NEW NONLINEAR CONJUGATE GRADIENT METHODS WITH GUARANTEED DESCENT FOR MULTI-OBJECTIVE OPTIMIZATION

A PREPRINT

Manuel Berkemeier.
Department of Computer Science
TU Dortmund University, Germany
manuel.berkemeier@tu-dortmund.de

Konstantin Sonntag.
Department of Mathematics
Paderborn University, Germany

Sebastian Peitz.
Department of Computer Science
TU Dortmund University, Germany

December 20, 2024

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 first-order critical points. 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.

Acknowledgments Parts of this research have been funded by the DFG Priority Program 2353 "Daring More Intelligence" (project ID 501834605).

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 a problem 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

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

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. [10]. To obtain multiple solutions, oftentimes so-called *evolutionary algorithms* are used, NSGA-II [7] being a prominent example. If derivatives are available, local descent-based optimization allows for finding individual critical solutions [40, 18, 21], or continuing along the manifold of solutions towards more favorable points [27, 35]. Preferences can also be encoded in scalarizations, see [11, 36]. For more complete overview, consider the surveys [42] and [12].

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 [17]. For convex objectives, it can be shown that faster first-order methods exist [51, 49]. 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 [41]. 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 [41].

Recently, Lucambio Pérez and Prudente [33] have adapted many of the popular nonlinear CG methods to the multi-objective setting. Their directions [22, 33] rely on strong Wolfe conditions being fulfilled. To this end, a suitable step-size algorithm is provided [34]. 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, and requires repeated gradient evaluations. Likewise, the method in [25] needs Wolfe conditions too.

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 certain single-objective methods with the same property. We show convergence of a subsequence of iterates to a critical point if a backtracking step-size rule is used, which is similar to the standard Armijo rule except for the fact that the acceptance threshold shrinks quadratically. There already are similar schemes for bi-objective optimization [14, 13]. The bi-objective algorithms also use the modified Armijo step-size. The directions in [3] are suited for more than two objectives and have the sufficient decrease property too, but the provided backtracking method is merely theoretical as it requires knowledge of the Lipschitz constant of the Jacobian. There are three algorithms in [23] based on coefficients similar to what we have in equation (12). One of the algorithms works with a backtracking procedure employing iterative estimates of the Lipschitz constant of the Jacobian. Sufficient decrease is ensured by having the decrease property as an additional (fulfillable) step-size criterion besides the Armijo condition. In [4], the authors present several directions with the sufficient decrease property. Their algorithm performs no backtracking, but performs quasi-Newton approximations of the objective Hessians to solve a convex program for the stepsize. Although in theory knowledge of Lipschitz constants is required, their experimental results are quite promising. The directions in [1] provide sufficient decrease, but appears to be no convergence proof for the backtracking algorithm.

2 Criticality, Steepest Descent and Sufficient Decrease

Given smooth objective functions, there is a 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

$$\boldsymbol{f}(\boldsymbol{x} + \sigma\boldsymbol{v}) \preceq_{\mathcal{K}} \boldsymbol{f}(\boldsymbol{x}) \qquad \forall \sigma \in (0, \bar{\sigma}) \quad \text{(see [24])}.$$

We will proceed to introduce the maps $\varphi(\bullet)$ and $\mathbb{D}(\bullet, \bullet)$ to facilitate working with the definitions of Pareto-optimality and -criticality. Adopting the notation from [24, 33], 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 $-\operatorname{int}(\mathcal{K})$ in terms of sublevel sets:

$$\boldsymbol{y} \in -\mathcal{K} \Leftrightarrow \varphi(\boldsymbol{y}) \leq 0 \quad \text{and} \quad \boldsymbol{y} \in -\operatorname{int}(\mathcal{K}) \Leftrightarrow \varphi(\boldsymbol{y}) < 0. \tag{2}$$

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

Lemma 3 (Lemma 3.1 in [24])**.** *Let $\boldsymbol{y}, \boldsymbol{y'} \in \mathbb{R}^K$. Then*

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

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

$$\mathbb{D}(\boldsymbol{x}, \boldsymbol{d}) = \mathbb{D}[\boldsymbol{x}](\boldsymbol{d}) = \varphi\left(\boldsymbol{\nabla f}(\boldsymbol{x}) \cdot \boldsymbol{d}\right) = \max_{\boldsymbol{w} \in C} \langle \boldsymbol{\nabla f}(\boldsymbol{x}) \cdot \boldsymbol{d}, \boldsymbol{w} \rangle,$$

then we can infer criticality from the function $\mathbb{D}$ as follows:

Lemma 4 ([24, Lemma 3.3], [50, Thm. 3.1])**.** *Suppose $\boldsymbol{f}$ is continuously differentiable on an open set $\Omega \subseteq \mathbb{R}^N$. For $\boldsymbol{x} \in \Omega$, consider the following optimization problem:*

$$\min_{\boldsymbol{d} \in \mathbb{R}^N} \mathbb{D}[\boldsymbol{x}](\boldsymbol{d}) + \frac{1}{2} \|\boldsymbol{d}\|_2^2. \tag{P}$$

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 $\mathbb{D}[\boldsymbol{x}](\boldsymbol{\delta}) < -\frac{1}{2} \|\boldsymbol{\delta}\|^2 < 0$, 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, if $W \subseteq \Omega$ is compact and $\boldsymbol{f}$ has Lipschitz-continuous gradients on W, then

1. *the steepest descent direction $\boldsymbol{x} \mapsto \boldsymbol{\delta}(\boldsymbol{x})$ is Hölder-continuous on W with exponent $1/2$,*
2. *and its norm $\boldsymbol{x} \mapsto \|\boldsymbol{\delta}(\boldsymbol{x})\|$ is Lipschitz on W.*

Besides, in [24] it is shown that

$$\alpha(\boldsymbol{x}) = -\frac{1}{2} \|\boldsymbol{\delta}(\boldsymbol{x})\|^2, \quad \text{and thus} \quad \mathbb{D}[\boldsymbol{x}](\boldsymbol{\delta}) = -\|\boldsymbol{\delta}(\boldsymbol{x})\|^2.$$

As polynomials are Lipschitz-continuous on compact sets, and the composition of Lipschitz-continuous functions is Lipschitz, the optimal value $\boldsymbol{x} \mapsto \alpha(\boldsymbol{x})$ is Lipschitz if $\boldsymbol{f}$ has Lipschitz-continuous gradients on a compact set.
In the single-objective case, with $\mathcal{K} = \mathbb{R}_{\geq 0}$, the solution is $\boldsymbol{\delta} = -\nabla f(\boldsymbol{x})$. In the multi-objective case, dualization shows the solution to be the element of smallest norm in the convex hull of negative gradients:

$$\boldsymbol{\delta}(\boldsymbol{x}) = -\boldsymbol{\nabla f}(\boldsymbol{x})^{\mathsf{T}} \cdot \boldsymbol{v}_{\boldsymbol{\delta}}, \quad \boldsymbol{v}_{\boldsymbol{\delta}} = \boldsymbol{v}_{\boldsymbol{\delta}}(\boldsymbol{x}) = \operatorname*{arg\,min}_{\boldsymbol{v} \in \operatorname{conv} C} \left\| \boldsymbol{\nabla f}(\boldsymbol{x})^{\mathsf{T}} \cdot \boldsymbol{v} \right\|_2^2. \tag{D}$$

The problem in (P) (or (D)) 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 (P), which is why we assume $\|\boldsymbol{y}\| = 1$ for all $\boldsymbol{y} \in C$ for the remainder of this work.

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 $\kappa_{\mathrm{sd}} > 0$ such that $\left\langle -\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}), \boldsymbol{d}^{(k)} \right\rangle \geq \kappa_{\mathrm{sd}} \left\| \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}) \right\|^2$, we qualify them accordingly in the multi-objective case:

Definition 5. The directions $\left\{\boldsymbol{d}^{(k)}\right\}$ are said to have the *sufficient decrease* property iff for all $k \in \mathbb{N}_0$ the following inequality holds:

$$-\mathbb{D}_k\left(\boldsymbol{d}^{(k)}\right) = -\varphi\left(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\cdot\boldsymbol{d}^{(k)}\right) \geq -\kappa_{\mathrm{sd}}\varphi\left(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\cdot\boldsymbol{\delta}^{(k)}\right) = -\kappa_{\mathrm{sd}}\mathbb{D}_k\left(\boldsymbol{\delta}^{(k)}\right) = \kappa_{\mathrm{sd}}\left\|\boldsymbol{\delta}^{(k)}\right\|^2. \quad \text{(dec)}$$

Instead of the dot product of a single gradient with the direction, we have used the following shorthand notation:

$$\mathbb{D}_k\left(\boldsymbol{d}\right) = \mathbb{D}\left[\boldsymbol{x}^{(k)}\right](\boldsymbol{d}) = \varphi(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\cdot\boldsymbol{d}).$$

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

Remark 6. *If a direction $\boldsymbol{d}^{(k)}$ has the sufficient decrease property $-\mathbb{D}_k(\boldsymbol{d}^{(k)}) \geq \kappa_{\mathrm{sd}}\left\|\boldsymbol{\delta}^{(k)}\right\|^2$ (and if $\boldsymbol{\delta}^{(k)} \neq \mathbf{0}$), then it is in the half-space containing $\boldsymbol{\delta}^{(k)}$, defined by the hyperplane orthogonal to $\boldsymbol{\delta}^{(k)}$. We see this by writing $\boldsymbol{\delta}^{(k)} = -\left(\boldsymbol{\nabla f}^{(k)}\right)^{\mathsf{T}}\boldsymbol{v}_{\boldsymbol{\delta}}$, with $\boldsymbol{v}_{\boldsymbol{\delta}} = \sum_i \lambda_i \boldsymbol{w}_i \in \operatorname{conv}(C)$ according to* (D)*. Then*

$$\left\langle \boldsymbol{d}^{(k)}, \boldsymbol{\delta}^{(k)}\right\rangle = \sum_i -\lambda_i \left\langle \boldsymbol{d}^{(k)}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{w}_i\right\rangle,$$

and each term is non-negative because of $\lambda_i \geq 0$ and

$$-\left\langle \boldsymbol{d}^{(k)}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{w}_i\right\rangle \geq -\max_{\boldsymbol{w}}\left\langle \boldsymbol{d}^{(k)}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{w}\right\rangle = -\mathbb{D}_k(\boldsymbol{d}^{(k)}) \geq \kappa_{\mathrm{sd}}\left\|\boldsymbol{\delta}^{(k)}\right\|^2.$$

3 Algorithm and Step-Size

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

3.1 Modified Armijo Step-Size

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

$$0 < \mathtt{c}_{\boldsymbol{e}} \leq \langle \boldsymbol{w}, \boldsymbol{e}\rangle \leq 1 \qquad \forall \boldsymbol{w} \in C \tag{3}$$

for some constant $\mathtt{c}_{\boldsymbol{e}} > 0$. We can find a suitable vector $\boldsymbol{e}$ 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 $\boldsymbol{e} = [1, \ldots, 1]^{\mathsf{T}}$.

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 [33], we get the strict modified Armijo condition:

Definition 7. 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

$$\boldsymbol{f}(\boldsymbol{x} + \sigma\boldsymbol{d}) - \boldsymbol{f}(\boldsymbol{x}) \preceq_{\mathcal{K}} -\mathtt{a}\sigma^2\left\|\boldsymbol{d}\right\|^2\boldsymbol{e}. \tag{4}$$

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

Proposition 8. *Suppose the conditions of Definition 7 hold. Then there is a suitable step-size $\sigma > 0$ satisfying* (4)*.*

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, with (2), there is some $\boldsymbol{w} \in C$ such that for all $\sigma > 0$:

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

A first order Taylor expansion leads to

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

Rearranging and dividing by $\sigma > 0$ gives

$$\begin{aligned}\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x})\boldsymbol{d}\rangle &> -\mathtt{a}\sigma \left\|\boldsymbol{d}\right\|^2 \langle \boldsymbol{w}, \boldsymbol{e}\rangle - \left\langle \boldsymbol{w}, \frac{\boldsymbol{R}(\sigma)}{\sigma}\right\rangle \\ &\geq -\mathtt{a}\sigma \left\|\boldsymbol{d}\right\|^2 - \left\langle \boldsymbol{w}, \frac{\boldsymbol{R}(\sigma)}{\sigma}\right\rangle, \end{aligned} \tag{5}$$

where, by definition of the total differential, $\dfrac{\boldsymbol{R}(\sigma)}{\sigma} \to \boldsymbol{0}$, as $\sigma \searrow 0$. Because $\boldsymbol{d}$ is a descent direction, the value on the left-hand side (LHS) in (5) 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 9. Under the same conditions as in Definition 7, the step-size $\sigma > 0$ satisfies the weak modified Armijo condition if

$$\varphi\left(\boldsymbol{f}(\boldsymbol{x} + \sigma\boldsymbol{d})\right) - \varphi\left(\boldsymbol{f}(\boldsymbol{x})\right) \leq -\mathtt{a}\sigma^2 \left\|\boldsymbol{d}\right\|^2 \mathtt{c}_{\boldsymbol{e}}. \tag{6}$$

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 $\preceq_{\mathcal{K}}$. But the sequence $\left\{\varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(k)})\right)\right\}_k$ will be monotonic. Likely, larger steps are taken with the weak condition.

Proposition 10. *The strict modified Armijo condition* (4) *implies the weak condition* (6)*.*

Proof. Suppose the strict Armijo condition (4) is fulfilled. With Lemma 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 definition (1) that

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

□

A step-size satisfying (4) or (6) 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 $\sigma_{k,0}$ an initial step-size bounded below by the constant $\mathtt{M} > 0$. The step-size σ_k can be found as

$$\sigma_k = \max_{j\in\mathbb{N}_0} \mathtt{b}^j \sigma_{k,0} \quad \text{such that (4) (or (6)) holds.} \tag{7}$$

3.2 Generic Algorithm and Zoutendijk Property

We are now in a position to state the complete procedure as Algorithm 1. In the following, we continue to establish

Algorithm 1: Algorithm with Generic Descent Direction and Backtracking

Data: $N \in \mathbb{N}, K \in \mathbb{N}, \boldsymbol{f}\colon \mathbb{R}^N \to \mathbb{R}^K, \boldsymbol{x}^{(0)} \in \mathbb{R}^N, \mathtt{a} \in (0,1), \mathtt{b} \in (0,1), \kappa_{\mathrm{sd}} > 0, \sigma_{k,0} \geq \mathtt{M} > 0$. A step-size condition $(*)$: either (4) or (6).
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 (dec);
  Compute a step-size σ_k satisfying $(*)$ by backtracking like in (7);
  Set $\boldsymbol{x}^{(k+1)} \leftarrow \boldsymbol{x}^{(k)} + \sigma_k \boldsymbol{d}^{(k)}$;
end

results to prove convergence 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.** That is, in iteration k we know that $\boldsymbol{x}_{(\ell)}$ was not critical, i.e. $\left\|\boldsymbol{\delta}_{(\ell)}\right\| > 0$ for all $\ell \leq k$. To state our main results, we introduce a set of additional 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*

$$\mathcal{F} = \begin{cases} \left\{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{f}(\boldsymbol{x}) \preceq_{\mathcal{K}} \boldsymbol{f}(\boldsymbol{x}^{(0)})\right\}, & \text{if (4) is used,} \\ \left\{\boldsymbol{x} \in \mathbb{R}^N : \varphi(\boldsymbol{f}(\boldsymbol{x})) \leq \varphi(\boldsymbol{f}(\boldsymbol{x}^{(0)}))\right\}, & \text{if (6) is used.} \end{cases}$$

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. *If $\{\boldsymbol{x}^{(k)}\} \subseteq \mathcal{F}$ is such that the value sequence $\left\{\varphi(\boldsymbol{f}(\boldsymbol{x}^{(k)}))\right\}$ is non-increasing, then this value sequence is uniformly bounded below by some $y_{\mathrm{lb}} \in \mathbb{R}$.*

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 sequence of continuous function values automatically follows. Thus, the following assumption can replace Assumptions 2 and 3:

Assumption 4. *The sublevel set $\mathcal{F}$ as defined in Assumption 1 is closed and bounded, i.e., compact.*

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

Lemma 11. *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 (6) holds:

$$\mathtt{ac}_{\boldsymbol{e}}\sigma_k^2 \left\|\boldsymbol{d}^{(k)}\right\|^2 \leq \varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(k)})\right) - \varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(k)} + \sigma_k \boldsymbol{d}^{(k)})\right).$$

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

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

Due to Assumption 3, the RHS simplifies:

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

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 derive a bound resembling the Zoutendijk condition often encountered in single-objective optimization. Note that – from now on – we will refrain from explicitly stating that the iterates are generated by Algorithm 1 most of the time.

