldsymbol{f}(\mathcal{F})$,

$$\{\boldsymbol{y}^{(k)}\}_k \subseteq \{\boldsymbol{f}(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{F}\}, \quad \boldsymbol{y}^{(k+1)} \preceq_{\mathcal{K}} \boldsymbol{y}^{(k)} \; \forall k,$$

   is bounded below by some $\boldsymbol{y}_{\mathrm{lb}} \in \mathbb{R}^K$ as per $\boldsymbol{y}_{\mathrm{lb}} \preceq_{\mathcal{K}} \boldsymbol{y}^{(k)}$ for all k. Proposition 3 then implies that $\left\{\varphi(\boldsymbol{y}^{(k)})\right\}_k$ is non-increasing and that it is bounded,

$$y_{\mathrm{lb}} := \varphi(\boldsymbol{y}_{\mathrm{lb}}) \leq \varphi(\boldsymbol{y}^{(k)}) \; \forall k.$$

2. If the weak condition (7) is used: For every sequence in $\boldsymbol{f}(\mathcal{F})$ with non-increasing φ values, i.e.,

$$\{\varphi(\boldsymbol{y}^{(k)})\}_k, \quad \varphi(\boldsymbol{y}^{(k+1)}) \preceq_{\mathcal{K}} \varphi(\boldsymbol{y}^{(k)}) \; \forall k,$$

   the sequence of values is bounded below by some $y_{\mathrm{lb}} \in \mathbb{R}^K$ as per $y_{\mathrm{lb}} \preceq_{\mathcal{K}} \varphi(\boldsymbol{y}^{(k)})$ for all k.

Our assumptions are rather granular to indicate what is strictly necessary for the results to work. Instead of Assumptions 2 and 3 we could also demand the sublevel set $\mathcal{F}$ to be bounded. Lipschitz-continuity of the Jacobian and boundedness of every (value) sequence automatically follows. It is rather a matter of taste that we try to have it optional whenever possible:

Assumption 4. The sublevel set $\mathcal{F}$ as defined in Assumption 1 is bounded.

Our first results concerns the step-length. It goes to zero because of the modified Armijo condition:

Lemma 10. *Consider a sequence $\left\{(\boldsymbol{x}^{(k)}, \boldsymbol{d}^{(k)}, \sigma_{(k)})\right\}_k$ produced by Algorithm 1. Suppose Assumptions 1 and 3 hold. Then*

$$\lim_{k \to \infty} \sigma_{(k)}^2 \left\|\boldsymbol{d}_k\right\|^2 = \lim_{k \to \infty} \sigma_{(k)} \left\|\boldsymbol{d}_k\right\| = 0.$$

Proof. By design, the weak Armijo condition (7) holds:

$$\varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(k)})\right) - \varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(k)} + \sigma_{(k)}\boldsymbol{d}^{(k)})\right) \geq \mathtt{a}\mathtt{c}_{\boldsymbol{e}}\sigma_{(k)}^2 \left\|\boldsymbol{d}^{(k)}\right\|^2.$$

Combining the constants into $\mathtt{c} > 0$ and summing up to $\kappa \in \mathbb{N}_0$ gives

$$\mathtt{c}\sum_{k=0}^{\kappa} \sigma_{(k)}^2 \left\|\boldsymbol{d}^{(k)}\right\|^2 \le \varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(0)})\right) - \varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(\kappa+1)} + \sigma^{(\kappa+1)}\boldsymbol{d}^{(\kappa+1)})\right)$$

Due to Assumption 3, the RHS simplifies:

$$\mathtt{c}\sum_{k=0}^{\kappa} \sigma_{(k)}^2 \left\|\boldsymbol{d}^{(k)}\right\|^2 \le \varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(0)})\right) - y_{\mathrm{lb}} = \text{const.}$$

We see that the LHS is a monotonically increasing sequence and bounded above. Due to the Monotone Convergence Theorem, it must be convergent, i.e.,

$$\sum_{k=0}^{\infty} \sigma_{(k)}^2 \left\|\boldsymbol{d}^{(k)}\right\|^2 < \infty$$

□

Next, we in derive a bound resembling the Zoutendijk condition often encountered in single-objective optimization. From now on, we will refrain from explicitly stating that the iterates are generated by Algorithm 1.

Proposition 11. *Suppose Assumptions 1 to 3 hold and that the step-sizes $\sigma_{(k)}$ in Algorithm 1 satisfy Eq.* (5). *Then following Zoutendijk-like condition follows:*

$$\sum_{k\in\mathbb{N}_0} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} = \sum_{k\in\mathbb{N}_0} \frac{\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})^2}{\left\|\boldsymbol{d}^{(k)}\right\|^2} < \infty \tag{9}$$

Proof. Let $k \in \mathbb{N}_0$ and consider two cases.

First, suppose $\sigma_{(k)} \neq \sigma_0^{(k)}$. Due to the backtracking procedure, the strict Armijo condition (5) must be violated for $\sigma_{(k)}\mathtt{b}^{-1} > \sigma_{(k)}$. (If the weak condition is used, and it is violated for some stepsize, then the strict condition cannot hold neither for that stepsize.) There thus is $\boldsymbol{w} \in C$ such that

$$\left\langle \boldsymbol{w}, \boldsymbol{f}\left(\boldsymbol{x} + \frac{\sigma_{(k)}}{\mathtt{b}}\boldsymbol{d}^{(k)}\right) - \boldsymbol{f}(\boldsymbol{x}) + \mathtt{a}\frac{\sigma_{(k)}^2}{\mathtt{b}^2}\left\|\boldsymbol{d}^{(k)}\right\|^2 \boldsymbol{e} \right\rangle > 0.$$

It follows, that

$$-\mathtt{a}\frac{\sigma_{(k)}^2}{\mathtt{b}^2}\left\|\boldsymbol{d}^{(k)}\right\|^2 \le \left\langle \boldsymbol{w}, -\mathtt{a}\frac{\sigma_{(k)}^2}{\mathtt{b}^2}\left\|\boldsymbol{d}^{(k)}\right\|^2 \boldsymbol{e} \right\rangle < \left\langle \boldsymbol{w}, \boldsymbol{f}\left(\boldsymbol{x} + \frac{\sigma_{(k)}}{\mathtt{b}}\boldsymbol{d}^{(k)}\right)\right\rangle - \langle \boldsymbol{w}, \boldsymbol{f}(\boldsymbol{x})\rangle\,.$$

Applying the mean-value-theorem on the RHS gives some $h \in (0,1)$ with

$$\begin{aligned}
-\mathtt{a}\frac{\sigma_{(k)}^2}{\mathtt{b}^2}\left\|\boldsymbol{d}^{(k)}\right\|^2 &\le \frac{\sigma_{(k)}}{\mathtt{b}}\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}\left(\boldsymbol{x}^{(k)} + h\frac{\sigma_{(k)}}{\mathtt{b}}\boldsymbol{d}^{(k)}\right)\boldsymbol{d}^{(k)}\right\rangle \\
&= \frac{\sigma_{(k)}}{\mathtt{b}}\left\langle \boldsymbol{w}, \left(\boldsymbol{\nabla f}\left(\boldsymbol{x}^{(k)} + h\frac{\sigma_{(k)}}{\mathtt{b}}\boldsymbol{d}^{(k)}\right) - \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\right)\boldsymbol{d}^{(k)} + \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k)}\right\rangle \\
&\le \frac{\sigma_{(k)}^2}{\mathtt{b}^2}\mathtt{L}_f\left\|\boldsymbol{w}\right\|\left\|\boldsymbol{d}^{(k)}\right\|^2 + \frac{\sigma_{(k)}}{\mathtt{b}}\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k)}),
\end{aligned}$$

where in the last line we have used the Cauchy-Schwarz inequality and the Lipschitz continuity of $\boldsymbol{\nabla f}$ (Assumption 2). Because C is compact, $\|\boldsymbol{w}\|$ is bounded and there is a constant $\mathtt{c} > 0$ such that

$$\sigma_{(k)} \ge \mathtt{c}\frac{-\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k)})}{\left\|\boldsymbol{d}^{(k)}\right\|^2}.$$

The RHS is positive, because $\boldsymbol{d}^{(k)}$ is a descent direction. The algorithm also ensures that the sufficient decrease conditions holds. Plugging the last equation into the weak Armijo condition (7), which must hold for $\sigma_{(k)}$, results in

$$\varphi(\boldsymbol{f}(\boldsymbol{x}^{(k)})) - \varphi\left(\boldsymbol{f}\left(\boldsymbol{x}^{(k)} + \sigma_{(k)}\boldsymbol{d}^{(k)}\right)\right) \ge \mathtt{acc}_{\boldsymbol{e}}\frac{\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k)})^2}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \overset{(3)}{\ge} \mathtt{acc}_{\boldsymbol{e}}\kappa_{\mathrm{sd}}\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2}. \tag{10}$$

Now, suppose $\sigma_{(k)} = \sigma_0^{(k)}$. By definition of $\boldsymbol{\delta}^{(k)}$ as the minimizer of (2), we have

$$\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)}) + \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}{2} \le \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\kappa_{\text{sd}}^{-1}\boldsymbol{d}^{(k)}) + \frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{2\kappa_{\text{sd}}^2}$$

The sufficient decrease condition gives $\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\kappa_{\text{sd}}^{-1}\boldsymbol{d}^{(k)}) \le \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})$, so it must hold that $\left\|\boldsymbol{\delta}^{(k)}\right\|^2 \le \frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{\kappa_{\text{sd}}^2}$. Thus,

$$\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \le \frac{1}{\kappa_{\text{sd}}^4}\left\|\boldsymbol{d}^{(k)}\right\|^2 \le \frac{1}{\kappa_{\text{sd}}^4 \mathtt{a}\left(\sigma_0^{(k)}\right)^2}\left(\varphi(\boldsymbol{f}(\boldsymbol{x}^{(k)})) - \varphi\left(\boldsymbol{f}\left(\boldsymbol{x}^{(k)} + \sigma_{(k)}\boldsymbol{d}^{(k)}\right)\right)\right)$$

As $\sigma_0^{(k)} \ge \mathtt{M} > 0$ for all k, we again obtain an expression

$$\bar{\mathtt{c}}\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \le \varphi(\boldsymbol{f}(\boldsymbol{x}^{(k)})) - \varphi\left(\boldsymbol{x}^{(k)} + \sigma_{(k)}\boldsymbol{d}^{(k)}\right) \tag{11}$$

for some constant $\bar{\mathtt{c}} > 0$.

Like in the proof of Lemma 10, we can deduce convergence of the series in (9) from (10) and (11) with Assumption 3, by realizing that the partial sums are again increasing and bounded. □

Later, the Zoutendijk property allows for a convenient way to prove convergence for certain direction schemes by means of contradiction, as made explicit with the following corollary.

Corollary 12. *Suppose, the criticality is uniformly bounded from below via*

$$\left\|\boldsymbol{\delta}^{(k)}\right\| \ge \varepsilon_{\text{crit}} > 0 \qquad \forall k \in \mathbb{N}_0. \tag{12}$$

If the directions $\left\{\boldsymbol{d}^{(k)}\right\}$ and step-sizes $\left\{\sigma_{(k)}\right\}$ are chosen so that Proposition 11 applies, ***and*** *it additionally holds that*

$$\sum_{k\in\mathbb{N}_0}\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} = \infty,$$

then $\liminf_{k\to\infty}\left\|\boldsymbol{\delta}^{(k)}\right\| = 0$. *We say, that the sequence $\left\{\boldsymbol{x}^{(k)}\right\}$ is critical. Furthermore, if the domain is bounded (Assumption 4), there is a subsequence of iterates $\{\boldsymbol{x}^{(k)}\}$ converging to a Pareto-critical point.*

If the directions $\left\{\boldsymbol{d}^{(k)}\right\}$ remain bounded, then this constitutes a special case, as can be easily verified.

Corollary 13. *Suppose that* (12) *holds and that Proposition 11 applies. If there is a constant $\mathtt{C}_{\boldsymbol{d}} > 0$ with*

$$\left\|\boldsymbol{d}^{(k)}\right\| \le \mathtt{C}_{\boldsymbol{d}} \qquad \forall k \in \mathbb{N}_0,$$

then the algorithm has a critical sequence.

4 Specific Directions with Guaranteed Descent

All that is left to do, is to actually provide directions $\left\{\boldsymbol{d}^{(k)}\right\}$ that can be used with Algorithm 1. The directions presented in this section are adapted from single-objective schemes, of which there are many. Hence, our list is by no means complete, and there are many more approaches to explore.

Insert a few references from single-objective optimization.

4.1 Two Flavors of Fletcher-Reeves

We will show two variants, derived from the single-objective scheme in [25]. The single-objective directions $\boldsymbol{d}^{(k)}$ are inspired by the classical Fletcher-Reeves (FR) recipe and given by

$$\boldsymbol{d}^{(k)} = \begin{cases} -\boldsymbol{g}^{(k)} & \text{if } k = 0, \\ -\theta_{(k)}\boldsymbol{g}^{(k)} + \beta_{(k)}\boldsymbol{d}^{(k-1)} & \text{if } k \ge 1, \end{cases} \quad \beta_{(k)} := \frac{\left\|\boldsymbol{g}^{(k)}\right\|^2}{\left\|\boldsymbol{g}^{(k-1)}\right\|^2}, \; \theta_{(k)} := \frac{\left\langle \boldsymbol{d}^{(k-1)}, \boldsymbol{g}^{(k)} - \boldsymbol{g}^{(k-1)}\right\rangle}{\left\|\boldsymbol{g}^{(k)}\right\|^2}, \tag{FR SO}$$

where $\boldsymbol{g}^{(k)} = \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})$.

FR Restart Variant

Unfortunately, simply replacing $-\boldsymbol{g}^{(k)}$ with the multi-objective steepest descent direction $\boldsymbol{\delta}^{(k)}$ from (2) and naively modifying $\theta_{(k)}$ does not work. For example, these definitions,

$$\tilde{\beta}_{(k)} = \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} = \frac{\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})}{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})},\ \tilde{\theta}_{(k)} = \frac{\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k-1)}) - \mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)})}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} = \frac{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)}) - \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k-1)})}{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})} \tag{13}$$

do not suffice to show the sufficient decrease property (3). $\tilde{\theta}_{(k)}$ might become negative (and this actually happens in practice). A negative coefficient can be avoided with Wolfe conditions. Supposing $k \geq 1$ and $\boldsymbol{d}^{(k-1)}$ is a descent-direction at $\boldsymbol{x}^{(k-1)}$, this direction satisfies the strong Wolfe conditions for $\sigma_{\text{sw}} \in (0,1)$ iff

$$\left|\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k-1)})\right| \leq \sigma_{\text{sw}} \left|\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)})\right| = -\sigma_{\text{sw}}\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)})$$

However, we do not enforce (strong) Wolfe conditions during line-search, but simply perform restarts if they are not satisfied. More specifically, we test for a modified version of the Wolfe conditions, that reads

$$\max\left\{\left|\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k-1)})\right|, \left|\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k-1)}\right\rangle\right|\right\} \leq -\sigma_{\text{sw}}\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)}). \tag{SW}$$

The restart directions are then defined by

$$\boldsymbol{d}^{(k)} = \begin{cases} \boldsymbol{\delta}^{(k)} & \text{if } k = 0, \\ \theta_{(k)}\boldsymbol{\delta}^{(k)} + \beta_{(k)}\boldsymbol{d}^{(k-1)}, & \text{if } k \geq 1, \end{cases} \quad (\theta_{(k)}, \beta_{(k)}) = \begin{cases} (\tilde{\theta}_{(k)}, \tilde{\beta}_{(k)}) & \text{if } \boldsymbol{d}^{(k-1)} \text{ satisfies (SW),} \\ (1, 0) & \text{else.} \end{cases} \tag{FR MO I}$$

Proposition 14. *The directions in* (FR MO I) *have the sufficient decrease property* (3) *with* $\kappa_{\text{sd}} = 1$.

Proof. For $k = 0$ there is nothing to show. We do a proof by induction to show the general case.

Let $k \geq 1$. For $(\theta_{(k)}, \beta_{(k)}) = (1, 0)$ the property also holds trivially. Assume thus that (SW) is satisfied for k.

Then $\theta_{(k)}$ is non-negative: By induction, we have $\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)}) \leq 0$ and the strong Wolfe conditions thus imply the standard Wolfe conditions

$$\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k-1)}) \geq \sigma_{\text{sw}}\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)}).$$

We obtain

$$\theta_{(k)}\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2 = \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k-1)}) - \mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)}) \geq \underbrace{(\sigma_{\text{sw}} - 1)}_{<0}\underbrace{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)})}_{\leq 0} \geq 0.$$

Now take any $\boldsymbol{w} \in C$ and use the fact that $\theta_{(k)}$ and $\beta_{(k)}$ are non-negative:

$$\begin{aligned}
\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k)}\right\rangle &= \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\left(\theta_{(k)}\boldsymbol{\delta}^{(k)} + \beta_{(k)}\boldsymbol{d}^{(k-1)}\right)\right\rangle \\
&= \theta_{(k)}\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{\delta}^{(k)}\right\rangle + \beta_{(k)}\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k-1)}\right\rangle \\
&\leq \theta_{(k)}\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)}) + \beta_{(k)}\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k-1)}) \\
&\overset{(13)}{=} \frac{\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})\left(\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)}) - \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k-1)}) + \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k-1)})\right)}{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})} \\
&= \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})\frac{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)})}{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})}.
\end{aligned}$$

By induction $\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)}) \leq \mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})$, and finally

$$\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k)}\right\rangle \leq \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})\ \forall \boldsymbol{w} \in C \quad \Rightarrow \quad \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k)}) \leq \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)}).$$

□

We now show that the directions also fit Corollary 12.