Lemma 12. *Suppose Assumptions 1 to 3 hold, that the directions have the sufficient decrease property* (dec) *with constant $\kappa_{\mathrm{sd}} > 0$, and that the step-sizes σ_k in Algorithm 1 satisfy Eq.* (4)*. Then the following Zoutendijk-like condition follows:*

$$\sum_{k\in\mathbb{N}_0} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} = \sum_{k\in\mathbb{N}_0} \frac{\left(\mathbb{D}_k(\boldsymbol{\delta}^{(k)})\right)^2}{\left\|\boldsymbol{d}^{(k)}\right\|^2} < \infty. \tag{ZD}$$

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

First, suppose $\sigma_k \neq \sigma_{k,0}$. Due to the backtracking procedure, the strict Armijo condition (4) must be violated for $\sigma_k \mathtt{b}^{-1} > \sigma_k$. (If the weak condition is used, and it is violated for some step-size, then the strict condition cannot hold neither for that step-size.) With (2) 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

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

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

$$\begin{aligned}
-\mathtt{a}\frac{\sigma_k^2}{\mathtt{b}^2}\left\|\boldsymbol{d}^{(k)}\right\|^2 &\leq \frac{\sigma_k}{\mathtt{b}}\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}\left(\boldsymbol{x}^{(k)}+h\frac{\sigma_k}{\mathtt{b}}\boldsymbol{d}^{(k)}\right)\boldsymbol{d}^{(k)}\right\rangle \\
&= \frac{\sigma_k}{\mathtt{b}}\left\langle \boldsymbol{w}, \left(\boldsymbol{\nabla f}\left(\boldsymbol{x}^{(k)}+h\frac{\sigma_k}{\mathtt{b}}\boldsymbol{d}^{(k)}\right)-\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\right)\boldsymbol{d}^{(k)}+\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k)}\right\rangle \\
&\leq \frac{\sigma_k}{\mathtt{b}}\left\langle \boldsymbol{w}, \left(\boldsymbol{\nabla f}\left(\boldsymbol{x}^{(k)}+h\frac{\sigma_k}{\mathtt{b}}\boldsymbol{d}^{(k)}\right)-\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\right)\boldsymbol{d}^{(k)}\right\rangle + \frac{\sigma_k}{\mathtt{b}}\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k)}\right\rangle \\
&\leq \frac{\sigma_k^2}{\mathtt{b}^2}\mathtt{L}_f\left\|\boldsymbol{w}\right\|\left\|\boldsymbol{d}^{(k)}\right\|^2 + \frac{\sigma_k}{\mathtt{b}}\mathbb{D}_k\left(\boldsymbol{d}^{(k)}\right),
\end{aligned}$$

where the first term in the last line is derived using the Cauchy-Schwarz inequality and the Lipschitz continuity of $\boldsymbol{\nabla f}$ (Assumption 2). We rearrange to get

$$-\frac{\mathbb{D}_k\left(\boldsymbol{d}^{(k)}\right)}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \leq \sigma_k \frac{\mathtt{L}_f\left\|\boldsymbol{w}\right\|+1}{\mathtt{b}}.$$

Because C is compact, $\|\boldsymbol{w}\|$ is bounded above. Rearranging our bound, we can thus find a constant $\mathtt{c}>0$ such that

$$\mathtt{c}\frac{-\mathbb{D}_k\left(\boldsymbol{d}^{(k)}\right)}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \leq \sigma_k.$$

The LHS is positive, because $\boldsymbol{d}^{(k)}$ is a descent direction. Furthermore, the algorithm ensures that the sufficient decrease conditions holds. Plugging the last inequality into the weak Armijo condition (6), which must hold for σ_k, results in

$$\varphi(\boldsymbol{f}(\boldsymbol{x}^{(k)})) - \varphi\left(\boldsymbol{f}\left(\boldsymbol{x}^{(k)}+\sigma_k\boldsymbol{d}^{(k)}\right)\right) \geq \mathtt{acc}_{\boldsymbol{e}}\frac{\left(\mathbb{D}_k\left(\boldsymbol{d}^{(k)}\right)\right)^2}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \overset{\text{(dec)}}{\geq} \mathtt{acc}_{\boldsymbol{e}}\kappa_{\mathrm{sd}}\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2}. \tag{8}$$

We now also want to find a bound like (8) for the case $\sigma_k = \sigma_{k,0}$. By definition of $\boldsymbol{\delta}^{(k)}$ as the minimizer of (P), we have

$$\mathbb{D}_k\left(\boldsymbol{\delta}^{(k)}\right) + \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}{2} \leq \mathbb{D}_k\left(\kappa_{\mathrm{sd}}^{-1}\boldsymbol{d}^{(k)}\right) + \frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{2\kappa_{\mathrm{sd}}^2}$$

The sufficient decrease condition (dec) gives

$$\kappa_{\mathrm{sd}}^{-1}\mathbb{D}_k\left(\boldsymbol{d}^{(k)}\right) = \mathbb{D}_k\left(\kappa_{\mathrm{sd}}^{-1}\boldsymbol{d}^{(k)}\right) \leq \mathbb{D}_k(\boldsymbol{\delta}^{(k)}),$$

so it must hold that

$$\left\|\boldsymbol{\delta}^{(k)}\right\|^2 \leq \frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{\kappa_{\mathrm{sd}}^2}.$$

Thus,

$$\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \leq \frac{1}{\kappa_{\mathrm{sd}}^4}\left\|\boldsymbol{d}^{(k)}\right\|^2 \leq \frac{1}{\kappa_{\mathrm{sd}}^4\mathtt{a}\left(\sigma_{k,0}\right)^2}\left(\varphi(\boldsymbol{f}(\boldsymbol{x}^{(k)})) - \varphi\left(\boldsymbol{f}\left(\boldsymbol{x}^{(k)}+\sigma_k\boldsymbol{d}^{(k)}\right)\right)\right)$$

As $\sigma_{k,0} \geq \mathtt{M} > 0$ for all k, we can again find a constant $\bar{\mathtt{c}} > 0$ and a bound similar to (8):

$$\bar{\mathtt{c}}\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \leq \varphi(\boldsymbol{f}(\boldsymbol{x}^{(k)})) - \varphi\left(\boldsymbol{x}^{(k)}+\sigma_k\boldsymbol{d}^{(k)}\right). \tag{9}$$

Like in the proof of Lemma 11, we can deduce convergence of the series in (ZD) from (8) and (9) with Assumption 3, by realizing that the sequence of partial sums is again increasing and bounded. □

The Zoutendijk property is common to many descent algorithms in single-objective and multi-objective optimization. It allows for a convenient way to prove that certain direction schemes converge to critical points by means of contradiction.

4 Directions with Guaranteed Descent

All that is left to do, is to actually provide directions $\{\boldsymbol{d}^{(k)}\}$ that can be used with Algorithm 1. The directions presented in this section are adapted from single-objective schemes, of which there are too many to list in this article. Hence, our list is by no means complete, and there are many more approaches to explore. Namely, we draw inspiration from [5, 55, 56]. We provide various translations of the referenced algorithms to the multi-objective case, that are meant to convergence to critical points, while avoiding excessive computational overhead.
Because the various schemes differ in their requirements, we explicitly state their convergence properties in the respective subsections. The main results are Theorems 16, 19, 25, 32, 35 and 38.

4.1 Projection Polak-Ribière-Polyak Scheme

To ensure sufficient decrease, Cheng [5] simply projects the residual term in a standard two-term Polak-Ribière-Polyak (PRP) scheme onto the orthogonal complement of the gradient, i.e., the null space of $\nabla f(\boldsymbol{x}^{(k)})^\mathsf{T}$. With multiple gradients, there usually is no non-trivial subspace orthogonal to all of them. However, the (convex) cone of non-ascent directions

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

is polar to the gradient cone

$$\boldsymbol{\nabla f}(\boldsymbol{x})^\mathsf{T}\,\mathcal{K}^* = \left\{\boldsymbol{\nabla f}(\boldsymbol{x})^\mathsf{T}\boldsymbol{v} : \boldsymbol{v} \in \mathcal{K}^*\right\}$$

and provides a suitable generalization. Note, that per (2) the properties of φ allow for a characterization of $\mathcal{D}$ via

$$\boldsymbol{d} \in \mathcal{D}(\boldsymbol{x}) \quad \Leftrightarrow \quad \mathbb{D}[\boldsymbol{x}](\boldsymbol{d}) = \varphi(\boldsymbol{\nabla f}(\boldsymbol{x})\cdot\boldsymbol{d}) \leq 0. \tag{10}$$

To further motivate the approach, let us revisit the single-objective definition of $\boldsymbol{d}^{(k)}$. Let $\boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)})$ be the single gradient in iteration k and denote for any vector $\boldsymbol{v} \in \mathbb{R}^N$ its orthogonal complement by $\ker(\boldsymbol{v}) \subseteq \mathbb{R}^N$. From [5] we take

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

with

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

where $\bar{\boldsymbol{d}}$ is the metric projection of $\beta^{(k)}\boldsymbol{d}^{(k-1)}$ onto $\ker(\boldsymbol{g}^{(k)})$, For a single vector $\boldsymbol{g}^{(k)}$, the projection onto its null space is given by the simple formula

$$\mathfrak{P}_{\ker(\boldsymbol{g}^{(k)})}\left(\beta_{(k)}\boldsymbol{d}^{(k-1)}\right) = \left(\boldsymbol{I}_{N\times N} - \frac{\boldsymbol{g}^{(k)}\left(\boldsymbol{g}^{(k)}\right)^\mathsf{T}}{\left\|\boldsymbol{g}^{(k)}\right\|^2}\right)\beta_{(k)}\boldsymbol{d}^{(k-1)}. \tag{11}$$

Moreover, as illustrated in Fig. 1, the sufficient decrease property follows suit:

$$\left\langle \boldsymbol{g}^{(k)}, \boldsymbol{d}^{(k)}\right\rangle = \left\langle \boldsymbol{g}^{(k)}, -\boldsymbol{g}^{(k)}\right\rangle + \underbrace{\left\langle \boldsymbol{g}^{(k)}, \bar{\boldsymbol{d}}^{(k)}\right\rangle}_{=0}.$$

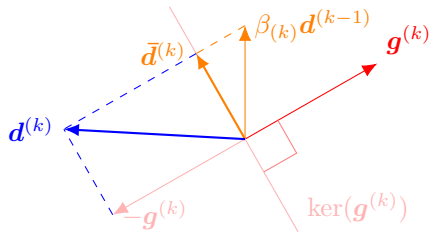

Figure 1: Projection Scheme in the single-objective setting. The standard residual term is projected onto the hyperplane orthogonal to the gradient, so that the final CG direction (blue) has sufficient decrease by construction.

To go to the multi-objective setting, we can use the PRP coefficient from [33]. Moreover, we would like to be able to use the cheap projection formula (11). This can be achieved by not projecting onto $\mathcal{D}(\boldsymbol{x}^{(k)})$, but some non-empty convex subset $S^{(k)}$ of $\mathcal{D}(\boldsymbol{x}^{(k)})$ that contains the origin.

Definition 13. For all $k \in \mathbb{N}_0$, let $S^{(k)}$ be a non-empty convex subset of the cone $\mathcal{D}(\boldsymbol{x}^{(k)})$. We define the direction scheme (PRPP) by

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

with coefficients

$$\beta_{(k)} = \frac{\mathbb{D}\left[\boldsymbol{x}^{(k-1)}\right]\left(\boldsymbol{\delta}^{(k)}\right) - \mathbb{D}\left[\boldsymbol{x}^{(k)}\right]\left(\boldsymbol{\delta}^{(k)}\right)}{-\mathbb{D}\left[\boldsymbol{x}^{(k-1)}\right]\left(\boldsymbol{\delta}^{(k-1)}\right)} = \frac{\mathbb{D}_{k-1}\left(\boldsymbol{\delta}^{(k)}\right) - \mathbb{D}_k\left(\boldsymbol{\delta}^{(k)}\right)}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2}, \quad k \geq 1. \tag{12}$$

As shown in the Appendix, the sufficient decrease property follows from (10):

Lemma 14. *Suppose Assumption 1 holds. The directions in Definition 13 have the sufficient decrease property* (dec) *with* $\kappa_{\mathrm{sd}} = 1$.

Remark 15. *We can always use* $S^{(k)} = \mathcal{D}(\boldsymbol{x}^{(k)})$*, but then the projection might be too expensive. To exploit the single-vector projection formula* (11) *from above, we can use a MiniMax approach, at least if* C *is discrete. For* $\boldsymbol{w} \in C$*, let*

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

and choose $\boldsymbol{w}^*$ *as the minimizer in*

$$\min_{\boldsymbol{w} \in C} \max_{\boldsymbol{v} \in C} \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) \bar{\boldsymbol{d}}(\boldsymbol{w}) \right\rangle = \min_{\boldsymbol{w} \in C} \mathbb{D}_k\left(\bar{\boldsymbol{d}}(\boldsymbol{w})\right).$$

If the optimal value at $\boldsymbol{w}^*$ *is less than or equal to* 0*, then, by the properties of* φ*, the vector* $\bar{\boldsymbol{d}}(\boldsymbol{w}^*)$*, 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* $\bar{\boldsymbol{d}}^{(k)} = \bar{\boldsymbol{d}}(\boldsymbol{w}^*)$ *and (a posteriori) choose* $S^{(k)}$ *to be the ray through* $\bar{\boldsymbol{d}}^{(k)}$*. If the optimal value is positive, then* $\beta_{(k)} \boldsymbol{d}^{(k-1)}$ *belongs to the polar cone of* $\mathcal{D}(\boldsymbol{x}^{(k)})$*. In this case,* $\bar{\boldsymbol{d}}^{(k)} = \boldsymbol{0}$ *is the projection onto* $S^{(k)} = \mathcal{D}(\boldsymbol{x}^{(k)})$*. The procedure is shown in Fig. 2.*

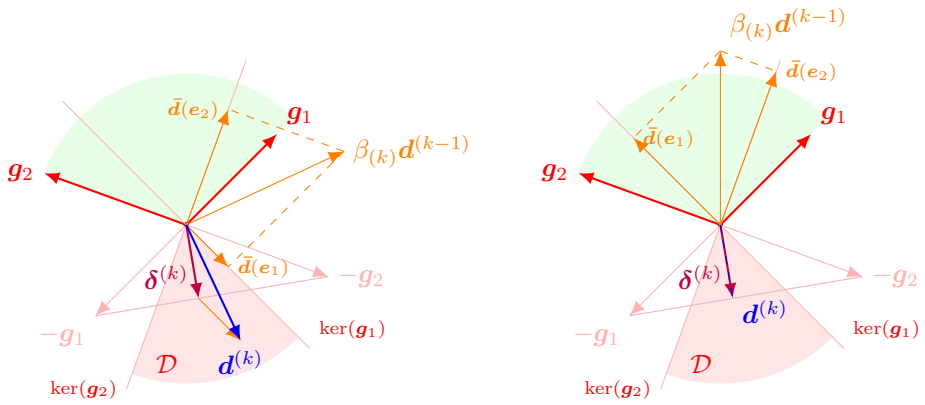

Figure 2: Illustration of the procedure in Remark 15. The gradients of the bi-objective problem in iteration k are denoted by $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$. On the left, the projection $\bar{\boldsymbol{d}}(\boldsymbol{e}_1)$ belongs to the cone $\mathcal{D}$ of non-ascent directions, and $\boldsymbol{d}^{(k)}$ is the sum of steepest descent direction $\boldsymbol{\delta}^{(k)}$ and projection $\bar{\boldsymbol{d}}(\boldsymbol{e}_1)$. On the right, the vector $\beta_{(k)} \boldsymbol{d}^{(k-1)}$ is polar to the cone $\mathcal{D}$ and thus $\boldsymbol{d}^{(k)} = \boldsymbol{\delta}^{(k)}$.

We finally find that the directions (PRPP) lead to convergence. The proof is in the Appendix again.

Theorem 16. *Suppose Assumptions 1 to 3 hold and that the criticality* $\left\|\boldsymbol{\delta}^{(k)}\right\|$ *is bounded below like in* ($\perp$)*. Then Algorithm 1 with directions defined by* (PRPP) *generates a critical subsequence.*

4.2 Three-Term Polak-Ribière-Polyak Scheme

Zhang et al. [55] define the following three-term 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 (P), to obtain a multi-objective scheme. In the following definition, we instead propose parameterized, multi-objective variants:

Definition 17. We define the direction scheme (PRP3) by

$$\boldsymbol{d}^{(k)} = \begin{cases} \boldsymbol{\delta}^{(k)} & \text{if } k = 0 \\ \boldsymbol{\delta}^{(k)} + \alpha_\beta \beta_{(k)}(\boldsymbol{w}_\beta) \boldsymbol{d}^{(k-1)} - \alpha_\theta \theta_{(k)}(\boldsymbol{w}_\theta) \boldsymbol{y}^{(k)}, & \text{if } k \geq 1, \end{cases} \tag{PRP3}$$

with parameter-dependent coefficients

$$\beta_{(k)}(\boldsymbol{w}) = \frac{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) \boldsymbol{y}^{(k)} \right\rangle}{\left\| \boldsymbol{\delta}^{(k-1)} \right\|^2} = \frac{\left\langle \boldsymbol{w}, \boldsymbol{a}^{(k)} \right\rangle}{\left\| \boldsymbol{\delta}^{(k-1)} \right\|^2} \quad \text{and} \quad \theta_{(k)}(\boldsymbol{w}) = \frac{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) \boldsymbol{d}^{(k-1)} \right\rangle}{\left\| \boldsymbol{\delta}^{(k-1)} \right\|^2} = \frac{\left\langle \boldsymbol{w}, \boldsymbol{b}^{(k)} \right\rangle}{\left\| \boldsymbol{\delta}^{(k-1)} \right\|}, \tag{13}$$

where

$$\boldsymbol{y}^{(k)} = \boldsymbol{\delta}^{(k-1)} - \boldsymbol{\delta}^{(k)}, \quad \boldsymbol{a}^{(k)} = \boldsymbol{\nabla f}^{(k)} \boldsymbol{y}^{(k)} \quad \text{and} \quad \boldsymbol{b}^{(k)} = \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k-1)}. \tag{14}$$

The parameter vectors $\boldsymbol{w}_\beta \in C$ and $\boldsymbol{w}_\theta \in C$, and the scalars $\alpha_\beta \in [0, 1]$ and $\alpha_\beta \in [0, 1]$, can be chosen as described below:

1. If C is not discrete, by setting $\alpha_\beta = \alpha_\theta = 1$, and solving the optimization problem

$$\boldsymbol{w}^* = \operatorname*{arg\,min}_{\boldsymbol{w} \in C} \max_{\boldsymbol{v} \in C} \left\langle \boldsymbol{w}, \boldsymbol{a}^{(k)} \right\rangle \left\langle \boldsymbol{v}, \boldsymbol{b}^{(k)} \right\rangle - \left\langle \boldsymbol{w}, \boldsymbol{b}^{(k)} \right\rangle \left\langle \boldsymbol{v}, \boldsymbol{a}^{(k)} \right\rangle \tag{15}$$

   to obtain $\boldsymbol{w}_\beta = \boldsymbol{w}_\theta = \boldsymbol{w}^*$.