Proposition 15. *Suppose Assumptions 1 to 3 hold and that the criticality* $\left\|\boldsymbol{\delta}^{(k)}\right\|$ *is bounded below like in* (12). *Then the Algorithm with directions defined by* (FR MO I) *generates a critical sequence.*

Proof. Denote by $\mathcal{P} \subseteq \mathbb{N}_0$ those iteration indices for which $\tilde{\theta}_{(k)} \geq 0$ and by $\mathcal{N} \subseteq \mathbb{N}_0$ the indices with $\tilde{\theta}_{(k)} < 0$. The case $\mathcal{P} = \emptyset$ reduces to $\boldsymbol{d}^{(k)} = \boldsymbol{\delta}^{(k)}$ for all k and

$$\sum_{k \in \mathbb{N}_0} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} = \sum_{k \in \mathbb{N}_0} \left\|\boldsymbol{\delta}^{(k)}\right\|^2 \geq \sum_{k \in \mathbb{N}_0} \varepsilon_{\text{crit}}^2 = \infty.$$

If $\mathcal{P} \neq \emptyset$, but still $|\mathcal{N}| = \infty$, then

$$\sum_{k \in \mathcal{N} \cup \mathcal{P}} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} = \underbrace{\sum_{k \in \mathcal{N}} \left\|\boldsymbol{\delta}^{(k)}\right\|^2}_{=\infty} + \underbrace{\sum_{k \in \mathcal{P}} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2}}_{\geq 0} = \infty.$$

Finally, assume that $|\mathcal{P}| = \infty$ and $|\mathcal{N}| < \infty$. Let k_0 be the maximal element in $\mathcal{N}$. For all $k > k_0$ it holds that $\theta_{(k)} = \tilde{\theta}_{(k)} \geq 0$ and $\beta_{(k)} = \tilde{\beta}_{(k)} \geq 0$. Squaring (FR MO I) for any $k \geq 1$ gives

$$\left\|\boldsymbol{d}^{(k)}\right\|^2 = \theta_{(k)}^2 \left\|\boldsymbol{\delta}^{(k)}\right\|^2 + \beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2 + 2\theta_{(k)}\beta_{(k)} \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k-1)} \right\rangle,$$

whilst multiplication with $\boldsymbol{\delta}^{(k)}$ results in

$$\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle = \theta_{(k)} \left\|\boldsymbol{\delta}^{(k)}\right\|^2 + \beta_{(k)} \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k-1)} \right\rangle,$$

giving

$$\left\|\boldsymbol{d}^{(k)}\right\|^2 = \beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2 - \theta_{(k)}^2 \left\|\boldsymbol{\delta}^{(k)}\right\|^2 + 2\theta_{(k)} \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle. \tag{14}$$

Let us take another look at $\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle$ and use this inequality:

$$\begin{aligned}
\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle &= \theta_{(k)} \left\|\boldsymbol{\delta}^{(k)}\right\|^2 + \beta_{(k)} \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k-1)} \right\rangle \\
&= \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \left(\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k-1)}) - \mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)}) + \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k-1)} \right\rangle \right) \\
&\leq \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \left(2 \max \left\{ \left| \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k-1)}) \right|, \left| \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k-1)} \right\rangle \right| \right\} - \mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)}) \right) \\
&\overset{\text{(SW)}}{\leq} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \left(-(1 + 2\sigma_{\text{sw}}) \mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)}) \right) \\
&= \left\|\boldsymbol{\delta}^{(k)}\right\|^2 \frac{(1 + 2\sigma_{\text{sw}}) \mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)})}{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})}, \qquad \text{for } k > k_0.
\end{aligned}$$

Because of the sufficient decrease property, $\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)}) \leq \mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})$, and $\sigma_{\text{sw}} \in (0, 1)$, we obtain

$$\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle \leq 3 \left\|\boldsymbol{\delta}^{(k)}\right\|^2. \tag{15}$$

Combining this with (14) gives

$$
\begin{aligned}
\frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} &= \frac{\beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} - \frac{\theta_{(k)}^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} + \frac{2\theta_{(k)} \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} \\
&\overset{(15)}{\leq} \frac{\beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} - \frac{\theta_{(k)}^2 \left\|\boldsymbol{\delta}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} + \frac{6\theta_{(k)} \left\|\boldsymbol{\delta}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} \\
&= \frac{\beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} - \frac{\theta_{(k)}^2 - 6\theta_{(k)}}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} \\
&= \frac{\beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} - \frac{(\theta_{(k)} - 3)^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} + \frac{9}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} \\
&\leq \frac{\beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} + \frac{9}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}. \qquad (16)
\end{aligned}
$$

In particular, for $k > k_0$, with $\beta_{(k)} = \left\|\boldsymbol{\delta}^{(k)}\right\|^2 / \left\|\boldsymbol{\delta}^{(k-1)}\right\|^2$, we find

$$
\frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} \leq \frac{\left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^4} + \frac{9}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} \leq \frac{\left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^4} + \frac{9}{\varepsilon_{\text{crit}}^2}.
$$

Recursion gives

$$
\frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} \leq \frac{\left\|\boldsymbol{d}^{(k_0)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k_0)}\right\|^4} + \sum_{i=k_0+1}^{k} \frac{9}{\varepsilon_{\text{crit}}^2} =: \mathtt{C}_0 + \frac{9(k-k_0)}{\varepsilon_{\text{crit}}^2} = 9\frac{\left(\frac{\mathtt{C}_0 \varepsilon_{\text{crit}}^2}{9} - k_0\right) + k}{\varepsilon_{\text{crit}}^2}.
$$

Summation of the reciprocals results in a divergent sum (because the harmonic series diverges):

$$
\sum_{k>k_0} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \geq \frac{1}{9} \sum_{k>k_0} \frac{\varepsilon_{\text{crit}}^2}{\left(\frac{\mathtt{C}_0 \varepsilon_{\text{crit}}^2}{9} - k_0\right) + k} = \infty.
$$

This concludes the proof, as Corollary 12 applies. □

4.2 FR Fractional-Linear Programming Variant

The single-objective directions have the nice property that $\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle \leq \left\|\boldsymbol{\delta}^{(k)}\right\|^2$ (in fact, equality holds). This gives both: The guaranteed descent and divergence of the series in Proposition 11. If we simply replace $\boldsymbol{g}^{(k)}$ with $-\boldsymbol{\delta}^{(k)}$, then we also obtain $\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle = \left\|\boldsymbol{\delta}^{(k)}\right\|^2$. But the descent property looks different in the multi-objective case, so instead we define

$$
\theta(\boldsymbol{w}) := \frac{\left\langle \boldsymbol{w}, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle - \left\langle \boldsymbol{w}, \boldsymbol{D}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle - (\mathtt{c}_{\text{FR}} - 1) \left\langle \boldsymbol{w}, \boldsymbol{D}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle}{-\mathtt{c}_{\text{FR}} \left\langle \boldsymbol{w}, \boldsymbol{D}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle} \quad \text{and}
$$

$$
\beta(\boldsymbol{w}) := \frac{-\left\langle \boldsymbol{w}, \boldsymbol{D}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle}{-\mathtt{c}_{\text{FR}} \left\langle \boldsymbol{w}, \boldsymbol{D}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle},
$$

with $\boldsymbol{D}^{(k)} = \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})$ and a constant $\mathtt{c}_{\text{FR}} > 1$. We then take $\boldsymbol{w}^*$ (neglecting an iteration index for the sake of readability) solving

$$
\min_{\boldsymbol{w} \in C} \frac{\left\langle \boldsymbol{w}, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle}{\left\langle \boldsymbol{w}, \boldsymbol{D}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle} \qquad (17)
$$

to define

$$
\boldsymbol{d}^{(k)} = \begin{cases} \boldsymbol{\delta}^{(k)} & \text{if } k = 0, \\ \theta_{(k)}(\boldsymbol{w}^*) \boldsymbol{\delta}^{(k)} + \beta_{(k)}(\boldsymbol{w}^*) \boldsymbol{d}^{(k-1)} & \text{if } k \geq 1. \end{cases} \qquad \text{(FR MO II)}
$$

Remark. This coefficient looks nicer and also works:

$$\theta(\boldsymbol{w}) := \frac{\left\langle \boldsymbol{w}, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle - \mathsf{c}_{\mathsf{FR}}\left\langle \boldsymbol{w}, \boldsymbol{D}^{(k-1)}\boldsymbol{d}^{(k-1)}\right\rangle}{-\mathsf{c}_{\mathsf{FR}}\left\langle \boldsymbol{w}, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle}$$

If C is discrete, (17) can be solved with a simple for-loop. Otherwise, it is a fractional-linear program and can be transformed to an LP. By noting that $\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle \le 0$ for all $\boldsymbol{v} \in C$ (including $\boldsymbol{w}^*$), we see from the minimizing property

$$\frac{\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle}{\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle} \le \frac{\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle}{\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle}$$

that

$$\frac{\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle}{\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle}\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle \ge \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle$$

and

$$\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle \le \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle \quad \text{for all } \boldsymbol{v} \in C. \tag{18}$$

We first use this to show sufficient decrease:

Proposition 16. *The directions in* (FR MO II) *have the sufficient decrease property* (3).

Proof. Again, the proof is by induction and we show that

$$\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k)}\right\rangle \le \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle \le 0 \qquad \forall k \ge 0,$$

which implies the sufficient decrease property. The case $k = 0$ is trivial. Let $k \ge 1$ and let $\boldsymbol{v} \in C$. Equation (18) shows that

$$\psi(\boldsymbol{w}^*, \boldsymbol{v}) := \left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle - \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle \le 0.$$

Using the definition (FR MO II) we get

$$\begin{aligned}\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k)}\right\rangle &= \frac{\left(-\mathsf{c}_{\mathsf{FR}}\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{d}^{(k-1)}\right\rangle - (\mathsf{c}_{\mathsf{FR}} - 1)\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle\right)\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle + \psi(\boldsymbol{w}^*, \boldsymbol{v})}{-\mathsf{c}_{\mathsf{FR}}\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle} \\ &\overset{(18)}{\le} \frac{\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{d}^{(k-1)}\right\rangle\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle}{\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle} + \frac{c-1}{c}\frac{\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle}{\mathsf{c}_{\mathsf{FR}}\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle} \\ &= \frac{\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{d}^{(k-1)}\right\rangle\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle}{\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle} + \left(1 - \frac{1}{c}\right)\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle.\end{aligned}$$

Because of

$$0 \le \frac{\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle}{\mathsf{c}_{\mathsf{FR}}\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle},$$

we can use the induction hypothesis

$$\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{d}^{(k-1)}\right\rangle \le \left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle$$

to finally get

$$\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k)}\right\rangle \le \left(\frac{1}{\mathsf{c}_{\mathsf{FR}}} + 1 - \frac{1}{\mathsf{c}_{\mathsf{FR}}}\right)\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle = \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle \le 0.$$

□

Proposition 17. *Suppose that Assumptions 1 and 4 hold. The Algorithm with directions* (FR MO II) *produces a critical sequence.*

Proof. For a proof by contradiction, assume that the criticality is bounded like in (12). We use the triangle inequality and Cauchy-Schwarz on (FR MO II) to get

$$\left\|\boldsymbol{d}^{(k)}\right\| \le \left|\theta_{(k)}\right|\left\|\boldsymbol{\delta}^{(k)}\right\| + \left|\beta_{(k)}\right|\left\|\boldsymbol{d}^{(k-1)}\right\|.$$

We first investigate $\left|\theta_{(k)}\right| \left\|\boldsymbol{\delta}^{(k)}\right\|$:

$$\begin{aligned}
\left|\theta_{(k)}\right| \left\|\boldsymbol{\delta}^{(k)}\right\| &\leq \left| \frac{\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{d}^{(k-1)}\right\rangle - (\mathsf{c}_{\mathsf{FR}} - 1)\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle}{-\mathsf{c}_{\mathsf{FR}}\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle} \right| \left\|\boldsymbol{\delta}^{(k)}\right\| \\
&\leq \frac{\left\|\boldsymbol{w}^*\right\| \left\|\boldsymbol{D}^{(k)} - \boldsymbol{D}^{(k-1)}\right\| \left\|\boldsymbol{d}^{(k-1)}\right\| \left\|\boldsymbol{\delta}^{(k-1)}\right\| \left\|\boldsymbol{\delta}^{(k)}\right\|}{\mathsf{c}_{\mathsf{FR}}\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle} + \frac{\mathsf{c}_{\mathsf{FR}} - 1}{\mathsf{c}_{\mathsf{FR}}} \left\|\boldsymbol{\delta}^{(k)}\right\| \\
&\leq \frac{\left(\sup_{\boldsymbol{w}\in C}\left\|\boldsymbol{w}\right\|\right) \mathtt{L}_f \left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right\| \left\|\boldsymbol{d}^{(k-1)}\right\| \left(\sup_{k\geq 0}\left\|\boldsymbol{\delta}^{(k-1)}\right\|\right)^2}{\mathsf{c}_{\mathsf{FR}}\varepsilon_{\mathrm{crit}}^2} + \frac{\mathsf{c}_{\mathsf{FR}} - 1}{\mathsf{c}_{\mathsf{FR}}} \sup_{k\geq 0}\left\|\boldsymbol{\delta}^{(k-1)}\right\| \\
&\leq \mathtt{C}_1 \left\|\sigma_{(k-1)}\boldsymbol{d}^{(k-1)}\right\| \left\|\boldsymbol{d}^{(k-1)}\right\| + \mathtt{C}_2 \qquad (19)
\end{aligned}$$

for some constants $\mathtt{C}_1, \mathtt{C}_2 > 0$, which we have derived using the boundedness of C and the sublevelset (Assumption 4).

To bound $\left|\beta_{(k)}\right| \left\|\boldsymbol{d}^{(k-1)}\right\|$, we note that because of the assumptions and [22], we know the steepest descent direction to be $\mathtt{H}$-Hölder continuous:

$$\left\|\boldsymbol{\delta}^{(k)} - \boldsymbol{\delta}^{(k-1)}\right\| = \left\|\boldsymbol{\delta}^{(k-1)} - \boldsymbol{\delta}^{(k)}\right\| \leq \mathtt{H}\left\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k)}\right\|^{\frac{1}{2}} = \mathtt{H}\left\|\sigma^{(k-1)}\boldsymbol{d}^{(k-1)}\right\|^{\frac{1}{2}}.$$

Hölder-continuity is show in [22] for the case $\mathcal{K} = \mathbb{R}^N$. Should work for other cones as well, but better check!

Now

$$\begin{aligned}
\left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle\right| - \left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle\right| &\leq \left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle\right| \\
&\leq \left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k)}\right\rangle + \left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k)}\right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle\right| \\
&\leq \left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k)}\right\rangle\right| + \left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k)}\right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle\right| \\
&\leq L_1\left\|\sigma_{(k-1)}\boldsymbol{d}^{(k-1)}\right\| + H_1\left\|\sigma_{(k-1)}\boldsymbol{d}^{(k-1)}\right\|^{\frac{1}{2}}.
\end{aligned}$$

The constant $L_1 > 0$ follows from the Lipschitz-continuity of the Jacobian and the boundedness of C and the sublevelset, the constant $H_1 > 0$ is derived from the Hölder-continuity of the steepest descent direction. It follows that

$$\left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle\right| \leq L_1\left\|\sigma_{(k-1)}\boldsymbol{d}^{(k-1)}\right\| + H_1\left\|\sigma_{(k-1)}\boldsymbol{d}^{(k-1)}\right\|^{\frac{1}{2}} + \left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle\right|$$

and

$$\frac{\left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle\right|}{\mathsf{c}_{\mathsf{FR}}\left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle\right|} \leq L_2\left\|\sigma_{(k-1)}\boldsymbol{d}^{(k-1)}\right\| + H_2\left\|\sigma_{(k-1)}\boldsymbol{d}^{(k-1)}\right\|^{\frac{1}{2}} + \frac{1}{\mathsf{c}_{\mathsf{FR}}}.$$

The constants L_2 and H_2 follow from the fact that $\mathsf{c}_{\mathsf{FR}}\left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle\right|$ is bounded below by $\mathsf{c}_{\mathsf{FR}}\varepsilon_{\mathrm{crit}}^2$.

Because the steps vanish according to Lemma 10, and because of $\mathsf{c}_{\mathsf{FR}}^{-1} \in (0, 1)$, there must be some $k_0 \in \mathbb{N}_0$ and a constant $\varepsilon_1 \in (0, 1)$ such that

$$\frac{\left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle\right|}{\mathsf{c}_{\mathsf{FR}}\left|\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle\right|} \leq L_2\left\|\sigma_{(k-1)}\boldsymbol{d}^{(k-1)}\right\| + H_2\left\|\sigma_{(k-1)}\boldsymbol{d}^{(k-1)}\right\|^{\frac{1}{2}} + \frac{1}{\mathsf{c}_{\mathsf{FR}}} \leq \varepsilon_1 < 1 \qquad \forall k \geq k_0.$$

That is,

$$\left|\beta_{(k)}\right| \left\|\boldsymbol{d}^{(k-1)}\right\| \leq \varepsilon_1\left\|\boldsymbol{d}^{(k-1)}\right\| \qquad \forall k \geq k_0.$$

For (19) we can make a similar argument: As $\varepsilon_1 < 1$, there is some $k_1 \geq k_0$ and $\varepsilon_2 > 0$ with $\mathtt{r} = \varepsilon_1 + \varepsilon_2 < 1$ such that

$$\left|\theta_{(k)}\right| \left\|\boldsymbol{\delta}^{(k)}\right\| \leq \varepsilon_2\left\|\boldsymbol{d}^{(k-1)}\right\| + \mathtt{C}_2 \qquad \forall k \geq k_0.$$

Finally,

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \left|\theta_{(k)}\right| \left\|\boldsymbol{\delta}^{(k)}\right\| + \left|\beta_{(k)}\right| \left\|\boldsymbol{d}^{(k-1)}\right\| \leq \mathtt{r}\left\|\boldsymbol{d}^{(k-1)}\right\| + \mathtt{C}_2 \quad \forall k \geq k_1.$$

Hence, $\boldsymbol{d}^{(k)}$ is bounded (there is a proof below and I have to re-sort the whole document). This concludes our proof. □

FR MaxiMin Variant

The other variant based on the single-objective scheme (FR SO) directly exploits the definition of $\mathfrak{f}_{\boldsymbol{x}}^{(k)}$ as a maximization problem over C. More precisely, for $k \geq 1$ and $\boldsymbol{w} \in C$, define the coefficients

$$\theta_{(k)}(\boldsymbol{w}) := \frac{\left\langle \boldsymbol{w}, \left(\nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) - \nabla \boldsymbol{f}(\boldsymbol{x}^{(k-1)})\right) \boldsymbol{d}^{(k-1)} \right\rangle}{-\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})} \quad \text{and} \quad \beta_{(k)}(\boldsymbol{w}) = \frac{-\left\langle \boldsymbol{w}, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \boldsymbol{\delta}^{(k)} \right\rangle}{-\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)})}. \tag{20}$$

Then, choose $\boldsymbol{v}^* = \boldsymbol{v}^*_{(k)}$ and $\boldsymbol{w}^* = \boldsymbol{w}^*_{(k)}$ to solve

$$\max_{\boldsymbol{v} \in C} \min_{\boldsymbol{w} \in C} \left\langle \boldsymbol{v}, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \left(\theta_{(k)}(\boldsymbol{w}) \boldsymbol{\delta}^{(k)} + \beta_{(k)}(\boldsymbol{w}) \boldsymbol{d}^{(k-1)} \right) \right\rangle. \tag{21}$$

The directions follow as

$$\boldsymbol{d}^{(k)} = \begin{cases} \boldsymbol{\delta}^{(k)} & \text{if } k = 0, \\ \theta_{(k)}(\boldsymbol{w}^*) \boldsymbol{\delta}^{(k)} + \beta_{(k)}(\boldsymbol{w}^*) \boldsymbol{d}^{(k-1)} & \text{if } k \geq 1. \end{cases} \tag{FR MO II}$$

Note, that (21) can be cheaply solved if C is discrete (i.e., $\mathcal{K}^*$ is finitely generated). It then boils down to a few For-loops. Alternatively, linear programming can be applied to the convex relaxation, as the problem cost function is linear in both arguments. Depending on how the problem is solved, a MiniMax problem might seem preferable, but equality does not necessarily hold here.