2. If C is discrete, by first solving two discrete problems,

$$\psi(\boldsymbol{v}_\theta, \boldsymbol{w}_\theta) = \max_{\boldsymbol{w} \in C} \min_{\boldsymbol{v} \in C} \psi(\boldsymbol{v}, \boldsymbol{w}) \leq \min_{\boldsymbol{w} \in C} \max_{\boldsymbol{v} \in C} \psi(\boldsymbol{w}, \boldsymbol{v}) = \psi(\boldsymbol{w}_\beta, \boldsymbol{v}_\beta), \tag{16}$$

   where ψ is the bilinear form induced by the dyadic product $\boldsymbol{M}^{(k)} = \boldsymbol{a}^{(k)} \cdot \boldsymbol{b}^{(k)\mathsf{T}} \in \mathbb{R}^{K \times K}$,

$$\psi(\boldsymbol{w}, \boldsymbol{v}) = \boldsymbol{w}^\mathsf{T} \boldsymbol{M}^{(k)} \boldsymbol{v} = \left\langle \boldsymbol{w}, \boldsymbol{M}^{(k)} \boldsymbol{v} \right\rangle. \tag{17}$$

   The scalars $\alpha_\beta \in [0, 1], \alpha_\theta \in [0, 1]$ are then chosen according to this decision tree:

   - If $\psi(\boldsymbol{v}_\theta, \boldsymbol{w}_\theta) = \psi(\boldsymbol{w}_\beta, \boldsymbol{v}_\beta)$, then use $\alpha_\beta = \alpha_\theta = 1$.
   - If equality requires a sign switch, i.e., either $\psi(\boldsymbol{v}_\theta, \boldsymbol{w}_\theta) < 0$ and $\psi(\boldsymbol{w}_\beta, \boldsymbol{v}_\beta) \geq 0$ or $\psi(\boldsymbol{v}_\theta, \boldsymbol{w}_\theta) \leq 0$ and $\psi(\boldsymbol{w}_\beta, \boldsymbol{v}_\beta) > 0$, then set $\alpha_\beta = \alpha_\theta = 0$.
   - If both values are different and positive, let $\alpha_\theta = 1$ and shrink the larger value $\psi(\boldsymbol{w}_\beta, \boldsymbol{v}_\beta)$ to obtain

$$\alpha_\beta = \frac{\psi(\boldsymbol{v}_\theta, \boldsymbol{w}_\theta)}{\psi(\boldsymbol{w}_\beta, \boldsymbol{v}_\beta)} \in (0, 1).$$

   - If both values are different and negative, let $\alpha_\beta = 1$ and grow the smaller value $\psi(\boldsymbol{v}_\theta, \boldsymbol{w}_\theta)$ to obtain

$$\alpha_\theta = \frac{\psi(\boldsymbol{w}_\beta, \boldsymbol{v}_\beta)}{\psi(\boldsymbol{v}_\theta, \boldsymbol{w}_\theta)} \in (0, 1).$$

These directions have the sufficient decrease property, which is again shown in the Appendix.

Lemma 18. *Suppose Assumption 1 holds. The directions in Definition 17 have the sufficient decrease property* (dec) *with* $\kappa_{\text{sd}} = 1$.

As seen in the Appendix, and stated in the assumptions of the following result, proving convergence for these specific directions requires a bounded domain.

Theorem 19. *Suppose that Assumptions 1 and 4 hold. The Algorithm with directions as in Definition 17 produces a critical sequence.*

4.3 Fletcher-Reeves Schemes

We next present several multi-objective direction schemes derived from the single-objective method in [56]. There, the single-objective directions $\boldsymbol{d}^{(k)}$ are inspired by the classical Fletcher-Reeves (FR) recipe and given by

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

where $\boldsymbol{g}^{(k)} = \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})$. In contrast to the standard FR method, there is a new factor $\theta_{(k)}$ to ensure the sufficient decrease condition. In fact, the single-objective directions are designed to satisfy

$$\left\langle \boldsymbol{d}^{(k)}, -\boldsymbol{g}^{(k)} \right\rangle = \left\| \boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}) \right\|^2 \qquad \forall k \in \mathbb{N}_0. \tag{18}$$

Unfortunately, simply replacing $-\boldsymbol{g}^{(k)}$ with the multi-objective steepest descent direction $\boldsymbol{\delta}^{(k)}$ from (P) and naively "translating" the coefficients $\theta_{(k)}$ and $\beta_{(k)}$ does not work.

In the next subsections, however, we present multi-objective alternatives derived from (FR SO). These variants are related in that they have the same form

$$\boldsymbol{d}^{(k)} = \begin{cases} \eta_0 \boldsymbol{\delta}^{(k)} & \text{if } k \in \mathcal{N}, \\ \theta_{(k)}\boldsymbol{\delta}^{(k)} + \beta_{(k)}\boldsymbol{d}^{(k-1)} & \text{if } k \in \mathcal{P}, \end{cases} \tag{FR MO1}$$

where $\eta_0 \in \mathbb{R}$ is constant, and $\mathcal{N}, \mathcal{P}$ are disjoint index sets. The directions are designed to have sufficient decrease with constant $\kappa_{\mathrm{sd}} > 0$ and converge to critical points according to the next proposition.

Lemma 20. *Suppose Assumptions 1 to 3 hold, and suppose we apply Algorithm 1 with the directions* $\left\{\boldsymbol{d}^{(k)}\right\}$ *defined by* (FR MO1) *for disjoint index sets* $\mathcal{N}$ *and* $\mathcal{P}$ *with* $\mathbb{N}_0 = \mathcal{N} \cup \mathcal{P}$*, and coefficients* $\left\{\beta_{(k)}\right\}_{k\in\mathcal{P}} \subseteq \mathbb{R}$ *and* $\left\{\theta_{(k)}\right\}_{k\in\mathcal{P}} \subseteq \mathbb{R}$*. Assume further, that the directions* $\left\{\boldsymbol{d}^{(k)}\right\}$ *have the sufficient decrease property with constant* $\kappa_{\mathrm{sd}} > 0$*, and that the criticality is bounded below like in* ($\perp$)*. If there are constants* $\mathtt{C}_{\mathrm{dp}} > 0$ *and* $\mathtt{C}_{\mathrm{o}} \geq 0$*, and an index* $\bar{k} \in \mathbb{N}_0$ *with*

$$\beta_{(k)}^2 \leq \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^4} \qquad \forall k \in \mathcal{P}, k \geq \bar{k}, \tag{19}$$

and

$$\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle \leq \mathtt{C}_{\mathrm{dp}} \left\|\boldsymbol{\delta}^{(k)}\right\|^2 + \mathtt{C}_{\mathrm{o}} \qquad \forall k \in \mathcal{P}, k \geq \bar{k}, \tag{20}$$

then the iteration sequence $\left\{\boldsymbol{x}^{(k)}\right\}$ *is critical.*

The proof is given in the Appendix.

4.3.1 Restarts with Modified Wolfe Condition

Show sufficient decrease becomes straightforward if $\theta_{(k)}$ is not negative. As we will see, a negative coefficient $\theta_{(k)}$ can be avoided with Wolfe-like conditions. Supposing $\boldsymbol{d}^{(k-1)}$ is a descent-direction at $\boldsymbol{x}^{(k-1)}$ (for $k \geq 1$), and $\sigma_{\mathrm{sw}} \in (0, 1)$ is a constant, then $\boldsymbol{d}^{(k-1)}$ satisfies the weak Wolfe condition iff

$$\mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right) \geq \sigma_{\mathrm{sw}} \mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right). \tag{WWC}$$

The condition (WWC) is already sufficient to have non-negative coefficients and to provide the sufficient decrease property. Note, that we do not enforce this condition when determining a step-size in iteration $k-1$. Rather, we use it as a restart criterion in iteration k with the unscaled direction $\boldsymbol{d}^{(k-1)}$.

Unfortunately, we must further modify the coefficients to obtain convergence to a critical point. That is, we employ a stricter test. The direction $\boldsymbol{d}^{(k-1)}$ satisfies the strong Wolfe condition iff

$$\left|\mathbb{D}_k(\boldsymbol{d}^{(k-1)})\right| \leq \sigma_{\mathrm{sw}} \left|\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right)\right|, \tag{SWC}$$

and the modified strong Wolfe condition iff

$$\max\left\{\left|\mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right)\right|, \left|\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k-1)} \right\rangle\right|\right\} \leq \sigma_{\mathrm{sw}} \left|\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right)\right|. \tag{MWC}$$

As $\boldsymbol{d}^{(k-1)}$ is a descent direction, we have $\left|\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right)\right| = -\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right)$, and

$$\text{(MWC)} \Rightarrow \text{(SWC)} \Rightarrow \text{(WWC)}.$$

The condition is used in the direction definition:

Definition 21. We define the direction scheme (FRR) by specifying the following coefficients and index sets for (FR MO1):

$$\begin{aligned}
\eta_0 &= 1, \\
\beta_{(k)} &= \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}{-\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right)} = \frac{-\left\|\boldsymbol{\delta}^{(k)}\right\|^2}{\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right)} = \frac{\mathbb{D}_k\left(\boldsymbol{\delta}^{(k)}\right)}{\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right)}, \\
\theta_{(k)} &= \frac{\mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right) - \mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right)}{-\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right)} = \frac{\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right) - \mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right)}{\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right)}, \\
\mathcal{N} &= \left\{k \in \mathbb{N}_0 : k = 0 \text{ or } \boldsymbol{d}^{(k-1)} \text{ does not satisfy (MWC)}\right\}, \\
\mathcal{P} &= \left\{k \in \mathbb{N}_0 : k \geq 1 \text{ and } \boldsymbol{d}^{(k-1)} \text{ satisfies (MWC)}\right\}.
\end{aligned} \tag{21}$$

Remark. *These directions also generalize a property of their single-objective ancestors: If $\boldsymbol{d}^{(k-1)}$ were to be scaled according to an exact line-search, then $\theta_{(k)} = 1$, and the resulting directions would be similar to the classical FR directions.*

The directions have sufficient decrease as per the next lemma, which is proven in the Appendix.

Lemma 22. *Assume that Assumption 2 holds. The directions in Definition 21 have the sufficient decrease property* (dec) *with $\kappa_{\mathrm{sd}} = 1$.*

We now show that the directions also fit Lemma 20. The proofs for both auxiliary results are in the Appendix.

Lemma 23. *Suppose Assumptions 1 to 3 hold and that the criticality $\left\|\boldsymbol{\delta}^{(k)}\right\|$ is bounded below like in* ($\perp$)*. Assume $\left\{\boldsymbol{d}^{(k)}\right\}_{k\in\mathbb{N}_0}$ are like in Definition 21. Then*

$$\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle \leq 3\left\|\boldsymbol{\delta}^{(k)}\right\|^2 \qquad \forall k \in \mathcal{P}. \tag{22}$$

Lemma 24. *Suppose Assumptions 1 to 3 hold and that the criticality $\left\|\boldsymbol{\delta}^{(k)}\right\|$ is bounded below like in* ($\perp$)*. Assume $\left\{\boldsymbol{d}^{(k)}\right\}_{k\in\mathbb{N}_0}$ are like in Definition 21. Then*

$$\beta_{(k)}^2 \leq \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^4} \qquad \forall k \in \mathcal{P}. \tag{23}$$

The convergence result is derived from Lemma 20 as a corollary by leveraging Lemmas 22 to 24.

Theorem 25. *Suppose Assumptions 1 to 3 hold and that we apply Algorithm 1 with directions as specified in Definition 21. Then the sequence $\left\{\boldsymbol{x}^{(k)}\right\}$ of iterates is critical.*

4.3.2 Denominator with Balancing Offset

When drawing inspiration from the single-objective case, it can be useful to rewrite the coefficient $\theta_{(k)}$. By equation (18), the single-objective coefficients are

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

What is more, (18) is used twice in the convergence analysis: It shows sufficient decrease, $\left\langle \boldsymbol{g}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle \leq -\left\|\boldsymbol{g}^{(k)}\right\|^2$, and later the other inequality, $-\left\|\boldsymbol{g}^{(k)}\right\|^2 \leq \left\langle \boldsymbol{g}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle$, helps in proving criticality. But with multiple objectives, the symmetries break. Incorporating a balancing term in the denominators provides a possible remedy.

Definition 26. Fix constants $\kappa > 0$ and $\mathtt{C}_\gamma \geq 0$. We define the direction scheme (FRBO) by specifying the coefficients and index sets in (FR MO1). We set $\eta_0 = \kappa, \mathcal{N} = \{0\}, \mathcal{P} = \mathbb{N}$, and, for $k \geq 1$,

$$\beta_{(k)} = -\gamma_{(k)}\left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)} \right\rangle, \quad \theta_{(k)} = \kappa + \gamma_{(k)}\mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right), \tag{24}$$

where the factor $\gamma_{(k)}$ is defined by

$$\gamma_{(k)} = \frac{\mathtt{C}_\gamma}{\frac{-\left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)} \right\rangle}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2 + \mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right)\left\|\boldsymbol{\delta}^{(k)}\right\|^2 - \left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)} \right\rangle \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k-1)} \right\rangle}, \tag{25}$$

and $\boldsymbol{w}_\beta \in C$ depends on the sign of $\mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right)$:

$$\boldsymbol{w}_\beta = \begin{cases} \arg\max_{\boldsymbol{w}\in C} \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle & \text{if } \mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right) \geq 0, \\ \arg\min_{\boldsymbol{w}\in C} \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle & \text{if } \mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right) < 0. \end{cases} \tag{26}$$

Before looking in detail at the properties of these coefficients, a few remarks:

- If $\mathtt{C}_\gamma = 0$, then the method reduces to a scaled steepest descent method, i.e., $\boldsymbol{d}^{(k)} = \kappa \boldsymbol{\delta}^{(k)}$ for all $k \in \mathbb{N}_0$.
- If we were to determine the step-size in iteration $k-1$ by *exact* line-search and scale $\boldsymbol{d}^{(k-1)}$ accordingly, then $\mathbb{D}_k(\boldsymbol{d}^{(k-1)}) = 0$, giving $\boldsymbol{d}^{(k)} = \kappa \boldsymbol{\delta}^{(k)} + \beta_{(k)} \boldsymbol{d}^{(k-1)}$, and the method would look similar to the standard FR method.
- The scheme generalizes the single-objective method, which is a special case with singleton set $C = \{1\}$.
- In the defining equations (24, 25, 26), we can replace $\mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right)$ with $\left|\mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right)\right|$ to obtain an alternative scheme. It has similar properties, and the convergence analysis is nearly identical. In the alternative scheme, $\beta_{(k)} = \gamma_{(k)} \left\|\boldsymbol{\delta}^{(k)}\right\|^2$, making implementation a bit simpler. In exchange, it does *not generalize the single-objective case as well*, because of the absolute value.