Proposition 18. *The directions in* (FR MO II) *have the sufficient decrease property* (3) *with* $\kappa_{\mathrm{sd}} = 1$.

Proof. Again, the proof is by induction. The property trivially holds for $k = 0$. Let $k \geq 1$. The definition (FR MO II) gives

$$\begin{aligned}
\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k)}) &= \left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \boldsymbol{d}^{(k)} \right\rangle \\
&= \left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \left(\theta_{(k)}(\boldsymbol{w}^*) \boldsymbol{\delta}^{(k)} + \beta_{(k)}(\boldsymbol{w}^*) \boldsymbol{d}^{(k-1)} \right) \right\rangle \\
&= \theta_{(k)}(\boldsymbol{w}^*) \left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \boldsymbol{\delta}^{(k)} \right\rangle + \beta_{(k)}(\boldsymbol{w}^*) \left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \boldsymbol{d}^{(k-1)} \right\rangle \\
&\overset{(21)}{\leq} \theta_{(k)}(\boldsymbol{v}^*) \left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \boldsymbol{\delta}^{(k)} \right\rangle + \beta_{(k)}(\boldsymbol{v}^*) \left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \boldsymbol{d}^{(k-1)} \right\rangle \\
&\overset{(20)}{=} \frac{\left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \boldsymbol{\delta}^{(k)} \right\rangle}{-\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})} \left(\left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \boldsymbol{d}^{(k-1)} \right\rangle - \left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k-1)}) \boldsymbol{d}^{(k-1)} \right\rangle - \left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \boldsymbol{d}^{(k-1)} \right\rangle \right) \\
&\leq \frac{\left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \boldsymbol{\delta}^{(k)} \right\rangle \left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k-1)}) \boldsymbol{d}^{(k-1)} \right\rangle}{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})} \\
&\leq \frac{\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)}) \mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)})}{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})}.
\end{aligned}$$

For the last line we have used the fact, that $\left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)}) \boldsymbol{\delta}^{(k)} \right\rangle \leq 0$ and $\left\langle \boldsymbol{v}^*, \nabla \boldsymbol{f}(\boldsymbol{x}^{(k-1)}) \boldsymbol{d}^{(k-1)} \right\rangle \leq 0$ (by induction), while repeatedly applying the definition of $\mathfrak{f}_{\boldsymbol{x}}$ as a maximum operator. Finally, sufficient decrease follows because of $\frac{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{d}^{(k-1)})}{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})} \leq 1$. □

To make the MaxiMin directions fit our convergence analysis, we now require a bounded sublevel-set. That makes the subsequent proof more akin to that of the directions in the next section than to the prior FR variant with restarts. Alternatively, we could impose angle conditions on the gradients, which feel more complicated and unconventional.

Proposition 19. *Suppose Assumptions 1 and 4 hold and that the criticality* $\left\| \boldsymbol{\delta}^{(k)} \right\|$ *is bounded below like in* (12)*. Then the Algorithm with modified Armijo-stepsizes according to* (5) *and with directions defined by* (FR MO II) *generates a critical subsequence.*

Proof. Assumption 4 implies Lipschitz continuity of the Jacobians as per Assumption 2. Furthermore, all gradients and Jacobians are uniformly bounded, say by $\mathtt{C}_g > 0$. Because $\boldsymbol{\delta}^{(k)}$ is a sum of negative gradients weighted by coefficients from the bounded set C, the steepest descent direction is also uniformly bounded with $\mathtt{C}_{\boldsymbol{\delta}} > 0$ like so:

$$\varepsilon_{\mathrm{crit}} \leq \left\| \boldsymbol{\delta}^{(k)} \right\| \leq \mathtt{C}_{\boldsymbol{\delta}}, \qquad \forall k \in \mathbb{N}_0.$$

We use the triangle inequality and Cauchy-Schwarz on (FR MO II) to get

$$\begin{aligned}\left\|\boldsymbol{d}^{(k)}\right\| &\leq \left|\theta_{(k)}\right| \left\|\boldsymbol{\delta}^{(k)}\right\| + \left|\beta_{(k)}\right| \left\|\boldsymbol{d}^{(k-1)}\right\| \\ &\leq \frac{1}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \left(\left|\left\langle \boldsymbol{w}^*, \left(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) - \boldsymbol{\nabla f}(\boldsymbol{x}^{(k-1)})\right) \boldsymbol{d}^{(k-1)}\right\rangle\right| \left\|\boldsymbol{\delta}^{(k)}\right\| + \left|\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{\delta}^{(k)}\right\rangle\right| \left\|\boldsymbol{d}^{(k-1)}\right\|\right) \\ &\leq \frac{\|\boldsymbol{w}^*\| \left\|\boldsymbol{\delta}^{(k)}\right\| \left\|\boldsymbol{d}^{(k-1)}\right\|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \left(\left\|\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) - \boldsymbol{\nabla f}(\boldsymbol{x}^{(k-1)})\right\| + \left\|\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\right\|.\right)\end{aligned}$$

We can now plug in all the bounds (and implicitly use the fact that C is compact to bound $\|\boldsymbol{w}^*\|$), before using the Lipschitz continuity of the Jacobians:

$$\begin{aligned}\left\|\boldsymbol{d}^{(k)}\right\| &\leq \tilde{\mathtt{C}} \left\|\boldsymbol{d}^{(k-1)}\right\| \left(\left\|\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) - \boldsymbol{\nabla f}(\boldsymbol{x}^{(k-1)})\right\| + \mathtt{C}_g\right) \\ &\leq \bar{\mathtt{C}} \left\|\boldsymbol{d}^{(k-1)}\right\| \left(\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right\| + \mathtt{C}_g\right).\end{aligned}$$

Now $\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right\| = \sigma_{(k-1)} \left\|\boldsymbol{d}^{(k-1)}\right\|$ and this goes to zero due to Lemma 10. Ab hier ist falsch! There is some $k_0 \in \mathbb{N}_0$ with $norm\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \leq \mathtt{r}$ for all $k \geq k_0$, where $\mathtt{r} > 0$ is small enough so as to meet

$$\mathtt{C} := \bar{\mathtt{C}}(\mathtt{r} + \mathtt{C}_g) \in (0, 1).$$

□

4.3 Three-Term Polak-Ribière-Polyak Scheme

Zhang et al. [24] define the following three-term Polak-Ribière-Polyak (PRP) directions for single-objective optimization:

$$\boldsymbol{d}^{(k)} = \begin{cases} -\boldsymbol{g}^{(k)} & \text{if } k = 0, \\ -\boldsymbol{g}^{(k)} + \beta_{(k)}\boldsymbol{d}^{(k-1)} - \theta_{(k)}\left(\boldsymbol{g}^{(k)} - \boldsymbol{g}^{(k-1)}\right) & \text{if } k \geq 1. \end{cases}$$

Here, $\boldsymbol{g}^{(k)} = \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})$ and the coefficients are defined by

$$\beta_{(k)} = \frac{\left\langle \boldsymbol{g}^{(k)}, \boldsymbol{g}^{(k)} - \boldsymbol{g}^{(k-1)}\right\rangle}{\left\|\boldsymbol{g}^{(k-1)}\right\|^2} \quad \text{and} \quad \theta_{(k)} = \frac{\left\langle \boldsymbol{g}^{(k)}, \boldsymbol{d}^{(k-1)}\right\rangle}{\left\|\boldsymbol{g}^{(k-1)}\right\|^2}.$$

These directions provide guaranteed descent as per $\left\langle -\boldsymbol{g}^{(k)}, \boldsymbol{d}^{(k)}\right\rangle = \left\|\boldsymbol{\delta}^{(k)}\right\|^2$.

Unfortunately, it is not sufficient to simply replace $-\boldsymbol{g}^{(k)}$ with $\boldsymbol{\delta}^{(k)}$, the minimizer in (2), to obtain a multi-objective scheme. If C is discrete, we cannot use the linear-fractional programming approach neither. To see this, take a look at the suggested multi-objective coefficients (for $k \geq 1$ and $\boldsymbol{w} \in C$):

$$\beta_{(k)}(\boldsymbol{w}) = \frac{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{y}^{(k)}\right\rangle}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \quad \text{and} \quad \theta_{(k)}(\boldsymbol{w}) = \frac{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k-1)}\right\rangle}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2}, \tag{22}$$

where $\boldsymbol{y}^{(k)} = \boldsymbol{\delta}^{(k-1)} - \boldsymbol{\delta}^{(k)}$. If we wanted to use a common $\boldsymbol{w}^*$ to define

$$\boldsymbol{d}^{(k)} = \boldsymbol{\delta}^{(k)} + \beta_{(k)}\boldsymbol{d}^{(k-1)} - \theta_{(k)}\boldsymbol{y}^{(k)}$$

and show sufficient decrease, this would require the following inequality to hold:

$$\left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{y}^{(k)}\right\rangle \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{D}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)}\boldsymbol{y}^{(k)}\right\rangle \leq 0 \qquad \forall \boldsymbol{v} \in C.$$

In the non-discrete case, $\boldsymbol{w}*$ could be chosen to be the MiniMax solution of the skew-symmetric bilinear form on the LHS. By von Neumann's theorem, it equals the MaxiMin and the optimal value is 0. But in the discrete case, we do not know, if the bilinear mapping describes a generalized rock-paper-scissors game (prohibiting the MiniMax) apporach) or not. A transformation into a linear-fractional problem is not possible, because we cannot assume a constant sign for either of the factors of the expression above.

References: von Neumann and Generalized Rock Paper Scissors

To circumvent these issues, we introduce two additional iteration-dependent scaling factors $\alpha_\beta \in [0,1]$ and $\alpha_\theta \in [0,1]$. After getting $(\boldsymbol{v}_\beta, \boldsymbol{w}_\beta)$ from solving

$$\min_{\boldsymbol{w}} \max_{\boldsymbol{v}} \left\langle \boldsymbol{w}, \boldsymbol{D}^{(k)} \boldsymbol{y}^{(k)} \right\rangle \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle$$

and $(\boldsymbol{v}_\theta, \boldsymbol{w}_\theta)$ from

$$\max_{\boldsymbol{w}} \min_{\boldsymbol{v}} \left\langle \boldsymbol{w}, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)} \boldsymbol{y}^{(k)} \right\rangle,$$

we only have

$$\psi(\boldsymbol{v}_\beta, \boldsymbol{w}_\beta) := \left\langle \boldsymbol{w}_\beta, \boldsymbol{D}^{(k)} \boldsymbol{y}^{(k)} \right\rangle \left\langle \boldsymbol{v}_\beta, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \ge \left\langle \boldsymbol{w}_\theta, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \left\langle \boldsymbol{v}_\theta, \boldsymbol{D}^{(k)} \boldsymbol{y}^{(k)} \right\rangle = \psi(\boldsymbol{w}_\theta, \boldsymbol{v}_\theta)$$

But we can determine α_β and α_θ to enforce equality:

$$\alpha_\beta \psi(\boldsymbol{v}_\beta, \boldsymbol{w}_\beta) = \alpha_\theta \psi(\boldsymbol{w}_\theta, \boldsymbol{v}_\theta)$$

Consider three cases:

- If equality would require a sign switch, $\psi(\boldsymbol{w}_\theta, \boldsymbol{v}_\theta) \le 0$ and $\psi(\boldsymbol{v}_\beta, \boldsymbol{w}_\beta) \ge 0$, then set $\alpha_\beta = \alpha_\theta = 0$.
- If both factors are positive, shrink the larger factor $\psi(\boldsymbol{v}_\beta, \boldsymbol{w}_\beta)$ by multiplication with
$$\alpha_\beta = \frac{\psi(\boldsymbol{w}_\theta, \boldsymbol{v}_\theta)}{\psi(\boldsymbol{v}_\beta, \boldsymbol{w}_\beta)} \in (0,1],$$
and set $\alpha_\theta = 1$.
- If both factors are negative, grow the smaller factor $\psi(\boldsymbol{w}_\theta, \boldsymbol{v}_\theta)$ by multiplication with
$$\alpha_\theta = \frac{\psi(\boldsymbol{v}_\beta, \boldsymbol{w}_\beta)}{\psi(\boldsymbol{w}_\theta, \boldsymbol{v}_\theta)} \in (0,1]$$
and set $\alpha_\beta = 1$.

Then (finally), let

$$\boldsymbol{d}^{(k)} = \begin{cases} \boldsymbol{\delta}^{(k)} & \text{if } k = 0 \\ \boldsymbol{\delta}^{(k)} + \alpha_\beta \beta_{(k)}(\boldsymbol{w}_\beta) \boldsymbol{d}^{(k-1)} - \alpha_\theta \theta_{(k)}(\boldsymbol{w}_\theta) \boldsymbol{y}^{(k)}, & \text{if } k \ge 1. \end{cases} \tag{PRP MO I}$$

Proposition 20. *The directions in* (PRP MO I) *have the sufficient decrease property* (3) *with* $\kappa_{\text{sd}} = 1$.

Proof. The case $k = 0$ is trivial. Let $k \ge 1$ and $\boldsymbol{v} \in C$. Because α_β is non-negative and $\boldsymbol{v}_\beta$ is a maximizer,

$$\begin{aligned} \alpha_\beta \psi(\boldsymbol{v}, \boldsymbol{w}_\beta) &= \alpha_\beta \left\langle \boldsymbol{w}_\beta, \boldsymbol{D}^{(k)} \boldsymbol{y}^{(k)} \right\rangle \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \\ &\le \alpha_\beta \max_{\boldsymbol{v}} \left\langle \boldsymbol{w}_\beta, \boldsymbol{D}^{(k)} \boldsymbol{y}^{(k)} \right\rangle \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \\ &= \alpha_\beta \left\langle \boldsymbol{w}_\beta, \boldsymbol{D}^{(k)} \boldsymbol{y}^{(k)} \right\rangle \left\langle \boldsymbol{v}_\beta, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle = \alpha_\beta \psi(\boldsymbol{v}_\beta, \boldsymbol{w}_\beta). \end{aligned}$$

Furthermore, because α_θ is nonnegative and $\boldsymbol{v}_\theta$ is a minimizer,

$$\begin{aligned} \alpha_\theta \psi(\boldsymbol{w}_\theta, \boldsymbol{v}_\theta) &= \alpha_\theta \left\langle \boldsymbol{w}_\theta, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \left\langle \boldsymbol{v}_\theta, \boldsymbol{D}^{(k)} \boldsymbol{y}^{(k)} \right\rangle \\ &\le \alpha_\theta \left\langle \boldsymbol{w}_\theta, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)} \boldsymbol{y}^{(k)} \right\rangle = \alpha_\theta \psi(\boldsymbol{w}_\theta, \boldsymbol{v}). \end{aligned}$$

By the way we have set up α_β and α_θ, we can combine both inequalities:

$$\alpha_\beta \psi(\boldsymbol{v}, \boldsymbol{w}_\beta) = \alpha_\beta \left\langle \boldsymbol{w}_\beta, \boldsymbol{D}^{(k)} \boldsymbol{y}^{(k)} \right\rangle \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \le \alpha_\theta \left\langle \boldsymbol{w}_\theta, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)} \boldsymbol{y}^{(k)} \right\rangle = \alpha_\theta \psi(\boldsymbol{w}_\theta, \boldsymbol{v}). \tag{23}$$

This enables us to discard those terms in the inner product:

$$\left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)} \boldsymbol{d}^{(k)} \right\rangle = \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle + \underbrace{\alpha_\beta \psi(\boldsymbol{v}, \boldsymbol{w}_\beta) - \alpha_\theta \psi(\boldsymbol{w}_\theta, \boldsymbol{v})}_{\le 0} \le \left\langle \boldsymbol{v}, \boldsymbol{D}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \le 0.$$

□

Proving convergence for these specific directions requires stricter assumptions than provided by Assumption 3:

Proposition 21. *Suppose that Assumptions 1 and 4 hold. The Algorithm with directions* (PRP MO I) *produces a critical sequence.*

Proof. For a proof by contradiction, assume that the criticality $\left\|\boldsymbol{\delta}^{(k)}\right\|$ is bounded below like in (12). Because of Assumptions 1 and 4, the norm of the steepest descent is also uniformly bounded above by a constant $\mathtt{C}_{\boldsymbol{\delta}} > 0$. If we can show the same for $\boldsymbol{d}^{(k)}$, that concludes the proof because of Corollary 13.

The proof below can be easily adapted to include the factors $\alpha_\beta, \alpha_\theta \in [0,1]$.

Assume $k \geq 1$. Apply the triangle inequality and Cauchy-Schwarz to the definition of $\boldsymbol{d}^{(k)}$:

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \left\|\boldsymbol{\delta}^{(k)}\right\| + \frac{2\left\|\boldsymbol{w}^* \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\right\| \left\|\boldsymbol{\delta}^{(k-1)} - \boldsymbol{\delta}^{(k)}\right\|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \left\|\boldsymbol{d}^{(k-1)}\right\|. \tag{24}$$

Because of our assumptions, the steepest descent direction is Hölder-continuous. Furthermore, the Jacobians must be bounded and because C is compact, there is some constant $\mathtt{C}_C > 0$ with $\left\|\boldsymbol{w}^* \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\right\| \leq \mathtt{C}_C$ for all k. Thus, (24) leads to

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \mathtt{C}_{\boldsymbol{\delta}} + \frac{2\mathtt{C}_C \mathtt{H} \left\|\sigma^{(k-1)}\boldsymbol{d}^{(k-1)}\right\|^{\frac{1}{2}}}{\varepsilon_{\mathrm{crit}}^2} \left\|\boldsymbol{d}^{(k-1)}\right\|. \tag{25}$$

With the Armijo condition, Lemma 10 is applicable and $\left\|\sigma^{(k-1)}\boldsymbol{d}^{(k-1)}\right\|^{\frac{1}{2}}$ vanishes. There must be $\mathtt{r} \in (0,1)$ and $k_0 \in \mathbb{N}_0$ with

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \mathtt{C}_{\boldsymbol{\delta}} + \mathtt{r}\left\|\boldsymbol{d}^{(k-1)}\right\| \forall k \geq k_0. \tag{26}$$