We can show that $\gamma_{(k)}$ in (25) is non-negative:

Lemma 27. *Suppose that Assumption 1 holds. If $k \geq 1$, $\boldsymbol{w} \in C$ is arbitrary, and $\boldsymbol{w}_\beta$ is as in* (26)*, then*

$$\Delta_{(k)}(\boldsymbol{w}) := \mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right) \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle - \left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \leq 0. \tag{27}$$

Corollary 28. *Suppose that Assumption 1 holds. Then the factor $\gamma_{(k)}$ in* (25) *is bounded below:*

$$0 \leq \gamma_{(k)} \qquad \forall k \in \mathbb{N}. \tag{28}$$

Both results are proven in the Appendix and can be used to show sufficient decrease:

Lemma 29. *Suppose that Assumption 1 holds. The directions in Definition 26 have the sufficient decrease property* (dec) *with $\kappa_{\mathrm{sd}} = \kappa > 0$.*

The proof can also be found in the Appendix. The coefficients are constructed in such a way that – if the criticality is bounded below – the product of $\boldsymbol{\delta}^{(k)}$ and $\boldsymbol{d}^{(k)}$ is bounded above by a multiple of $\left\|\boldsymbol{\delta}^{(k)}\right\|^2$ and a constant offset:

Lemma 30. *Suppose that Assumptions 1 and 4 hold and that the criticality is bounded as in* ($\perp$)*. The directions in Definition 26 fulfill*

$$\left\langle \boldsymbol{d}^{(k)}, \boldsymbol{\delta}^{(k)} \right\rangle \leq \kappa \left\|\boldsymbol{\delta}^{(k)}\right\|^2 + \mathtt{C}_\gamma \qquad \forall k \in \mathbb{N}_0. \tag{29}$$

The proof is in the Appendix. This final lemma will allow us to derive the convergence result easily:

Lemma 31. *Suppose that Assumption 1 holds. The directions in Definition 26 fulfill*

$$\beta_{(k)} \leq \mathtt{C}_\gamma \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \qquad \forall k \in \mathbb{N}_0. \tag{30}$$

Combining Lemmas 29 to 31 with Lemma 20 finalizes the convergence analysis for these coefficients:

Theorem 32. *Suppose that Assumptions 1 and 4 hold and that we apply Algorithm 1 with the directions in Definition 26. Then the sequence $\left\{\boldsymbol{x}^{(k)}\right\}$ of iterates is critical.*

4.3.3 FR Fractional Programming Variants

In this subsection, we introduce two more multi-objective direction schemes inspired by the single-objective directions in (FR SO). They are structurally different from (FR MO1). In both cases, we now use parameterized coefficients to build $\boldsymbol{d}^{(k)}$, and we choose the optimal parameters by minimizing a certain fractional objective. To be precise, both variants look like this:

$$\boldsymbol{d}^{(k)} = \begin{cases} \eta_0 \boldsymbol{\delta}^{(k)} & \text{if } k = 0, \\ \theta_{(k)}(\boldsymbol{w}^*) \boldsymbol{\delta}^{(k)} + \beta_{(k)}(\boldsymbol{w}^*) \boldsymbol{d}^{(k-1)} & \text{if } k \geq 1, \end{cases} \tag{FR MO2}$$

with real coefficients η_0, $\theta_{(k)}$ and $\beta_{(k)}$, and, $\boldsymbol{w}^*$ solving

$$\min_{\boldsymbol{w}\in C} \frac{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}\left(\boldsymbol{x}^{(k)}\right)\boldsymbol{d}^{(k-1)}\right\rangle}{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}\left(\boldsymbol{x}^{(k)}\right)\boldsymbol{\delta}^{(k)}\right\rangle}. \tag{31}$$

If C is discrete, then (31) can be solved with a simple for-loop. If C is a polyhedron, then we have a fractional-linear program, and it can be transformed to an linear program (LP). By noting that $\left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle < 0$ for all $\boldsymbol{v} \in C$ (including $\boldsymbol{w}^*$), we see from the minimizing property

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

that

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

and

$$\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle\left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle \leq \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle \quad \text{for all } \boldsymbol{v} \in C. \tag{32}$$

This inequality is used to show sufficient decrease for suitable coefficients $\beta(\boldsymbol{w}^*)$ and $\theta(\boldsymbol{w}^*)$.

Fractional-Linear Programming Variant I

Definition 33. We derive the scheme (FRF1) from (FR MO2) by fixing a constant $\mathsf{c}_{\text{FR}} > 1$, and defining the coefficients

$$\begin{aligned} \eta_0 &= \mathsf{c}_{\text{FR}}, \\ \theta(\boldsymbol{w}) &= \frac{\mathsf{c}_{\text{FR}}\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k-1)}\boldsymbol{d}^{(k-1)}\right\rangle - \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle}{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k-1)}\boldsymbol{d}^{(k-1)}\right\rangle} \quad \text{and} \\ \beta(\boldsymbol{w}) &= \frac{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle}{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k-1)}\boldsymbol{d}^{(k-1)}\right\rangle}. \end{aligned} \tag{33}$$

The parameters $\boldsymbol{w}^*$ are obtained from (31).

The following sufficient decrease property is derived using (32), as shown in the Appendix.

Lemma 34. *Assume Assumption 1 holds. The directions in Definition 33 have the sufficient decrease property* (dec) *with* $\kappa_{\text{sd}} = \mathsf{c}_{FR} > 1$.

Not only do these directions provide descent, they also lead to convergence:

Theorem 35. *Suppose that Assumptions 1 and 4 hold. The Algorithm with directions (FRF1) as in Definition 33 produces a critical sequence.*

The proof has been moved to the Appendix.

Fractional-Linear Programming Variant II

The coefficients presented next look a bit more complicated than those in (33).

Definition 36. The direction scheme (FRF2) is based on (FR MO2) with coefficients

$$\begin{aligned} \eta_0 &= 1, \\ \theta(\boldsymbol{w}) &:= \frac{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle - \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k-1)}\boldsymbol{d}^{(k-1)}\right\rangle - (\mathsf{c}_{\text{FR}}-1)\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle}{-\mathsf{c}_{\text{FR}}\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle} \quad \text{and} \\ \beta(\boldsymbol{w}) &:= \frac{-\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle}{-\mathsf{c}_{\text{FR}}\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k-1)}\boldsymbol{\delta}^{(k-1)}\right\rangle}, \end{aligned} \tag{34}$$

where, again, $\boldsymbol{w}^*$ is the optimizer in (31) and $\mathsf{c}_{\text{FR}} > 1$ is a constant.

The core difference between both variants (FRF1) and (FRF2) lies in the way that $\boldsymbol{d}^{(k)}$ is bounded. In (33), the terms in $\theta_{(k)}$ have different scaling factors, whilst in (34) an offset term is added.

We have the following results that are both proven in the Appendix.

Lemma 37. *Assume that Assumption 1 holds. The directions in Definition 36 have the sufficient decrease property* (dec) *with* $\kappa_{\text{sd}} = 1$.

Theorem 38. *Suppose that Assumptions 1 and 4 hold. The Algorithm with directions (FRF2) as in Definition 36 produces a critical sequence.*

5 Experiments

5.1 Bi-Objective Rosenbrock Problem

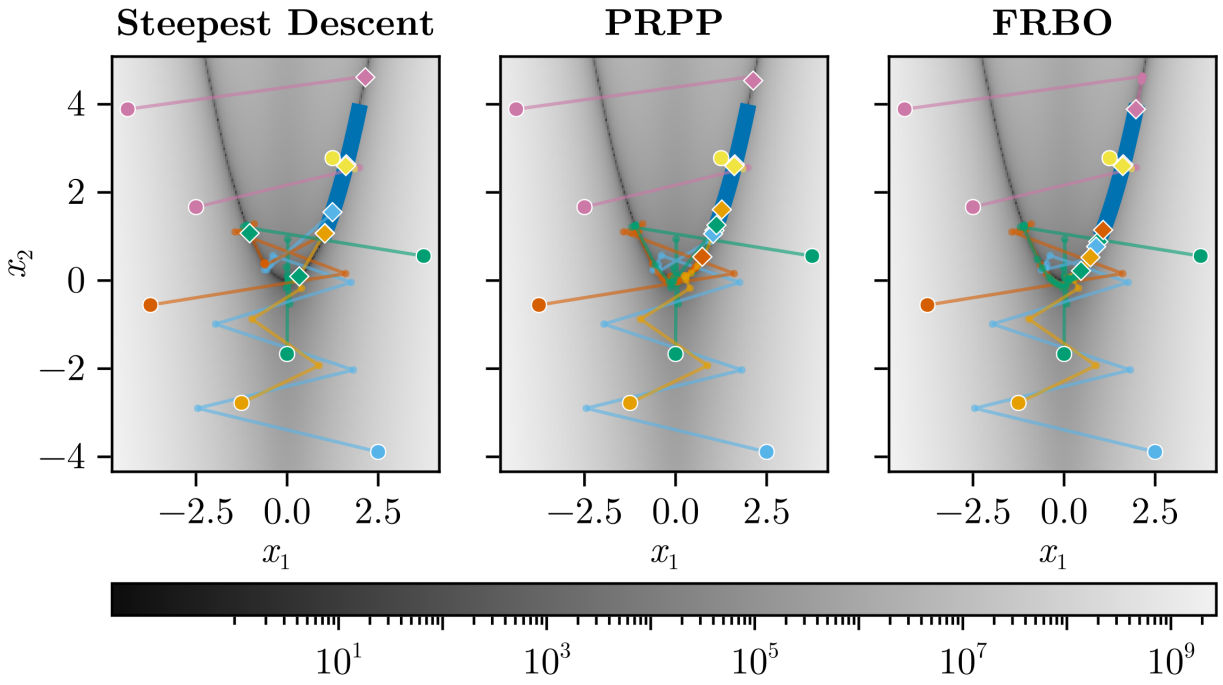

Figure 3: Optimization trajectories using steepest descent directions, and the directions in Definition 13 and Definition 26. The critical set is depicted by a blue line. Circles show starting points and diamonds show end points after at most 30 iterations. In the background, the criticality values of (RB2D) are plotted.

To illustrate the behavior of Algorithm 1 when used with different direction schemes, we start with a bi-objective example. The Rosenbrock function

$$f_{a,b}(x_1, x_2) = b(x_2 - x_1^2)^2 + (a - x_1)^2$$

has its global minimum at (a, a^2) with $f(a, a^2) = 0$. We can construct a bi-objective minimization problem. Fix real numbers $a_1 \leq a_2$ and $b_1 > 0, b_2 > 0$, define $f_1 = f_{a_1,b_1}$ and $f_2 = f_{a_2,b_2}$, and consider

$$\min_{\boldsymbol{x} \in \mathbb{R}^2} \begin{bmatrix} f_1(\boldsymbol{x}) \\ f_2(\boldsymbol{x}) \end{bmatrix}. \tag{RB2D}$$

The Pareto-set is the parabolic segment $\{[x, x^2] : x \in [a_1, a_2]\}$. Like the single-objective problem, the bi-objective problem has a flat valley surrounding the globally optimal set. The valley is easily found, but progress towards the optimal points can be difficult.

For our tests, we use $a_1 = 1, a_2 = 2$ and $b_1 = b_2 = 100$. In Fig. 3 we compare the steepest descent algorithm with strict standard Armijo backtracking against the nonlinear CG directions defined in Definition 13 and Definition 26 with strict modified Armijo backtracking. For each configuration and each of the 8 starting points, at most 30 iterations are performed. The plots show optimization trajectories and a log-scale criticality surface for (RB2D). All algorithms can find the valley, but the nonlinear CG algorithms have more final solutions clustered near the optimal set. They appear to make better progress along the valley.

Name	N	K	lower bounds	upper bounds	Ref.
BK1	2	2	$[-5, \dots, -5]$	$[10, \dots, 10]$	[2, 28, 4]
DD1a	5	2	$[-1, \dots, -1]$	$[1, \dots, 1]$	[6, 16, 4]
DD1b	5	2	$[-10, \dots, -10]$	$[10, \dots, 10]$	[6, 16, 4]
DD1c	5	2	$[-20, \dots, -20]$	$[20, \dots, 20]$	[6, 16, 4]
DGO1	1	2	$[-10, \dots, -10]$	$[13, \dots, 13]$	[9, 28, 4]
Far1	2	2	$[-1, \dots, -1]$	$[1, \dots, 1]$	[15, 28, 4]
FDSa	10	3	$[-2, \dots, -2]$	$[2, \dots, 2]$	[16, 4]
FDSb	200	3	$[-2, \dots, -2]$	$[2, \dots, 2]$	[16, 4]
FDSc	500	3	$[-2, \dots, -2]$	$[2, \dots, 2]$	[16, 4]
FDSd	1000	3	$[-2, \dots, -2]$	$[2, \dots, 2]$	[16, 4]
FF1	2	2	$[-1, \dots, -1]$	$[1, \dots, 1]$	[19, 28, 4]
Hil1	2	2	$[0, \dots, 0]$	$[1, \dots, 1]$	[26, 4]
IKK1	2	3	$[-50, \dots, -50]$	$[50, \dots, 50]$	[29, 28, 4]
JOS1a	50	2	$[-100, \dots, -100]$	$[100, \dots, 100]$	[30, 28, 4]
JOS1b	500	2	$[-100, \dots, -100]$	$[100, \dots, 100]$	[30, 28, 4]
JOS1c	1000	2	$[-100, \dots, -100]$	$[100, \dots, 100]$	[30, 28, 4]
KW2	2	2	$[-3, \dots, -3]$	$[3, \dots, 3]$	[31, 4]
Lov1	2	2	$[-10, \dots, -10]$	$[10, \dots, 10]$	[32, 4]
Lov3	2	2	$[-100, \dots, -100]$	$[100, \dots, 100]$	[32, 4]
Lov4	2	2	$[-20, \dots, -20]$	$[20, \dots, 20]$	[32, 4]
Lov5	3	2	$[-2, \dots, -2]$	$[2, \dots, 2]$	[32, 4]
MGH16	4	5	$[-25, -5, -5, -1]$	$[25, 5, 5, 1]$	[39, 38, 4]
MGH26	4	4	$[-1, \dots, -1]$	$[1, \dots, 1]$	[39, 38, 4]
MMR5a	50	2	$[-5, \dots, -5]$	$[5, \dots, 5]$	[37, 4]
MMR5b	200	2	$[-5, \dots, -5]$	$[5, \dots, 5]$	[37, 4]
MMR5c	500	2	$[-5, \dots, -5]$	$[5, \dots, 5]$	[37, 4]
MOP2	2	2	$[-2, \dots, -2]$	$[2, \dots, 2]$	[20, 28, 54, 4]
MOP3	2	2	$[-\pi, \dots, -\pi]$	$[\pi, \dots, \pi]$	[43, 28, 54, 4]
MOP5	2	3	$[-30, \dots, -30]$	$[30, \dots, 30]$	[43, 28, 54, 4]
PNR	2	2	$[-2, \dots, -2]$	$[2, \dots, 2]$	[44, 4]
SLCDT2	10	2	$[-1, \dots, -1]$	$[1, \dots, 1]$	[46, 4]
SP1	2	2	$[-100, \dots, -100]$	$[100, \dots, 100]$	[47, 28, 4]
SSFYY2	1	2	$[-100, \dots, -100]$	$[100, \dots, 100]$	[48, 28, 4]
TOI9	4	4	$[-1, \dots, -1]$	$[1, \dots, 1]$	[52, 38, 4]
TOI10	4	3	$[-2, \dots, -2]$	$[2, \dots, 2]$	[52, 38, 4]
VU1	2	2	$[-3, \dots, -3]$	$[3, \dots, 3]$	[45, 28, 4]
RB2D	2	2	$[-5, \dots, -5]$	$[5, \dots, 5]$	

Table 1: Test problems, their properties and references. We compiled the suite similar to [4]. Many of the problems are described in the review [28].

5.2 Test Suite

Next, we compare several CG schemes on a multitude of test problems. Most problems are taken from [4] and listed in Table 1. For each test problem, we sample 100 random starting points from within the box described by the lower and upper bound vectors. Then, for each algorithm configuration and each starting point, we perform at most $\max\{1000, 10 * N\}$ iterations. We stop early, if the criticality falls below 10^{-6}, that is, if $\left\|\boldsymbol{\delta}^{(k)}\right\|^2 < 10^{-6}$. Additionally, we stop if a relative step size criterion is met, i.e., if $\left\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right\|_\infty \leq 10^{-10} \left\|\boldsymbol{x}^{(k-1)}\right\|_\infty$. For all runs, we record the number of iterations and the number of function calls. From these records we have computed median values in Table 2 and Table 3. Moreover, Table 4 shows the actual percentage of problems solved with respect to the test $\left\|\boldsymbol{\delta}^{(k)}\right\|^2 < 10^{-6}$.
The column "SD" refers to steepest descent with strict Armijo backtracking. All other algorithms use the strict modified backtracking. The Armijo constants are $\mathtt{a} = 10^{-4}$ and $\mathtt{b} = 0.5$.

- "PRPP" is defined in Definition 13.
- "PRP3" is defined in Definition 17. We used the discrete variant with $C = \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_K\}$.

- "FRR" is defined in Definition 21. We used $\sigma_{\text{sw}} = 0.9$.
- "FRF1a" is defined in Definition 33. We used $\mathtt{c}_{\text{FR}} = 1.1$.
- "FRF1b" is defined in Definition 33. We used $\mathtt{c}_{\text{FR}} = 10$.
- "FRF2a" is defined in Definition 36. We used $\mathtt{c}_{\text{FR}} = 1.1$.
- "FRF2b" is defined in Definition 36. We used $\mathtt{c}_{\text{FR}} = 10$.
- "FRBOa" is defined in Definition 26. We used $\kappa = 1, \mathtt{C}_\gamma = 1$.
- "FRBOb" is defined in Definition 26. We used $\kappa = 1, \mathtt{C}_\gamma = 0.1$.

What can be seen from Table 2 is that steepest descent almost always is worst when it comes to the number of iterations. Surprisingly, this is not necessarily true for the number of function calls, as becomes apparent in Table 3. Here, "FRF1a", "FRF2a" and "FRBOa" seem to do particularly bad. This could be due to boundedness assumptions failing in practice. In any case, the direction norm remains large, prohibiting successful backtracking. From looking at "FRF1b", however, we see that hyperparameter tuning might be necessary. This scheme does very well with respect to the number of iterations and function calls. And Table 4 shows that it is still able to solve most problems. Nearly all schemes fail on the high dimensional test problems. Besides "FRF1b", only "PRPP" manages to solve those. "PRPP" also seems like a good all-rounder, and it does not require tuning hyperparameters.

The Julia code for all experiments can be found at `https://github.com/manuelbb-upb/NonlinearCGCode`.

6 Conclusion

With Algorithm 1 we have provided a very generic algorithm framework to approximate critical points of unconstrained problems like (MOP). Further, in Section 4 we listed several possible CG direction schemes with guaranteed descent. For all of those schemes convergence proofs are given. Our experiments have shown, that many of the CG schemes perform better than simple steepest descent. However, some variants appear to require further hyperparameter tweaking.

Besides improving the directions introduced in this work, one could try to translate other single-objective algorithms to the multi-objective settings. Moreover, the backtracking could be improved with initial step-size guesses leveraging second order approximations (similar to the techniques in [4]). It would also be worthwhile to compare the "simple" backtracking algorithms against methods such as those in [33], which use a more involved step-size technique.

Number of Iterations, mode=all, tol=1e-06											
Problem	max	SD	PRP3	PRPP	FRR	FRF1a	FRF1b	FRF2a	FRF2b	FRBOa	FRBOb
BK1	1000	27	16.5	15	25	19	4	20.5	27	16	25.5
DD1a	1000	8	8	6	10	30.5	6	29.5	7	109.5	7
DD1b	1000	96	91	34.5	96	54	7.5	59	95.5	181.5	95.5
DD1c	1000	182	179.5	60	182	124	15	139	182	112.5	182
DGO1	1000	2	2	2	2	1	1	2	2	2	2
Far1	1000	24.5	19	16	25.5	187	18	192.5	21	1000	21
FDSa	1000	29	24	19	27	101	11	102	28	303	28
FDSb	2000	66	57	43	68.5	310	27	310	64	348.5	63
FDSc	5000	72.5	63	46	74	339	30	338	69	263.5	68
FDSd	10000	79	67.5	49	79	359	33	355.5	74	203	73
FF1	1000	22.5	19	12	20	19.5	5	21.5	22	31.5	22
Hil1	1000	7	6	6	7	34.5	7	19	7	36	7
IKK1	1000	83.5	34.5	35.5	83.5	78	6	86.5	83.5	38	78
JOS1a	1000	819	811	762	819	744	79	819	819	771.5	802.5
JOS1b	5000	2134	2131	1990	2134	1939.5	210	2134	2134	1902	2112.5
JOS1c	10000	4098	4096	3805.5	4098	3725	407	4098	4098	3799.5	4067
KW2	1000	22	10	16	26	133	9.5	140.5	22	150.5	22
Lov1	1000	39	25.5	28.5	39	31.5	4	35.5	39	23	36.5
Lov3	1000	459	459	95	459	414.5	43	456.5	459	439	459
Lov4	1000	81.5	59	63.5	82.5	74	6	82	81.5	50	78
Lov5	1000	656.5	656.5	146	656.5	596.5	64	656.5	656.5	657	656.5
MGH16	1000	58	40	54.5	59	56.5	12	61	58	35	55
MGH26	1000	8	6	6	8	7.5	4	8	8	6	8
MMR5a	1000	1000	1000	621	1000	1000	267.5	1000	1000	1000	1000
MMR5b	2000	2000	2000	1812.5	2000	2000	762	2000	2000	2000	2000
MMR5c	5000	5000	5000	3643.5	5000	5000	1566	5000	5000	5000	5000
MOP2	1000	17	11	15	16.5	29	9.5	38	17	28.5	17
MOP3	1000	17	14	11	15	16	7	15	17	18	17
MOP5	1000	0	0	0	0	0	0	0	0	0	0
PNR	1000	6.5	6	5	6	6	4	6	6	6	6
SLCDT2	1000	18	15	10	14	20.5	8	22	17	28.5	17
SP1	1000	1000	1000	749.5	1000	987	107	1000	1000	365.5	1000
SSFYY2	1000	135	109	135	135	122.5	11	135	135	89	129.5
TOI9	1000	7	7	7	7.5	10	5	10	8	9	7
TOI10	1000	70	20.5	20	28	22	17	25	46	19	51
VU1	1000	116.5	108.5	29.5	116.5	761.5	59	763	106	1000	106
RB2D	1000	89.5	68.5	69	75	71.5	9.5	69.5	89.5	48	89.5

Table 2: Number of iterations of different descent algorithms on test problem suite. Fewer iterations are better, and in each row the cells are shaded proportional to the performance on the respective test problem. The column "max" gives the maximum number of iterations allowed, and its shading is constant.

Number of Func. Calls, mode=all, tol=1e-06										
Problem	SD	PRP3	PRPP	FRR	FRF1a	FRF1b	FRF2a	FRF2b	FRBOa	FRBOb
BK1	29	21.5	17	27	24	11	26	29	22.5	27.5
DD1a	10	10	10	20	176	29	161.5	10	1137.5	10
DD1b	98	93	38	98	111.5	10	112	97.5	1346.5	97.5
DD1c	184	181.5	63.5	184	130	17	144.5	184	298	184
DGO1	4	9.5	4	4	2	5	4	4	4	4
Far1	58.5	57.5	60	83	1803	120	1792	68	14584	68
FDSa	33.5	28.5	31	30.5	644.5	37.5	630.5	30	3127.5	31
FDSb	71	66	76	72	2098.5	87.5	2077.5	67.5	2961	67
FDSc	79.5	76	84	80	2305	99	2267.5	74	1972.5	74
FDSd	89.5	84.5	93	89	2444.5	110	2396	84.5	1386.5	83
FF1	24.5	21	17.5	22	53	16.5	51	24	155	24
Hil1	18.5	19.5	19	25.5	255	49	146	21	310	21.5
IKK1	85.5	38	37.5	85.5	80	8.5	88.5	85.5	40	80
JOS1a	821	813	764	821	746	81	821	821	773.5	804.5
JOS1b	2136	2133	1992	2136	1941.5	212	2136	2136	1904	2114.5
JOS1c	4100	4098	3807.5	4100	3727	409	4100	4100	3801.5	4069
KW2	31	28	50.5	53	322	37	365.5	42	1748	42.5
Lov1	41	29.5	30.5	41	34	7.5	38	41	26	38.5
Lov3	461	461	97.5	461	416.5	45	458.5	461	443.5	461
Lov4	87	71	69.5	85.5	77	11	85.5	85.5	56	81.5
Lov5	658.5	658.5	148.5	658.5	598.5	66	658.5	658.5	659	658.5
MGH16	60	48.5	59	61	58.5	15	64.5	60	37.5	57
MGH26	10.5	11	8	10.5	9.5	10.5	10	10	8	10
MMR5a	1001	1001	862.5	1001	2800.5	355	2472.5	1001	7846	1001
MMR5b	2001	2001	1883.5	2001	2001	764	2001	2001	6739	2001
MMR5c	5001	5001	3669.5	5001	5001	1568	5001	5001	5001	5001
MOP2	19	19	26	24	98.5	33	126	20	89	19.5
MOP3	21.5	20	15	21	34.5	25	31.5	20	44	20
MOP5	1	1	1	1	1	1	1	1	1	1
PNR	10	11	9	11	10.5	24	11	10	12	10
SLCDT2	20	17	13	19	40	32.5	43	19	67	19
SP1	1001	1001	751.5	1001	989	109	1001	1001	370	1001
SSFYY2	137	112	137	137	124.5	12	137	137	91	131.5
TOI9	20	21.5	20.5	22.5	32	30.5	40	19.5	34	19
TOI10	118.5	53	37	56	36	64.5	60	84	37.5	73
VU1	118.5	126	69	122	5556.5	220.5	5485.5	108	11421.5	108
RB2D	91.5	76.5	80	82	73.5	23.5	72	91.5	70	91.5

Table 3: Number of objective function calls of different descent algorithms on test problem suite. Fewer calls are better, and in each row the cells are shaded proportional to the performance on the respective test problem.

Solved Percentage, mode=all, tol=1e-06										
Problem	SD	PRP3	PRPP	FRR	FRF1a	FRF1b	FRF2a	FRF2b	FRBOa	FRBOb
BK1	100	100	100	100	100	100	100	100	100	100
DD1a	100	100	100	100	100	100	100	100	100	100
DD1b	100	100	100	100	100	100	100	100	92	100
DD1c	100	100	100	100	100	100	100	100	81	100
DGO1	100	100	100	100	100	100	100	100	100	100
Far1	100	100	100	100	87	100	86	100	30	100
FDSa	100	100	100	100	100	100	100	100	94	100
FDSb	100	100	100	100	100	100	100	100	100	100
FDSc	100	100	100	100	100	100	100	100	100	100
FDSd	100	100	100	100	100	100	100	100	99	100
FF1	100	100	100	100	99	100	99	100	63	100
Hil1	100	100	100	100	100	100	100	100	75	100
IKK1	80	80	80	80	79	77	80	80	80	80
JOS1a	100	100	100	100	100	100	100	100	100	100
JOS1b	100	100	100	100	100	100	100	100	100	100
JOS1c	100	100	100	100	100	100	100	100	100	100
KW2	100	100	100	100	99	100	100	100	65	100
Lov1	100	100	100	100	100	100	100	100	100	100
Lov3	100	100	100	100	100	100	100	100	100	100
Lov4	97	94	94	97	96	73	96	97	94	97
Lov5	100	100	100	100	100	100	100	100	100	100
MGH16	100	100	100	99	100	100	100	100	100	100
MGH26	100	100	100	100	100	100	100	100	94	100
MMR5a	0	0	100	0	5	100	2	0	1	0
MMR5b	0	0	94	0	0	100	0	0	4	0
MMR5c	0	0	100	0	0	100	0	0	9	0
MOP2	100	100	100	100	100	100	100	100	82	100
MOP3	100	100	100	100	100	100	100	100	97	100
MOP5	100	100	100	100	100	100	100	100	100	100
PNR	100	100	100	100	100	100	100	100	100	100
SLCDT2	100	100	100	100	100	100	100	100	84	100
SP1	43	46	67	43	52	100	43	43	100	47
SSFYY2	100	100	100	100	100	100	100	100	100	100
TOI9	93	90	91	98	96	93	99	91	93	92
TOI10	97	98	100	100	95	99	97	98	81	98
VU1	100	100	100	100	65	100	64	100	3	100
RB2D	96	97	100	100	96	100	97	96	100	96

Table 4: Percentage of solved problems by different descent algorithms on test problem suite. Higher values are better, and in each row the cells are shaded proportional to the performance on the respective test problem.

A Convergence Analysis

A.1 Sufficient Convergence Criteria

As stated before, we show the criticality of an iteration sequence $\left\{\boldsymbol{x}^{(k)}\right\}_{k\in\mathbb{N}_0}$ by means of contradiction: According to Lemma 12, the Zoutendijk condition must hold. If the hypothesis

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

breaks the Zoutendijk property, then it cannot be true, and it must hold that

$$\liminf_{k\to\infty} \left\|\boldsymbol{\delta}^{(k)}\right\| = 0. \tag{$\star$}$$

We want to derive criteria to show ($\star$). The Zoutendijk property implies that

$$\lim_{k\to\infty} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} = 0.$$

It hence breaks, if $\left\|\boldsymbol{\delta}^{(k)}\right\|$ (or an infinite subsequence) is uniformly bounded below and $\left\|\boldsymbol{d}^{(k)}\right\|$ is uniformly bounded above:

Corollary 39. *Assume that Algorithm 1 is applied under the same conditions as in Lemma 12. Suppose further that* ($\perp$) *holds. If there is an index $k_\sharp \in \mathbb{N}_0$ and a constant $\mathtt{C}_{\boldsymbol{d}} > 0$ with*

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \mathtt{C}_{\boldsymbol{d}} \qquad \forall k \geq k_\sharp,$$

then the algorithm has a critical sequence, that is, ($\star$) *holds.*

It might not be obvious that a set of directions $\left\{\boldsymbol{d}^{(k)}\right\}$ is uniformly bounded. We might be able to show that the norm increases no more than a fraction from "iteration to iteration". To this end, we could apply the following Lemma to the sequence $\left\{\left\|\boldsymbol{d}^{(k)}\right\|\right\}$:

Lemma 40. *Suppose $\left\{n^{(k)}\right\}_{k\in\mathbb{N}_0}$ is a sequence of non-negative real numbers. If there is an index $k_\sharp \in \mathbb{N}$, a constant $\mathtt{C} \geq 0$ and a constant $\mathtt{r} \in (0,1)$ with*

$$n^{(k)} \leq \mathtt{C} + \mathtt{r} n^{(k-1)} \qquad \forall k \geq k_\sharp, \tag{35}$$

then there is a constant $\mathtt{C}' \geq 0$ such that

$$n^{(k)} \leq \mathtt{C}' \qquad \forall k \in \mathbb{N}_0.$$

Proof. Let $k \geq k_\sharp$. Repeated application of (35) gives a geometric sum:

$$\begin{aligned} n^{(k)} &\leq \mathtt{C} + \mathtt{r} n^{(k-1)} \\ &\leq \mathtt{C} + \mathtt{r}\left(\mathtt{C} + \mathtt{r}\left\|n^{(k-2)}\right\|\right) \leq \dots \\ &\leq \mathtt{C}\left(1 + \mathtt{r} + \dots + \mathtt{r}^{k-k_\sharp+1}\right) + \mathtt{r}^{k-k_\sharp} n^{(k_\sharp)} \\ &\overset{\mathtt{r}<1}{\leq} \frac{\mathtt{C}}{1-\mathtt{r}} + n^{(k_\sharp)} =: \mathtt{C}^\sharp. \end{aligned}$$

We obtain an overall upper bound (for all $k \in \mathbb{N}_0$) via

$$\mathtt{C}' = \max\left\{\mathtt{C}^\sharp, \max_{k<k_\sharp}\left\{n^{(k)}\right\}\right\} \geq 0.$$

□

Of course, Corollary 39 is rather trivial. It is not necessary to have a fixed uniform upper bound for $\left\|\boldsymbol{d}^{(k)}\right\|$. If instead $\left\|\boldsymbol{d}^{(k)}\right\|^2$ were no larger than a positive multiple of k, then the sum in (ZD) would be bounded below by a divergent harmonic series:

Corollary 41. *Assume that Algorithm 1 is applied under the same conditions as in Lemma 12. Suppose further that* ($\perp$) *holds. If there is an index* $k \in \mathbb{N}_0$ *and constants* $\mathtt{C}' \geq 0, \mathtt{C}^\circ \geq 0$ *such that*

$$\frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} \leq \mathtt{C}' + \mathtt{C}^\circ k, \quad 0 < \mathtt{C}' + \mathtt{C}^\circ k, \qquad \forall k \geq k_\sharp, \tag{36}$$

then the algorithm has a critical sequence, that is, ($\star$) *holds.*

Proof. As indicated, the Zoutendijk condition breaks because the reciprocals are bounded below and

$$\sum_{k\in\mathbb{N}_0} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \geq \sum_{k\geq k_\sharp} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \overset{(36)}{\geq} \sum_{k\geq k_\sharp} \frac{1}{\mathtt{C}' + \mathtt{C}^\circ k} = \infty,$$

where the RHS diverges because it is an infinite sum of a harmonic progression with only finitely many terms removed. □

We can again derive a bound like (36) if the fraction on the LHS does not increase, except for a constant offset:

Lemma 42. *Suppose* $\left\{n^{(k)}\right\}_{k\in\mathbb{N}_0}$ *is a sequence of non-negative real numbers. If there is an index* $k_\sharp \in \mathbb{N}$*, a constant* $\mathtt{C} \geq 0$ *and a constant* $\mathtt{r} \in (0, 1]$ *such that*

$$n^{(k)} \leq \mathtt{C} + \mathtt{r} n^{(k-1)} \qquad \forall k \geq k_\sharp,$$

then there are constants $\mathtt{C}^\circ \geq 0$*,* $\mathtt{C}' \geq 0$ *such that*

$$n^{(k)} \leq \mathtt{C}' + \mathtt{C}^\circ k, \quad \mathtt{C}' + \mathtt{C}^\circ k > 0, \qquad \forall k \geq k_\sharp.$$

Proof. If $\mathtt{r} < 1$, then by Lemma 40 there is a positive upper bound $\mathtt{C}' > 0$ for $\left\{n^{(k)}\right\}_{k\geq k_\sharp}$, and we get the desired result with any $\mathtt{C}^\circ \geq 0$.

Now let $k \geq k_\sharp$ and assume $\mathtt{r} = 1$. We then apply "$n^{(k)} \leq \mathtt{C} + n^{(k-1)}$" recursively to obtain

$$n^{(k)} \leq n^{(k_\sharp)} + \sum_{\ell=k_\sharp}^{k} \mathtt{C} = n^{(k_\sharp)} + \mathtt{C}(k - k_\sharp + 1) = n^{(k_\sharp)} + (1 - k_\sharp)\mathtt{C} + \mathtt{C}k.$$

We can find a positive constant $\mathtt{C}'$ satisfying

$$\mathtt{C}' \geq n^{(k_\sharp)} + (1 - k_\sharp)\mathtt{C},$$

and take $\mathtt{C}^\circ = \mathtt{C} \geq 0$ to get the desired bound. □

In what follows we will use these results to show convergence for explicit direction schemes.