Repeated application of (26) leads to a geometric sum (for $k \geq k_0$):

$$\begin{aligned}\left\|\boldsymbol{d}^{(k)}\right\| &\leq \mathtt{C}_{\boldsymbol{\delta}} + \mathtt{r}\left\|\boldsymbol{d}^{(k-1)}\right\| \leq \mathtt{C}_{\boldsymbol{\delta}} + \mathtt{r}\left(\mathtt{C}_{\boldsymbol{\delta}} + \mathtt{r}\left\|\boldsymbol{d}^{(k-2)}\right\|\right) \leq \dots \\ &\leq \mathtt{C}_{\boldsymbol{\delta}}\left(1 + \mathtt{r} + \dots + \mathtt{r}^{k-k_0-1}\right) + \mathtt{r}^{k-k_0}\left\|\boldsymbol{d}^{(k_0)}\right\| \\ &\leq \frac{\mathtt{C}_{\boldsymbol{\delta}}}{1-\mathtt{r}} + \left\|\boldsymbol{d}^{(k_0)}\right\|.\end{aligned}$$

Hence, for all k the directions $\{\boldsymbol{d}^{(k)}\}$ are also bounded by

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \max\left\{\left\|\boldsymbol{d}^{(0)}\right\|, \left\|\boldsymbol{d}^{(1)}\right\|, \dots, \left\|\boldsymbol{d}^{(k_0-1)}\right\|, \frac{\mathtt{C}_{\boldsymbol{\delta}}}{1-\mathtt{r}} + \left\|\boldsymbol{d}^{(k_0)}\right\|\right\} =: \mathtt{C}_{\boldsymbol{d}}.$$

Corollary 13 finally gives convergence. □

4.4 Cone-Projection PRP Scheme

To ensure sufficient decrease, Cheng [1] simply project the residual term in a standard two-term PRP scheme onto the null space of $\nabla f(\boldsymbol{x}^{(k)})$. With multiple gradients, there usually is no single orthogonal space for all of them. However, the (convex) cone of non-ascent directions

$$\mathcal{D}(\boldsymbol{x}) = \left\{\boldsymbol{v} \in \mathbb{R}^N : \boldsymbol{\nabla f}(\boldsymbol{x})\boldsymbol{v} \in -\mathcal{K}\right\}$$

is polar to the gradient cone $\boldsymbol{\nabla f}(\boldsymbol{x})^{\mathsf{T}}\mathcal{K}^*$ and provides a suitable generalization. Note, that the properties of φ characterize $\mathcal{D}$ via

$$\boldsymbol{v} \in \mathcal{D}(\boldsymbol{x}) \Leftrightarrow \varphi(\boldsymbol{\nabla f}(\boldsymbol{x})\boldsymbol{v}) \leq 0.$$

To motivate the approach, let us revisit the single-objective definition of $\boldsymbol{d}^{(k)}$. To this end, let $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$ and denote for any vector $\boldsymbol{v} \in \mathbb{R}^N$ its null space/orthogonal complement by $\ker(\boldsymbol{v})$. From [1] we take

$$\boldsymbol{d}^{(k)} = \begin{cases} -\boldsymbol{g}^{(k)} & \text{if } k = 0, \\ -\boldsymbol{g}^{(k)} + \bar{\boldsymbol{d}}^{(k)} & \text{if } k \geq 1, \end{cases}$$

where

$$\bar{\boldsymbol{d}}^{(k)} = \mathfrak{P}_{\ker(\boldsymbol{g}^{(k)})}\left(\beta_{(k)}\boldsymbol{d}^{(k-1)}\right), \quad \beta_{(k)} = \frac{\langle \boldsymbol{g}^{(k)}, \boldsymbol{g}^{(k)} - \boldsymbol{g}^{(k-1)}\rangle}{\left\|\boldsymbol{g}^{(k-1)}\right\|^2}.$$

The parameter $\beta_{(k)}$ is the usual PRP parameter. $\mathfrak{P}_{\bullet}(\bullet)$ is the metric projection operator and for a single vector, the projection onto its null space is given by the simple formula

$$\mathfrak{P}_{\ker(\boldsymbol{g}^{(k)})}\left(\beta_{(k)}\boldsymbol{d}^{(k-1)}\right) = \left(\boldsymbol{I}_{N\times N} - \frac{\boldsymbol{g}^{(k)}\left(\boldsymbol{g}^{(k)}\right)^{\mathsf{T}}}{\left\|\boldsymbol{g}^{(k)}\right\|^2}\right)\beta_{(k)}\boldsymbol{d}^{(k-1)}.$$

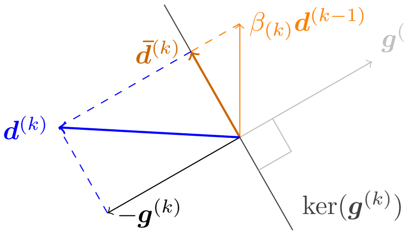
$\beta_{(k)}\boldsymbol{d}^{(k-1)}$
$\boldsymbol{g}^{(k)}$
$\bar{\boldsymbol{d}}^{(k)}$
$\boldsymbol{d}^{(k)}$
$-\boldsymbol{g}^{(k)}$
$\ker(\boldsymbol{g}^{(k)})$

Figure 1: Projection Scheme in the single-objective setting. The standard residual term is projected onto the plane orthogonal to the gradient, so that the fincal CG direction (blue) has sufficient decrease by construction.

To go to the multi-objective setting, use the PRP coefficient as in [14]. (If we used a coefficient similar to (22) we would again need Assumption 4 in the convergence proof.)

$$\beta_{(k)} = \frac{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k)}) - \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})}{-\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})}, \quad k \geq 1. \tag{27}$$

Moreover, for all $k \geq 1$, let $M^{(k)}$ be a non-empty convex subset of $\mathcal{D}(\boldsymbol{x}^{(k)})$, containing the origin. Lastly, define

$$\boldsymbol{d}^{(k)} = \begin{cases}\boldsymbol{\delta}^{(k)} & \text{if } k = 0,\\ \boldsymbol{\delta}^{(k)} + \bar{\boldsymbol{d}}^{(k)} & \text{if } k \geq 1,\end{cases} \qquad \bar{\boldsymbol{d}}^{(k)} = \mathfrak{P}_{M^{(k)}}\left(\beta_{(k)}\boldsymbol{d}^{(k-1)}\right). \tag{PRP MO II}$$

Remark. We can always use $M^{(k)} = \mathcal{D}(\boldsymbol{x}^{(k)})$, but then the projection might be too expensive. To exploit the single-vector projection formula from above, we can use a MiniMax approach, at least if C is discrete. Choose $\boldsymbol{w}^*$ as the minimizer in

$$\min_{\boldsymbol{w}\in C}\max_{\boldsymbol{v}\in C}\left\langle \boldsymbol{v}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\mathfrak{P}_{\ker(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})^{\mathsf{T}}\boldsymbol{w})}\left(\beta_{(k)}\boldsymbol{d}^{(k-1)}\right)\right\rangle$$

If the optimal at $\boldsymbol{w}$ is less than or equal to 0, then, by the properties of φ, the vector

$$\mathfrak{P}_{\ker(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})^{\mathsf{T}}\boldsymbol{w}^*)}\left(\beta_{(k)}\boldsymbol{d}^{(k-1)}\right),$$

a projection onto the hyperplane $\ker(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})^{\mathsf{T}}\boldsymbol{w}^*)$, is contained in $\mathcal{D}(\boldsymbol{x}^{(k)})$, and we can use this as $\bar{\boldsymbol{d}}^{(k)}$. If the optimal value is positive, then $\beta_{(k)}\boldsymbol{d}^{(k-1)}$ belongs to the polar cone of $\mathcal{D}(\boldsymbol{x}^{(k)})$ and $\bar{\boldsymbol{d}}^{(k)} = \boldsymbol{0}$ is the projection onto $\mathcal{D}(\boldsymbol{x}^{(k)})$.

Proposition 22. *The directions in* (PRP MO II) *have the sufficient decrease property* (3) *with* $\kappa_{\text{sd}} = 1$.

Proof. For $k = 0$ the property is trivially satisfied. Let $k \geq 1$. As $\bar{\boldsymbol{d}}^{(k)}$ is a projection onto $M^{(k)} \subseteq \mathcal{D}(\boldsymbol{x}^{(k)})$,

$$\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k)}) = \varphi(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\left(\boldsymbol{\delta}^{(k)} + \bar{\boldsymbol{d}}^{(k)}\right)) \leq \varphi(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{\delta}^{(k)}) + \underbrace{\varphi(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\bar{\boldsymbol{d}}^{(k)})}_{\leq 0} \leq \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)}).$$

□

Proposition 23. *Suppose Assumptions 1 and 3 hold and that the criticality* $\left\|\boldsymbol{\delta}^{(k)}\right\|$ *is bounded below like in* (12). *Then the Algorithm with modified Armijo-stepsizes according to* (5) *and with directions defined by* (PRP MO II) *generates a critical subsequence.*

Proof. First, note that the projection onto a convex set is non-expansive. If the origin is contained in the convex set $M^{(k)}$, then

$$\|\mathfrak{P}_{M^{(k)}}(\boldsymbol{v})\| \leq \|\boldsymbol{v}\| \qquad \forall \boldsymbol{v} \in \mathbb{R}^N.$$

Let $k \geq 1$. We find that

$$\left\|\boldsymbol{d}^{(k)}\right\| = \left\|\boldsymbol{\delta}^{(k)} + \bar{\boldsymbol{d}}^{(k)}\right\| \leq \left\|\boldsymbol{\delta}^{(k)}\right\| + \left\|\bar{\boldsymbol{d}}^{(k)}\right\| \leq \left\|\boldsymbol{\delta}^{(k)}\right\| + |\beta_{(k)}| \left\|\boldsymbol{d}^{(k-1)}\right\|. \tag{28}$$

Per Assumption 1, the Jacobian of $\boldsymbol{f}$ is Lipschitz. Suppose first that $\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k)}) \geq \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})$. Then

$$\begin{aligned}\left|\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k)}) - \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})\right| &= \mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k)}) - \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)}) \leq \varphi((\boldsymbol{\nabla f}(\boldsymbol{x}^{(k-1)}) - \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}))\boldsymbol{\delta}^{(k)}) \\ &\leq \mathtt{C}_C \mathtt{L}_f \left\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k)}\right\| \left\|\boldsymbol{\delta}^{(k)}\right\| = \mathtt{C}_C \mathtt{L}_f \sigma_{(k-1)} \left\|\boldsymbol{d}^{(k-1)}\right\| \left\|\boldsymbol{\delta}^{(k)}\right\|,\end{aligned}$$

where the existence of $\mathtt{C}_C > 0$ follows from the compactness of C. We find the same bound for the case $\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k)}) < \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})$. Looking at the definition (27), we see that there must be a constant $\mathtt{C}_\beta > 0$ with

$$|\beta_{(k)}| \leq \frac{\mathtt{C}_\beta \sigma_{(k-1)} \left\|\boldsymbol{d}^{(k-1)}\right\| \left\|\boldsymbol{\delta}^{(k)}\right\|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2}.$$

Combining this with (28) results in

$$\begin{aligned}\frac{\left\|\boldsymbol{d}^{(k)}\right\|}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} &\leq \frac{1}{\left\|\boldsymbol{\delta}^{(k)}\right\|} + \frac{\mathtt{C}_\beta \sigma_{(k-1)} \left\|\boldsymbol{d}^{(k-1)}\right\|}{\left\|\boldsymbol{\delta}^{(k)}\right\|} \frac{\left\|\boldsymbol{d}^{(k-1)}\right\|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \\ &\overset{(12)}{\leq} \frac{1}{\varepsilon_{\text{crit}}} + \frac{\mathtt{C}_\beta \sigma_{(k-1)} \left\|\boldsymbol{d}^{(k-1)}\right\|}{\varepsilon_{\text{crit}}} \frac{\left\|\boldsymbol{d}^{(k-1)}\right\|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2}.\end{aligned} \tag{29}$$

Due to Lemma 10, the step-length goes to zero and there is a k_0 such that $\frac{\mathtt{C}_\beta \sigma_{(k-1)} \left\|\boldsymbol{d}^{(k-1)}\right\|}{\varepsilon_{\text{crit}}} < \mathtt{r}$ for some $r \in (0,1)$ and all $k > k_0$. Analogous to what is done in the proof of Proposition 21, we can recurse

$$\frac{\left\|\boldsymbol{d}^{(k)}\right\|}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} \leq \frac{1}{\varepsilon_{\text{crit}}} + r \frac{\left\|\boldsymbol{d}^{(k-1)}\right\|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2}$$

to deduce that $\frac{\left\|\boldsymbol{d}^{(k)}\right\|}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}$ is uniformly bounded above for all $k \in \mathbb{N}_0$, e.g., by $\mathtt{A} > 0$. Then

$$\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \geq \frac{1}{A^2} > 0 \ \forall k \in \mathbb{N}_0 \quad \Rightarrow \quad \sum_{k \in \mathbb{N}_0} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} = \infty,$$

in contradiction to Proposition 11! Existence of a critical sequence follows by Corollary 12. □

Fix broken bibtex entries...

References

[1] Wanyou Cheng. A Two-Term PRP-Based Descent Method. *Numerical Functional Analysis and Optimization*, 28(11-12):1217–1230, December 2007. ISSN 0163-0563, 1532-2467. doi:10.1080/01630560701749524.

[2] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, April 2002. ISSN 1941-0026. doi:10.1109/4235.996017.

[3] Matthias Ehrgott. *Multicriteria Optimization*. Springer, Berlin ; New York, 2nd ed edition, 2005. ISBN 978-3-540-21398-7.

[4] Gabriele Eichfelder. Twenty Years of Continuous Multiobjective Optimization. page 34.

[5] Gabriele Eichfelder. *Adaptive Scalarization Methods in Multiobjective Optimization*. Vector Optimization. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-79157-7 978-3-540-79159-1. doi:10.1007/978-3-540-79159-1.

[6] Y Elboulqe and M El Maghri. An Explicit Three-Term Polak–Ribi'ere–Polyak Conjugate Gradient Method for Bicriteria Optimization.

[7] J. Fliege, A. I. F. Vaz, and L. N. Vicente. Complexity of gradient descent for multiobjective optimization. *Optimization Methods and Software*, 34(5):949–959, September 2019. ISSN 1055-6788, 1029-4937. doi:10.1080/10556788.2018.1510928.

[8] Jörg Fliege and Benar Fux Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical Methods of Operations Research (ZOR)*, 51(3):479–494, August 2000. ISSN 1432-2994, 1432-5217. doi:10.1007/s001860000043.

[9] Ellen H. Fukuda and Luis Mauricio Graña Drummond. A SURVEY ON MULTIOBJECTIVE DESCENT METHODS. *Pesquisa Operacional*, 34(3):585–620, December 2014. ISSN 0101-7438. doi:10.1590/0101-7438.2014.034.03.0585.

[10] M. L. N. Gonçalves and L. F. Prudente. On the extension of the Hager–Zhang conjugate gradient method for vector optimization. *Computational Optimization and Applications*, 76(3):889–916, July 2020. ISSN 0926-6003, 1573-2894. doi:10.1007/s10589-019-00146-1.

[11] M.L.N. Gonçalves, F.S. Lima, and L.F. Prudente. A study of Liu-Storey conjugate gradient methods for vector optimization. *Applied Mathematics and Computation*, 425:127099, July 2022. ISSN 00963003. doi:10.1016/j.amc.2022.127099.

[12] L.M. Graña Drummond and B.F. Svaiter. A steepest descent method for vector optimization. *Journal of Computational and Applied Mathematics*, 175(2):395–414, March 2005. ISSN 03770427. doi:10.1016/j.cam.2004.06.018.

[13] Claus Hillermeier. *Nonlinear Multiobjective Optimization: A Generalized Homotopy Approach.* Springer Basel AG, Basel, 2001. ISBN 978-3-0348-8280-4.

[14] L. R. Lucambio Pérez and L. F. Prudente. Nonlinear Conjugate Gradient Methods for Vector Optimization. *SIAM Journal on Optimization*, 28(3):2690–2720, January 2018. ISSN 1052-6234, 1095-7189. doi:10.1137/17M1126588.

[15] L. R. Lucambio Pérez and L. F. Prudente. A Wolfe Line Search Algorithm for Vector Optimization. *ACM Transactions on Mathematical Software*, 45(4):1–23, December 2019. ISSN 0098-3500, 1557-7295. doi:10.1145/3342104.

[16] Adanay Martín and Oliver Schütze. Pareto Tracer: A predictor–corrector method for multi-objective optimization problems. *Engineering Optimization*, 50(3):516–536, March 2018. ISSN 0305-215X, 1029-0273. doi:10.1080/0305215X.2017.1327579.

[17] Kaisa Miettinen. *Nonlinear Multiobjective Optimization.* Springer Verlag, 2013. ISBN 978-1-4613-7544-9.

[18] H. Mukai. Algorithms for multicriterion optimization. *IEEE Transactions on Automatic Control*, 25(2):177–186, April 1980. ISSN 1558-2523. doi:10.1109/TAC.1980.1102298.

[19] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization.* Springer Series in Operations Research. Springer, New York, 2nd ed edition, 2006. ISBN 978-0-387-30303-1.

[20] Sebastian Peitz and Michael Dellnitz. A Survey of Recent Trends in Multiobjective Optimal Control—Surrogate Models, Feedback Control and Objective Reduction. *Mathematical and Computational Applications*, 23(2):30, June 2018. ISSN 2297-8747. doi:10.3390/mca23020030.

[21] Konstantin Sonntag and Sebastian Peitz. Fast Multiobjective Gradient Methods with Nesterov Acceleration via Inertial Gradient-like Systems, July 2022.

[22] Benar F. Svaiter. The multiobjective steepest descent direction is not Lipschitz continuous, but is Hölder continuous. *Operations Research Letters*, 46(4):430–433, July 2018. ISSN 01676377. doi:10.1016/j.orl.2018.05.008.

[23] Hiroki Tanabe, Ellen H. Fukuda, and Nobuo Yamashita. An accelerated proximal gradient method for multiobjective optimization. *Computational Optimization and Applications*, June 2023. ISSN 0926-6003, 1573-2894. doi:10.1007/s10589-023-00497-w.

[24] Li Zhang, Weijun Zhou, and Dong-Hui Li. A descent modified Polak–Ribière–Polyak conjugate gradient method and its global convergence. *IMA Journal of Numerical Analysis*, 26(4):629–640, October 2006. ISSN 1464-3642, 0272-4979. doi:10.1093/imanum/drl016.

[25] Li Zhang, Weijun Zhou, and Donghui Li. Global convergence of a modified Fletcher–Reeves conjugate gradient method with Armijo-type line search. *Numerische Mathematik*, 104(4):561–572, September 2006. ISSN 0029-599X, 0945-3245. doi:10.1007/s00211-006-0028-z.