A.2 Various Proofs

A.2.1 Projection Polak-Ribière-Polyak Scheme

Proof of Lemma 14. For $k = 0$ the property is trivially satisfied. Let $k \geq 1$. We use Lemma 3 and the fact that $\bar{\boldsymbol{d}}^{(k)}$ is a projection onto $S^{(k)} \subseteq \mathcal{D}(\boldsymbol{x}^{(k)})$ to get

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

□

Proof of Theorem 16. First, note that the projection onto a convex set is non-expansive. It follows that if the origin is contained in the convex set $S^{(k)}$, then

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

Let $k \geq 1$. Using the triangle-inequality and the non-expansiveness, we find that

$$\left\|\boldsymbol{d}^{(k)}\right\| = \left\|\boldsymbol{\delta}^{(k)} + \bar{\boldsymbol{d}}^{(k)}\right\| \leq \left\|\boldsymbol{\delta}^{(k)}\right\| + \left\|\bar{\boldsymbol{d}}^{(k)}\right\| \leq \left\|\boldsymbol{\delta}^{(k)}\right\| + |\beta_{(k)}| \left\|\boldsymbol{d}^{(k-1)}\right\|. \tag{37}$$

We turn to $\left|\beta_{(k)}\right|$. Suppose first that $\mathbb{D}\left[\boldsymbol{x}^{(k-1)}\right]\left(\boldsymbol{\delta}^{(k)}\right) \geq \mathbb{D}\left[\boldsymbol{x}^{(k)}\right]\left(\boldsymbol{\delta}^{(k)}\right)$. Due to Assumption 2 we know the Jacobian to be Lipschitz with constant $\mathtt{L}_f > 0$. The set C is compact, so there is a constant $\mathtt{C}'$ with $\|\boldsymbol{w}\| \leq \mathtt{C}'$ for all $\boldsymbol{w} \in C$. Thus,

$$\begin{aligned}\left|\mathbb{D}\left[\boldsymbol{x}^{(k-1)}\right]\left(\boldsymbol{\delta}^{(k)}\right) - \mathbb{D}\left[\boldsymbol{x}^{(k)}\right]\left(\boldsymbol{\delta}^{(k)}\right)\right| &= \mathbb{D}\left[\boldsymbol{x}^{(k-1)}\right]\left(\boldsymbol{\delta}^{(k)}\right) - \mathbb{D}\left[\boldsymbol{x}^{(k)}\right]\left(\boldsymbol{\delta}^{(k)}\right) \\ &\overset{\text{Lemma 3}}{\leq} \varphi\left(\left(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k-1)}) - \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\right)\cdot\boldsymbol{\delta}^{(k)}\right) \\ &\leq \left\langle \boldsymbol{w}_\phi, \left(\boldsymbol{\nabla f}(\boldsymbol{x}^{(k-1)}) - \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\right)\cdot\boldsymbol{\delta}^{(k)}\right\rangle \\ &\leq \mathtt{C}'\mathtt{L}_f \left\|\boldsymbol{x}^{(k-1)} - \boldsymbol{x}^{(k)}\right\| \left\|\boldsymbol{\delta}^{(k)}\right\| \\ &= \mathtt{C}'\mathtt{L}_f \left\|\sigma_{k-1}\boldsymbol{d}^{(k-1)}\right\| \left\|\boldsymbol{\delta}^{(k)}\right\|,\end{aligned}$$

where we have used the Cauchy-Schwartz inequality to obtain an upper bound for the inner product. We find the same bound for the case $\mathbb{D}\left[\boldsymbol{x}^{(k-1)}\right]\left(\boldsymbol{\delta}^{(k)}\right) < \mathbb{D}\left[\boldsymbol{x}^{(k)}\right]\left(\boldsymbol{\delta}^{(k)}\right)$. Looking at the definition (12), we see that there must be a constant $\mathtt{C}^\sharp > 0$ with

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

Combining this with (37) results in

$$\begin{aligned}\frac{\left\|\boldsymbol{d}^{(k)}\right\|}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} &\leq \frac{1}{\left\|\boldsymbol{\delta}^{(k)}\right\|} + \frac{\mathtt{C}^\sharp \left\|\sigma_{k-1}\boldsymbol{d}^{(k-1)}\right\|}{\left\|\boldsymbol{\delta}^{(k)}\right\|} \frac{\left\|\boldsymbol{d}^{(k-1)}\right\|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \\ &\overset{(\perp)}{\leq} \frac{1}{\varepsilon_{\text{crit}}} + \frac{\mathtt{C}^\sharp \left\|\sigma_{k-1}\boldsymbol{d}^{(k-1)}\right\|}{\varepsilon_{\text{crit}}} \frac{\left\|\boldsymbol{d}^{(k-1)}\right\|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2}.\end{aligned} \tag{38}$$

Due to Lemma 11, the steps goes to zero and there is a $k_0 \in \mathbb{N}$ and $\mathtt{r} \in (0,1)$ such that

$$\frac{\mathtt{C}^\sharp \left\|\sigma_{k-1}\boldsymbol{d}^{(k-1)}\right\|}{\varepsilon_{\text{crit}}} \leq \mathtt{r} \qquad \forall k \geq k_0.$$

We can thus weaken (38) for $k \geq k_0$ to

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

allowing for Lemma 40, and deduce that $\dfrac{\left\|\boldsymbol{d}^{(k)}\right\|}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}$ is uniformly bounded above, say by $\sqrt{\mathtt{C}'} > 0$. Thus,

$$\frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} \leq \mathtt{C}' \qquad \forall k \in \mathbb{N}_0$$

and the existence of a critical sequence follows with Corollary 41. □

A.2.2 Three-Term Polak-Ribière-Polyak Scheme

Proof of Lemma 18. The case $k = 0$ is trivial. Hence, let $k \geq 1$, and take $\boldsymbol{v} \in C$. The definition (PRP3), together with (17), gives

$$\left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{d}^{(k)}\right\rangle = \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle + \frac{1}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2}\left(\alpha_\beta\psi(\boldsymbol{w}_\beta, \boldsymbol{v}) - \alpha_\theta\psi(\boldsymbol{v}, \boldsymbol{w}_\theta)\right).$$

We are done if we can show

$$\alpha_\beta\psi(\boldsymbol{w}_\beta, \boldsymbol{v}) - \alpha_\theta\psi(\boldsymbol{v}, \boldsymbol{w}_\theta) \leq 0. \tag{39}$$

Consider the first variant. That is, $\alpha_\beta = \alpha_\theta = 1$. Denote the objective of (15) by

$$\Gamma(\boldsymbol{w}, \boldsymbol{v}) := \left\langle \boldsymbol{w}, \boldsymbol{a}^{(k)}\right\rangle\left\langle \boldsymbol{v}, \boldsymbol{b}^{(k)}\right\rangle - \left\langle \boldsymbol{w}, \boldsymbol{b}^{(k)}\right\rangle\left\langle \boldsymbol{v}, \boldsymbol{a}^{(k)}\right\rangle = \psi(\boldsymbol{w}, \boldsymbol{v}) - \psi(\boldsymbol{v}, \boldsymbol{w}). \tag{40}$$

As we take $\boldsymbol{w}_\beta = \boldsymbol{w}_\theta = \boldsymbol{w}^*$, the condition (39) becomes

$$\Gamma(\boldsymbol{w}^*, \boldsymbol{v}) = \psi(\boldsymbol{w}^*, \boldsymbol{v}) - \psi(\boldsymbol{v}, \boldsymbol{w}^*) \leq 0. \tag{41}$$

Now Γ is a skew-symmetric bilinear form,

$$\Gamma(\boldsymbol{w}, \boldsymbol{v}) = -\Gamma(\boldsymbol{v}, \boldsymbol{w}) \qquad \forall (\boldsymbol{v}, \boldsymbol{w}),$$

and the MiniMax solution $\boldsymbol{w}^*$ thus satisfies (41) if a MiniMax theorem holds, i.e., if

$$\min_{\boldsymbol{w}\in C} \max_{\boldsymbol{v}\in C} \Gamma(\boldsymbol{w}, \boldsymbol{v}) = \max_{\boldsymbol{v}\in C} \min_{\boldsymbol{w}\in C} \Gamma(\boldsymbol{w}, \boldsymbol{v}) = 0. \tag{42}$$

The set C is compact and convex. Thus, von Neumann's MiniMax theorem [53] applies, and (42) holds. The problem can be transformed to a LP if C is a polyhedron.

Now consider the discrete case. If there was strong MiniMax theorem for (15) with discrete argument set C, we could just use the same approach as before. A strong, discrete MiniMax theorem would be equivalent to the existence of a pure Nash equilibrium in the symmetric zero-sum two-player game with strategy set C and payoff function Γ, or equivalent to the payoff matrix $[\Gamma(\boldsymbol{w}, \boldsymbol{v})]_{\boldsymbol{w}\in C, \boldsymbol{v}\in C}$ having a saddle point. Unfortunately, it does not generally hold true for $|C| > 2$, because then there are Rock-Paper-Scissors games [8].

Generally, only the following max-min inequality holds:

$$\max_{\boldsymbol{v}\in C} \min_{\boldsymbol{w}\in C} \psi(\boldsymbol{w}, \boldsymbol{v}) \leq \min_{\boldsymbol{w}\in C} \max_{\boldsymbol{v}\in C} \psi(\boldsymbol{w}, \boldsymbol{v}).$$

By renaming variables, this becomes (16). It is easily verified that the non-negative factors α_β and α_θ enforce equality:

$$\alpha_\beta \psi(\boldsymbol{v}_\beta, \boldsymbol{w}_\beta) = \alpha_\theta \psi(\boldsymbol{w}_\theta, \boldsymbol{v}_\theta). \tag{43}$$

Because α_β is non-negative, and $\boldsymbol{v}_\beta$ is a maximizer of $\psi(\boldsymbol{w}_\beta, \bullet)$, we find

$$\alpha_\beta \psi(\boldsymbol{w}_\beta, \boldsymbol{v}) \leq \alpha_\beta \max_{\boldsymbol{v}} \psi(\boldsymbol{w}_\beta, \boldsymbol{v}) = \alpha_\beta \psi(\boldsymbol{w}_\beta, \boldsymbol{v}_\beta).$$

Furthermore, because α_θ is non-negative and $\boldsymbol{v}_\theta$ is a minimizer for $\psi(\bullet, \boldsymbol{w}_\theta)$:

$$\alpha_\theta \psi(\boldsymbol{v}_\theta, \boldsymbol{w}_\theta) \leq \alpha_\theta \psi(\boldsymbol{v}, \boldsymbol{w}_\theta).$$

With (43) we can combine both inequalities:

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

This implies (39). As $\boldsymbol{v}$ was arbitrary we indeed find

$$\mathbb{D}_k(\boldsymbol{d}^{(k)}) = \max_{\boldsymbol{v}} \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k)} \right\rangle \leq \max_{\boldsymbol{v}} \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle = \mathbb{D}_k(\boldsymbol{\delta}^{(k)}).$$

□

Proof of Theorem 19. For a proof by contradiction, assume that the criticality $\left\|\boldsymbol{\delta}^{(k)}\right\|$ is bounded below like in ($\perp$). Because of Assumptions 1 and 4, the norm of the steepest descent is also uniformly bounded *above* by a constant $\mathtt{C}_{\boldsymbol{\delta}} > 0$, i.e.,

$$\left\|\boldsymbol{\delta}^{(k)}\right\| \leq \mathtt{C}_{\boldsymbol{\delta}} \qquad \forall k \in \mathbb{N}_0. \tag{44}$$

If we can show the same for $\boldsymbol{d}^{(k)}$, that concludes the proof because of Corollary 39.

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

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \left\|\boldsymbol{\delta}^{(k)}\right\| + \frac{\left|\alpha_\beta \left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) \boldsymbol{y}^{(k)} \right\rangle\right|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \left\|\boldsymbol{d}^{(k-1)}\right\| + \frac{\left|\alpha_\theta \left\langle \boldsymbol{w}_\theta, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) \boldsymbol{d}^{(k-1)} \right\rangle\right|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \left\|\boldsymbol{y}^{(k)}\right\| \tag{45}$$

As C is compact, there is some constant $\mathtt{C}' > 0$ with $\|\boldsymbol{w}\| \leq \mathtt{C}'$ for all $\boldsymbol{w} \in C$. The Jacobian is continuous and $\boldsymbol{x}^{(k)}$ is from a bounded set, so there is $\mathtt{C}^\circ > 0$ with $\left\|\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\right\| \leq \mathtt{C}^\circ$ for all k. Hence, using Cauchy-Schwartz and $|\alpha_\beta| \leq 1$:

$$\left|\alpha_\beta \left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) \boldsymbol{y}^{(k)} \right\rangle\right| \leq |\alpha_\beta| \, \|\boldsymbol{w}_\beta\| \left\|\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) \boldsymbol{y}^{(k)}\right\| \leq \mathtt{C}' \left\|\boldsymbol{\nabla f}^{(k)}\right\| \left\|\boldsymbol{y}^{(k)}\right\| \overset{(14)}{\leq} \mathtt{C}'\mathtt{C}^\circ \left\|\boldsymbol{\delta}^{(k-1)} - \boldsymbol{\delta}^{(k)}\right\|. \tag{46}$$

Likewise,

$$\left|\alpha_\theta \left\langle \boldsymbol{w}_\theta, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k-1)} \right\rangle\right| \leq \mathtt{C}'\mathtt{C}^\circ \left\|\boldsymbol{d}^{(k-1)}\right\|. \tag{47}$$

Combining (46), (47), as well as (44) and ($\perp$) with (45) results in

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \mathtt{C}_{\boldsymbol{\delta}} + \frac{\mathtt{C}'\mathtt{C}^\circ \left\|\boldsymbol{\delta}^{(k-1)} - \boldsymbol{\delta}^{(k)}\right\|}{\varepsilon_{\text{crit}}^2} \left\|\boldsymbol{d}^{(k-1)}\right\|. \tag{48}$$

Because of our assumptions, the steepest descent direction is $\mathtt{H}$-Hölder-continuous according to Lemma 4. Thus, (48) leads to

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \mathtt{C}_{\boldsymbol{\delta}} + \frac{2\mathtt{C}'\mathtt{C}^\circ\mathtt{H}\sqrt{\left\|\sigma^{(k-1)}\boldsymbol{d}^{(k-1)}\right\|}}{\varepsilon_{\text{crit}}^2} \left\|\boldsymbol{d}^{(k-1)}\right\|.$$

Due to the Armijo condition, Lemma 11 is applicable and $\sqrt{\left\|\sigma^{(k-1)}\boldsymbol{d}^{(k-1)}\right\|}$ vanishes. So there must be $\mathtt{r} \in (0,1)$ and $k_0 \in \mathbb{N}_0$ with

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \mathtt{C}_{\boldsymbol{\delta}} + \mathtt{r}\left\|\boldsymbol{d}^{(k-1)}\right\| \qquad \forall k \geq k_0. \tag{49}$$

We can invoke Lemma 40 (showing $\left\|\boldsymbol{d}^{(k)}\right\|$ to be uniformly bounded) and Corollary 39 to finish the proof. □

A.2.3 Fletcher-Reeves Schemes

Proof of Lemma 20. Whenever $k \in \mathcal{N}$ we have $\boldsymbol{d}^{(k)} = \eta_0\boldsymbol{\delta}^{(k)}$ with $\left\|\boldsymbol{\delta}^{(k)}\right\|^2 \geq \varepsilon_{\text{crit}}^2 > 0$. Thus, if $|\mathcal{N}| = \infty$ then

$$\sum_{k\in\mathcal{N}\cup\mathcal{P}} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} = \underbrace{\sum_{k\in\mathcal{N}} \eta_0^{-2}\left\|\boldsymbol{\delta}^{(k)}\right\|^2}_{=\infty} + \underbrace{\sum_{k\in\mathcal{P}} \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2}}_{\geq 0} = \infty.$$

Next assume $|\mathcal{N}| < \infty$. Also assume $\mathtt{C}_{\text{o}} > 0$. Because of $\varepsilon_{\text{crit}}^2 \leq \left\|\boldsymbol{\delta}^{(k)}\right\|^2$, it follows that $\mathtt{C}_{\text{o}}\frac{\varepsilon_{\text{crit}}^2}{\mathtt{C}_{\text{o}}} \leq \left\|\boldsymbol{\delta}^{(k)}\right\|^2$, or equivalently,

$$\mathtt{C}_{\text{o}} \leq \frac{\mathtt{C}_{\text{o}}}{\varepsilon_{\text{crit}}^2}\left\|\boldsymbol{\delta}^{(k)}\right\|^2.$$

Thus, we deduce from (20) the existence of a constant $\mathtt{C}' > 0$ such that

$$\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle \leq \mathtt{C}'\left\|\boldsymbol{\delta}^{(k)}\right\|^2 \qquad \forall k \in \mathcal{P}, k \geq \bar{k}. \tag{50}$$

Of course, this bound holds with $\mathtt{C}' = \mathtt{C}_{\text{dp}}$ if $\mathtt{C}_{\text{o}} = 0$.

Let k_0 be the maximal element in $\mathcal{N}$. For all $k > k_0$ it holds that $k \in \mathcal{P}$ and $\boldsymbol{d}^{(k)} = \theta_{(k)}\boldsymbol{\delta}^{(k)} + \beta_{(k)}\boldsymbol{d}^{(k-1)}$. Squaring this expression 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

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

Combining both leads to

$$\left\|\boldsymbol{d}^{(k)}\right\|^2 = \beta_{(k)}^2\left\|\boldsymbol{d}^{(k-1)}\right\|^2 - \theta_{(k)}^2\left\|\boldsymbol{\delta}^{(k)}\right\|^2 + 2\theta_{(k)}\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)}\right\rangle. \tag{51}$$

We propose that there is an index $k_1 \geq k_0$ and a constant $\mathtt{C}^\dagger \geq 0$ such that

$$\frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} \leq \beta_{(k)}^2\frac{\left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} + \frac{\mathtt{C}^\dagger}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} \qquad \forall k > k_1. \tag{52}$$

To prove this bound, first consider the case that $\theta_{(k)} \leq 0$ for some $k > k_0$. Then (51) can be weakened to

$$\left\|\boldsymbol{d}^{(k)}\right\|^2 \leq \beta_{(k)}^2\left\|\boldsymbol{d}^{(k-1)}\right\|^2,$$

because $-\theta_{(k)}^2 \left\|\boldsymbol{\delta}^{(k)}\right\|^2 \leq 0$ as well as $2\theta_{(k)} \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle \leq 0$ (see Remark 6). Hence, for non-positive $\theta_{(k)}$, the bound (52) holds for any index $k_1 \geq k_0$ and any real number $\mathtt{C}^\dagger \geq 0$.

Let $k_1 = \max\left\{k_0, \bar{k}\right\}$ and $k > k_1$. Now consider the case $\theta_{(k)} \geq 0$. Substituting our earlier bound for the product (50), into (51) gives

$$\begin{aligned}\left\|\boldsymbol{d}^{(k)}\right\|^2 &\leq \beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2 - \theta_{(k)}^2 \left\|\boldsymbol{\delta}^{(k)}\right\|^2 + 2\theta_{(k)}\mathtt{C}' \left\|\boldsymbol{\delta}^{(k)}\right\|^2 \\ &= \beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2 + \left\|\boldsymbol{\delta}^{(k)}\right\|^2 \left(2\theta_{(k)}\mathtt{C}' - \theta_{(k)}^2\right) \\ &= \beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2 + \left\|\boldsymbol{\delta}^{(k)}\right\|^2 \left(\mathtt{C}'^2 - \left(\mathtt{C}' - \theta_{(k)}\right)^2\right).\end{aligned}$$

We can divide by $\left\|\boldsymbol{\delta}^{(k)}\right\|^4$ and dismiss the non-positive term:

$$\begin{aligned}\frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} &\leq \beta_{(k)}^2 \frac{\left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} + \frac{\mathtt{C}'^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} - \frac{\left(\mathtt{C}' - \theta_{(k)}\right)^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} \\ &\leq \beta_{(k)}^2 \frac{\left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} + \frac{\mathtt{C}'^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2},\end{aligned} \tag{53}$$

All in all, equation (52) is valid with $\mathtt{C}^\dagger = \mathtt{C}'^2 \geq 0$ and $k_1 = \max\left\{k_0, \bar{k}\right\}$.

For all $k > k_1$ the bound (19) holds, too. Plug it into (53) to obtain

$$\frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} \leq \frac{\left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} + \frac{\mathtt{C}^\dagger}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} \overset{(\perp)}{\leq} \frac{\left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^4} + \frac{\mathtt{C}^\dagger}{\varepsilon_{\text{crit}}^2} \quad \forall k > k_1.$$

By Lemma 42 the fraction does not grow too quick so that Corollary 41 can be used to conclude the proof. □

Proof of Lemma 22. For $k = 0$ or $\boldsymbol{d}^{(k-1)}$ violating (MWC) there is nothing to show. We do a proof by induction to show the general case.

Let $k \geq 1$ and suppose $\boldsymbol{d}^{(k-1)}$ satisfies (MWC). As always we can assume $\left\|\boldsymbol{\delta}^{(k-1)}\right\| < 0$ and, by induction, we have $\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right) < 0$. The strong Wolfe conditions thus imply the standard Wolfe conditions (WWC). We obtain

$$\theta_{(k)} \underbrace{\left(-\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right)\right)}_{>0} = \mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right) - \mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right) \geq \underbrace{(\sigma_{\text{sw}} - 1)}_{<0} \underbrace{\mathbb{D}_{k-1}\left(\boldsymbol{d}^{(k-1)}\right)}_{\leq 0} \geq 0.$$

In conclusion, we see that $\theta_{(k)}$ is non-negative.

Now take any $\boldsymbol{w} \in C$ and use the fact that $\theta_{(k)}$ and $\beta_{(k)}$ are non-negative, which allows us to pull them out of the sublinear $\max$ operator in $\mathbb{D}_k$:

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

□

A.3 Restarts with Modified Wolfe Condition

Proof of Lemma 23. Let $k \in \mathcal{P}$. Like before, we plug in the definitions (21):

$$\begin{aligned}
\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle &= \theta_{(k)} \left\| \boldsymbol{\delta}^{(k)} \right\|^2 + \beta_{(k)} \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k-1)} \right\rangle \\
&= \frac{\left\| \boldsymbol{\delta}^{(k)} \right\|^2}{-\mathbb{D}_{k-1} \boldsymbol{d}^{(k-1)}} \left(\mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) - \mathbb{D}_{k-1} \left(\boldsymbol{d}^{(k-1)} \right) + \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k-1)} \right\rangle \right) \\
&\leq \frac{\left\| \boldsymbol{\delta}^{(k)} \right\|^2}{-\mathbb{D}_{k-1} \left(\boldsymbol{d}^{(k-1)} \right)} \left(2 \max \left\{ \left| \mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) \right|, \left| \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k-1)} \right\rangle \right| \right\} - \mathbb{D}_{k-1} \left(\boldsymbol{d}^{(k-1)} \right) \right) \\
&\overset{\text{(MWC)}}{\leq} (1 + 2\sigma_{\text{sw}}) \frac{\left\| \boldsymbol{\delta}^{(k)} \right\|^2}{\mathbb{D}_{k-1} \left(\boldsymbol{d}^{(k-1)} \right)} \mathbb{D}_{k-1} \left(\boldsymbol{d}^{(k-1)} \right).
\end{aligned}$$

Because of $\sigma_{\text{sw}} \in (0, 1)$, we obtain (22). □

Proof of Lemma 24. For these directions, the sufficient decrease property is $0 \leq -\mathbb{D}_{k-1} \left(\boldsymbol{\delta}^{(k-1)} \right) \leq -\mathbb{D}_{k-1} \left(\boldsymbol{d}^{(k-1)} \right)$. Taking squares gives

$$0 \leq \left(\mathbb{D}_{k-1} \left(\boldsymbol{\delta}^{(k-1)} \right) \right)^2 \leq \left(\mathbb{D}_{k-1} \left(\boldsymbol{d}^{(k-1)} \right) \right)^2.$$

For the reciprocals, the relation switches:

$$\left(\beta_{(k)} \right)^2 = \frac{\left\| \boldsymbol{\delta}^{(k)} \right\|^4}{\left(\mathbb{D}_{k-1} (\boldsymbol{d}^{(k-1)}) \right)^2} \leq \frac{\left\| \boldsymbol{\delta}^{(k)} \right\|^4}{\left(\mathbb{D}_{k-1} (\boldsymbol{\delta}^{(k-1)}) \right)^2} = \frac{\left\| \boldsymbol{\delta}^{(k)} \right\|^4}{\left\| \boldsymbol{\delta}^{(k-1)} \right\|^4}.$$

□

A.4 Denominator with Balancing Offset

Proof of Lemma 27. Let $k \geq 1$ and $\boldsymbol{w} \in C$ be arbitrary. If $\mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) \geq 0$, then

$$\mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \leq \mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) \arg\max_{\boldsymbol{w} \in C} \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle = \mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) \left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle.$$

If otherwise $\mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) < 0$, then

$$\mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \leq \mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) \arg\min_{\boldsymbol{w} \in C} \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle = \mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) \left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle.$$

Moreover, $\boldsymbol{\delta}^{(k)}$ is a descent direction with $-\left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \geq 0$, and thus

$$\begin{aligned}
-\left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle &\leq -\left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \arg\max_{\boldsymbol{w} \in C} \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \\
&= -\left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right).
\end{aligned}$$

Combining both inequalities results in

$$\begin{aligned}
\Delta_{(k)}(\boldsymbol{w}) &\overset{(27)}{=} \mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle - \left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \\
&\leq \mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) \left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle - \left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \mathbb{D}_k \left(\boldsymbol{d}^{(k-1)} \right) \qquad = 0.
\end{aligned}$$

□

Proof of Corollary 28. Let $k \geq 1$. Write

$$\gamma_{(k)} = \frac{\mathsf{C}_\gamma}{\frac{-\left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle}{\left\| \boldsymbol{\delta}^{(k)} \right\|^2} \left\| \boldsymbol{\delta}^{(k-1)} \right\|^2 + \Gamma_{(k)}},$$

with

$$\Gamma_{(k)} := \mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right)\left\|\boldsymbol{\delta}^{(k)}\right\|^2 - \left\langle \boldsymbol{w}_\beta, \mathbb{D}_k\boldsymbol{\delta}^{(k)}\right\rangle\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k-1)}\right\rangle. \tag{54}$$

If $\Gamma^{(k)}$ is not negative, then $\gamma_{(k)} \geq 0$, because the enumerator $\mathtt{C}_\gamma$ is not negative, and the other summand in the denominator is not negative neither. According to (D) we can write $\boldsymbol{\delta}^{(k)} = -(\boldsymbol{\nabla f}^{(k)})^\mathsf{T}\boldsymbol{v}_{\boldsymbol{\delta}}$ with $\boldsymbol{v}_{\boldsymbol{\delta}} \in \operatorname{conv}(C)$. Hence,

$$\Gamma_{(k)} = -\mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right)\left\langle \boldsymbol{v}_{\boldsymbol{\delta}}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle + \left\langle \boldsymbol{w}_\beta, \mathbb{D}_k\boldsymbol{\delta}^{(k)}\right\rangle\left\langle \boldsymbol{v}_{\boldsymbol{\delta}}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle. \tag{55}$$

The expression $\Delta_{(k)}(\boldsymbol{w})$ in (27) is linear in its argument. Write $\boldsymbol{v}_{\boldsymbol{\delta}} = \sum_i \lambda_i \boldsymbol{w}_i$ with $\boldsymbol{w}_i \in C$ and convex coefficients $\lambda_i \geq 0$. Then

$$\Gamma_{(k)} = -\Delta_{(k)}(\boldsymbol{v}_{\boldsymbol{\delta}}) = -\Delta_{(k)}\left(\sum_i \lambda_i \boldsymbol{w}_i\right) = \sum_i -\lambda_i \cdot \Delta_{(k)}(\boldsymbol{w}_i) \geq 0,$$

as each term is non-negative according to Lemma 27.

□

Proof of Lemma 29. The case $k = 0$ is trivial. Let $k \geq 1$ and $\boldsymbol{w} \in C$ be arbitrary. Use Lemma 27 and Corollary 28 to derive the sufficient decrease property from (24):

$$\begin{aligned}
\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{d}^{(k)}\right\rangle &= \theta_{(k)}\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle + \beta_{(k)}\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{d}^{(k-1)}\right\rangle \\
&= \kappa\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle + \underbrace{\gamma_{(k)} \cdot \Delta_{(k)}(\boldsymbol{w})}_{\leq 0} \\
&\leq \kappa\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle \\
&\overset{\kappa>0}{\leq} \kappa \max_{\boldsymbol{w}\in C}\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle = \kappa\mathbb{D}_k(\boldsymbol{\delta}^{(k)}).
\end{aligned}$$

As $\boldsymbol{w} \in C$ was arbitrary, we have

$$\mathbb{D}_k(\boldsymbol{d}^{(k)}) = \max_{\boldsymbol{w}\in C}\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k)}\boldsymbol{d}^{(k)}\right\rangle \leq \kappa\mathbb{D}_k(\boldsymbol{\delta}^{(k)}) = -\kappa\left\|\boldsymbol{\delta}^{(k)}\right\|^2.$$

□

Proof of Lemma 30. The case $k = 0$ is trivial because of $\boldsymbol{d}_0 = \kappa\boldsymbol{\delta}_0$ and $\mathtt{C}_\gamma \geq 0$.
Let $k \geq 1$. The definitions (FR MO1) and (24) give

$$\begin{aligned}
\left\langle \boldsymbol{d}^{(k)}, \boldsymbol{\delta}^{(k)}\right\rangle &= \left(\kappa + \gamma_{(k)}\mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right)\right)\left\|\boldsymbol{\delta}^{(k)}\right\|^2 - \gamma_{(k)}\left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle\left\langle \boldsymbol{d}^{(k-1)}, \boldsymbol{\delta}^{(k)}\right\rangle \\
&= \kappa\left\|\boldsymbol{\delta}^{(k)}\right\|^2 + \gamma_{(k)}\left(\mathbb{D}_k\left(\boldsymbol{d}^{(k-1)}\right)\left\|\boldsymbol{\delta}^{(k)}\right\|^2 - \left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle\left\langle \boldsymbol{d}^{(k-1)}, \boldsymbol{\delta}^{(k)}\right\rangle\right) \\
&\overset{(54)}{=} \kappa\left\|\boldsymbol{\delta}^{(k)}\right\|^2 + \gamma_{(k)}\Gamma_{(k)}.
\end{aligned}$$

Recall that $\Gamma^{(k)} \geq 0$, and

$$\gamma_{(k)} = \frac{\mathtt{C}_\gamma}{\frac{-\left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2 + \Gamma_{(k)}} \geq 0.$$

As

$$\bar{\Gamma}^{(k)} = \frac{-\left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2 \geq 0,$$

the inequality follows immediately if $\Gamma^{(k)} > 0$. For $\Gamma^{(k)} = 0$, we use the assumptions to find positive constants bounding $\bar{\Gamma}^{(k)}$ from above and below. So even if $\Gamma^{(k)} = 0$ the value of $\gamma^{(k)}$ is well-defined and finite. In this case, $\gamma^{(k)}\Gamma^{(k)} = 0 \leq \mathtt{C}_\gamma$. □

Proof of Lemma 31. Similar to before we find the bound

$$\gamma_{(k)} = \frac{\mathtt{C}_\gamma}{\frac{-\left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2 + \Gamma_{(k)}} \leq \frac{\mathtt{C}_\gamma}{\frac{-\left\langle \boldsymbol{w}_\beta, \boldsymbol{\nabla f}^{(k)}\boldsymbol{\delta}^{(k)}\right\rangle}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \qquad \forall k \in \mathbb{N}.$$

Using this in the definition (24) directly gives the desired bound (30). □

A.5 Fractional-Linear Programming Variant I

Proof of Lemma 34. We show the even stronger property

$$\left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k)} \right\rangle \leq \mathsf{c}_{\mathsf{FR}} \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \overset{\mathsf{c}_{\mathsf{FR}}>1}{\leq} \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \leq 0 \qquad \forall \boldsymbol{v} \in C, \forall k \geq 0, \tag{56}$$

which implies sufficient decrease.

The case $k = 0$ is trivial. Let $k \geq 1$ and take $\boldsymbol{v} \in C$. Equation (32) shows that

$$\psi(\boldsymbol{w}^*, \boldsymbol{v}) := \left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle - \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \leq 0. \tag{57}$$

The expression actually occurs if we plug the definitions (FR MO2) and (33) into $\left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k)} \right\rangle$:

$$\begin{aligned}
\left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k)} \right\rangle &= \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \left(\theta\left(\boldsymbol{w}^*\right) \boldsymbol{\delta}^{(k)} + \beta\left(\boldsymbol{w}^*\right) \boldsymbol{d}^{(k-1)} \right) \right\rangle \\
&= \theta\left(\boldsymbol{w}^*\right) \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle + \beta\left(\boldsymbol{w}^*\right) \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \\
&= \frac{\mathsf{c}_{\mathsf{FR}} \left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle}{\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle} - \frac{\psi(\boldsymbol{w}^*, \boldsymbol{v})}{\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle}
\end{aligned} \tag{58}$$

With $\boldsymbol{\delta}^{(k-1)}$ a descent direction at $\boldsymbol{x}^{(k-1)}$, the denominator is negative, and (57) provides

$$-\frac{\psi(\boldsymbol{w}^*, \boldsymbol{v})}{\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle} \leq 0.$$

We can thus dismiss this term in (58) to get

$$\left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k)} \right\rangle \leq \mathsf{c}_{\mathsf{FR}} \left\langle \boldsymbol{v}, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle.$$

This concludes the proof, as $\boldsymbol{v} \in C$ was arbitrary.
(If the denominator were $\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle$ instead, we could have made an inductive argument.) □

Proof of Appendix A.5. For a proof by contradiction, assume that the criticality is bounded like in ($\perp$). Let $k \geq 1$. We use the triangle inequality and Cauchy-Schwarz on (FR MO2) to obtain

$$\left\| \boldsymbol{d}^{(k)} \right\| \leq \left| \theta_{(k)} \right| \left\| \boldsymbol{\delta}^{(k)} \right\| + \left| \beta_{(k)} \right| \left\| \boldsymbol{d}^{(k-1)} \right\|. \tag{59}$$

We first investigate $\left| \theta_{(k)} \right| \left\| \boldsymbol{\delta}^{(k)} \right\|$. Use $\mathsf{c}_{\mathsf{FR}} = 1 + (\mathsf{c}_{\mathsf{FR}} - 1)$ in (33):

$$\theta_{(k)} = \frac{\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle}{\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle} + (\mathsf{c}_{\mathsf{FR}} - 1) \frac{\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle}{\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle}.$$

As $(\mathsf{c}_{\mathsf{FR}} - 1) > 0$, the triangle inequality gives

$$\left| \theta_{(k)} \right| \leq \left| \frac{\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle}{\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle} \right| + (\mathsf{c}_{\mathsf{FR}} - 1). \tag{60}$$

Assuming compatible norms, we can use the Cauchy-Schwartz inequality for the enumerator and then invoke the $\mathtt{L}_f$-Lipschitz-property of the Jacobian. Additionally, we note that C is compact, so there is a positive constant with

$$\|\boldsymbol{w}\| \leq \mathtt{C}' \qquad \forall \boldsymbol{w} \in C. \tag{61}$$

Thus,

$$\begin{aligned}
\left| \left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k)} \boldsymbol{d}^{(k-1)} \right\rangle \right| &= \left| \left\langle \boldsymbol{w}^*, \left(\boldsymbol{\nabla f}^{(k-1)} - \boldsymbol{\nabla f}^{(k)} \right) \boldsymbol{d}^{(k-1)} \right\rangle \right| \\
&\leq \|\boldsymbol{w}^*\| \left\| \left(\boldsymbol{\nabla f}^{(k-1)} - \boldsymbol{\nabla f}^{(k)} \right) \boldsymbol{d}^{(k-1)} \right\| \\
&\leq \mathtt{C}' \left\| \boldsymbol{\nabla f}^{(k-1)} - \boldsymbol{\nabla f}^{(k)} \right\| \left\| \boldsymbol{d}^{(k-1)} \right\| \\
&\leq \mathtt{C}' \mathtt{L}_f \left\| \sigma_{k-1} \boldsymbol{d}^{(k-1)} \right\| \left\| \boldsymbol{d}^{(k-1)} \right\|.
\end{aligned} \tag{62}$$

The sufficient decrease property (56), together with ($\perp$) provides a constant bound for the denominator:

$$\frac{1}{\left|\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle\right|} \leq \frac{1}{\mathtt{c}_{\mathrm{FR}} \left|\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle\right|} = \frac{1}{\mathtt{c}_{\mathrm{FR}} \left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \leq \frac{1}{\mathtt{c}_{\mathrm{FR}} \varepsilon_{\mathrm{crit}}^2}. \tag{63}$$

By Assumption 4, there must be a positive upper bound for the (continuous) steepest descent directions:

$$\left\|\boldsymbol{\delta}^{(k)}\right\| \leq \mathtt{C}^{\sharp} \qquad \forall k \in \mathbb{N}_0. \tag{64}$$

Altogether, (60, 62, 64) give

$$\left|\theta_{(k)}\right| \left\|\boldsymbol{\delta}^{(k)}\right\| \leq \frac{\mathtt{C}' \mathtt{L}_f \mathtt{C}^{\sharp} \left\|\sigma_{k-1} \boldsymbol{d}^{(k-1)}\right\|}{\mathtt{c}_{\mathrm{FR}} \varepsilon_{\mathrm{crit}}^2} + \frac{(\mathtt{c}_{\mathrm{FR}} - 1)\, \mathtt{C}^{\sharp}}{\mathtt{c}_{\mathrm{FR}} \varepsilon_{\mathrm{crit}}^2}. \tag{65}$$

Next, we want to establish a similar upper bound for $|\beta_{(k)}| \left\|\boldsymbol{d}^{(k-1)}\right\|$. The assumptions make $\boldsymbol{d}^{(k)}$ H-Hölder continuous by Lemma 4, i.e., there is $\mathtt{H} > 0$ such that

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

Using the reverse triangle inequality we get

$$\begin{aligned}
\left|\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle\right| - \left|\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle\right| &\leq \left|\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle\right| \\
&\leq \left|\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{\delta}^{(k)} \right\rangle + \left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{\delta}^{(k)} \right\rangle - \left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle\right| \\
&\leq \left|\left\langle \boldsymbol{w}^*, \left(\boldsymbol{\nabla f}^{(k)} - \boldsymbol{\nabla f}^{(k-1)}\right) \boldsymbol{\delta}^{(k)} \right\rangle\right| + \left|\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \left(\boldsymbol{\delta}^{(k)} - \boldsymbol{\delta}^{(k-1)}\right) \right\rangle\right| \\
&\leq \mathtt{C}^{\heartsuit} \left\|\sigma_{k-1} \boldsymbol{d}^{(k-1)}\right\| + \mathtt{C}^{\diamondsuit} \sqrt{\left\|\sigma_{k-1} \boldsymbol{d}^{(k-1)}\right\|}.
\end{aligned}$$

The constants $\mathtt{C}^{\heartsuit} > 0$ and $\mathtt{C}^{\diamondsuit} > 0$ stem from the Lipschitz-continuity of the Jacobian, the Hölder-continuity of the steepest descent direction, and the boundedness of C and $\mathcal{F}$.
It follows that

$$\left|\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle\right| \leq \mathtt{C}^{\heartsuit} \left\|\sigma_{k-1} \boldsymbol{d}^{(k-1)}\right\| + \mathtt{C}^{\diamondsuit} \sqrt{\left\|\sigma_{k-1} \boldsymbol{d}^{(k-1)}\right\|} + \left|\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle\right|$$

and eventually

$$\frac{\left|\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle\right|}{\mathtt{c}_{\mathrm{FR}} \left|\left\langle \boldsymbol{w}^*, \boldsymbol{\nabla f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle\right|} \leq \frac{\mathtt{C}^{\heartsuit}}{\varepsilon_{\mathrm{crit}}^2 \mathtt{c}_{\mathrm{FR}}} \left\|\sigma_{k-1} \boldsymbol{d}^{(k-1)}\right\| + \frac{\mathtt{C}^{\diamondsuit}}{\varepsilon_{\mathrm{crit}}^2 \mathtt{c}_{\mathrm{FR}}} \left\|\sigma_{k-1} \boldsymbol{d}^{(k-1)}\right\|^{\frac{1}{2}} + \frac{1}{\mathtt{c}_{\mathrm{FR}}}.$$

Because of (63), this is an upper bound for $\beta_{(k)}$:

$$\beta_{(k)} \leq \frac{\mathtt{C}^{\heartsuit}}{\varepsilon_{\mathrm{crit}}^2 \mathtt{c}_{\mathrm{FR}}} \left\|\sigma_{k-1} \boldsymbol{d}^{(k-1)}\right\| + \frac{\mathtt{C}^{\diamondsuit}}{\varepsilon_{\mathrm{crit}}^2 \mathtt{c}_{\mathrm{FR}}} \left\|\sigma_{k-1} \boldsymbol{d}^{(k-1)}\right\|^{\frac{1}{2}} + \frac{1}{\mathtt{c}_{\mathrm{FR}}} \tag{66}$$

With (65) and (66) equation (59) becomes

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \left(\mathtt{C}^{\dagger} \left\|\sigma_{k-1} \boldsymbol{d}^{(k-1)}\right\| + \mathtt{C}^{\ddagger} \sqrt{\left\|\sigma_{k-1} \boldsymbol{d}^{(k-1)}\right\|} + \frac{1}{\mathtt{c}_{\mathrm{FR}}}\right) \left\|\boldsymbol{d}^{(k-1)}\right\| + \mathtt{C}^{\wr}, \tag{67}$$

with some positive constants $\mathtt{C}^{\dagger}, \mathtt{C}^{\ddagger}, \mathtt{C}^{\wr} > 0$. According to Lemma 11, the steps vanish, i.e.,

$$\left\|\sigma_{k-1} \boldsymbol{d}^{(k-1)}\right\| \xrightarrow{k \to \infty} 0,$$

and because of $\mathtt{c}_{\mathrm{FR}}^{-1} \in (0, 1)$, there must be some $k_0 \in \mathbb{N}_0$ and a constant $\mathtt{r} \in (0, 1)$ such that finally

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \mathtt{r} \left\|\boldsymbol{d}^{(k-1)}\right\| + \mathtt{C}^{\wr}.$$

According to Lemma 40, $\left\|\boldsymbol{d}^{(k)}\right\|$ is uniformly bounded. The proof is concluded by invoking Corollary 39. □

A.6 Fractional-Linear Programming Variant II

Proof of Lemma 37. Like before, the proof is by induction, and we show the stronger inequality

$$\left\langle \boldsymbol{v}, \nabla \boldsymbol{f}^{(k)} \boldsymbol{d}^{(k)} \right\rangle \leq \left\langle \boldsymbol{v}, \nabla \boldsymbol{f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle \leq 0 \qquad \forall \boldsymbol{v} \in C, \forall k \geq 0.$$

The case $k = 0$ is trivial. Let $k \geq 1$ and $\boldsymbol{v} \in C$. Using the definition (FR MO2) and (34) we get

$$\begin{aligned}
\left\langle \boldsymbol{v}, \nabla \boldsymbol{f}^{(k)} \boldsymbol{d}^{(k)} \right\rangle &= \frac{\left(-\left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle - (\mathsf{c}_{\mathrm{FR}} - 1) \left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle\right) \left\langle \boldsymbol{v}, \nabla \boldsymbol{f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle + \psi(\boldsymbol{w}^*, \boldsymbol{v})}{-\mathsf{c}_{\mathrm{FR}} \left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle} \\
&\overset{(57)}{\leq} \frac{\left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle \left\langle \boldsymbol{v}, \nabla \boldsymbol{f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle}{\mathsf{c}_{\mathrm{FR}} \left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle} + \frac{\mathsf{c}_{\mathrm{FR}} - 1}{\mathsf{c}_{\mathrm{FR}}} \frac{\left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle \left\langle \boldsymbol{v}, \nabla \boldsymbol{f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle}{\left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle} \\
&= \frac{\left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle \left\langle \boldsymbol{v}, \nabla \boldsymbol{f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle}{\mathsf{c}_{\mathrm{FR}} \left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle} + \left(1 - \frac{1}{\mathsf{c}_{\mathrm{FR}}}\right) \left\langle \boldsymbol{v}, \nabla \boldsymbol{f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle.
\end{aligned}$$

Now the induction hypothesis gives

$$\underbrace{\frac{\left\langle \boldsymbol{v}, \nabla \boldsymbol{f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle}{\mathsf{c}_{\mathrm{FR}} \left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle}}_{\geq 0} \left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{d}^{(k-1)} \right\rangle \leq \frac{\left\langle \boldsymbol{v}, \nabla \boldsymbol{f}^{(k)} \boldsymbol{\delta}^{(k)} \right\rangle}{\mathsf{c}_{\mathrm{FR}} \left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle} \left\langle \boldsymbol{w}^*, \nabla \boldsymbol{f}^{(k-1)} \boldsymbol{\delta}^{(k-1)} \right\rangle,$$

leading to

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

□

Proof of Theorem 38. Like before, we want to use (59) to bound $\left\|\boldsymbol{d}^{(k)}\right\|$. It is easy to see that (66) is still valid, and we can bound $\left|\beta_{(k)}\right| \left\|\boldsymbol{d}^{(k-1)}\right\|$. Further, we have

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

Thus, there is a bound like (65). The rest of the proof is identical to that of Theorem 35. □

References

[1] Yushan Bai, Jiawei Chen, and Kaiping Liu. A modified Polak-Ribiere-Polyak type conjugate gradient method with two stepsize strategies for vector optimization, April 2024.

[2] To Tanh Binh and Ulrich Korn. An evolution strategy for the multiobjective optimization. In *Proceedings of the 2nd International Conference on Genetic Algorithms*, pages 23–28, June 1996.

[3] Wang Chen, Xinmin Yang, and Yong Zhao. Memory gradient method for multiobjective optimization, June 2022.

[4] Wang Chen, Yong Zhao, Liping Tang, and Xinmin Yang. Conjugate gradient methods without line search for multiobjective optimization, October 2024.

[5] 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.

[6] Indraneel Das and J. E. Dennis. Normal-Boundary Intersection: A New Method for Generating the Pareto Surface in Nonlinear Multicriteria Optimization Problems. *SIAM Journal on Optimization*, 8(3):631–657, August 1998. ISSN 1052-6234, 1095-7189. doi:10.1137/S1052623496307510.

[7] 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.

[8] Peter Duersch, Jörg Oechssler, and Burkhard C. Schipper. Pure strategy equilibria in symmetric two-player zero-sum games. *International Journal of Game Theory*, 41(3):553–564, August 2012. ISSN 0020-7276, 1432-1270. doi:10.1007/s00182-011-0302-x.

[9] D. Dumitrescu, Crina Groşan, and Mihai Oltean. A New Evolutionary Approach for Multiobjective Optimization. *Studia Universitatis Babeş-Bolyai, Informatica*, XLV(1):51–67, January 2000.

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

[11] 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.

[12] Gabriele Eichfelder. Twenty years of continuous multiobjective optimization in the twenty-first century. *EURO Journal on Computational Optimization*, 9:100014, 2021. ISSN 2192-4406. doi:10.1016/j.ejco.2021.100014.

[13] Y Elboulqe and M El Maghri. An explicit spectral Fletcher–Reeves conjugate gradient method for bi-criteria optimization. *IMA Journal of Numerical Analysis*, page drae003, April 2024. ISSN 0272-4979, 1464-3642. doi:10.1093/imanum/drae003.

[14] Y Elboulqe and M El Maghri. An Explicit Three-Term Polak–Ribi‘ere–Polyak Conjugate Gradient Method for Bicriteria Optimization, 2023.

[15] M. Farina. A neural network based generalized response surface multiobjective evolutionary algorithm. In *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*, volume 1, pages 956–961, Honolulu, HI, USA, 2002. IEEE. ISBN 978-0-7803-7282-5. doi:10.1109/CEC.2002.1007054.

[16] J. Fliege, L. M. Graña Drummond, and B. F. Svaiter. Newton's Method for Multiobjective Optimization. *SIAM Journal on Optimization*, 20(2):602–626, January 2009. ISSN 1052-6234, 1095-7189. doi:10.1137/08071692X.

[17] 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.

[18] 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.

[19] Carlos M. Fonseca and Peter J. Fleming. An Overview of Evolutionary Algorithms in Multiobjective Optimization. *Evolutionary Computation*, 3(1):1–16, March 1995. ISSN 1063-6560, 1530-9304. doi:10.1162/evco.1995.3.1.1.

[20] C.M. Fonseca and P.J. Fleming. Multiobjective genetic algorithms made easy: Selection sharing and mating restriction. In *First International Conference on Genetic Algorithms in Engineering Systems: Innovations and Applications*, pages 45–52, September 1995. doi:10.1049/cp:19951023.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] Qing-Rui He, Chun-Rong Chen, and Sheng-Jie Li. Spectral conjugate gradient methods for vector optimization problems. *Computational Optimization and Applications*, 86(2):457–489, November 2023. ISSN 0926-6003, 1573-2894. doi:10.1007/s10589-023-00508-w.

[26] C. Hillermeier. Generalized Homotopy Approach to Multiobjective Optimization. *Journal of Optimization Theory and Applications*, 110(3):557–583, September 2001. ISSN 0022-3239, 1573-2878. doi:10.1023/A:1017536311488.

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

[28] S. Huband, P. Hingston, L. Barone, and L. While. A review of multiobjective test problems and a scalable test problem toolkit. *IEEE Transactions on Evolutionary Computation*, 10(5):477–506, October 2006. ISSN 1089-778X. doi:10.1109/TEVC.2005.861417.

[29] K. Ikeda, H. Kita, and S. Kobayashi. Failure of Pareto-based MOEAs: Does non-dominated really mean near to optimal? In *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546)*, volume 2, pages 957–962 vol. 2, May 2001. doi:10.1109/CEC.2001.934293.

[30] Yaochu Jin, Markus Olhofer, and Bernhard Sendhoff. Dynamic Weighted Aggregation for Evolutionary Multi-Objective Optimization: Why Does It Work and How?

[31] I.Y. Kim and O.L. De Weck. Adaptive weighted-sum method for bi-objective optimization: Pareto front generation. *Structural and Multidisciplinary Optimization*, 29(2):149–158, February 2005. ISSN 1615-147X, 1615-1488. doi:10.1007/s00158-004-0465-1.

[32] Alberto Lovison. Singular Continuation: Generating Piece-wise Linear Approximations to Pareto Sets via Global Analysis. *SIAM Journal on Optimization*, 21(2):463–490, April 2011. ISSN 1052-6234, 1095-7189. doi:10.1137/100784746.

[33] 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.

[34] 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.

[35] 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.

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

[37] E. Miglierina, E. Molho, and M.C. Recchioni. Box-constrained multi-objective optimization: A gradient-like method without "a priori" scalarization. *European Journal of Operational Research*, 188(3):662–682, August 2008. ISSN 03772217. doi:10.1016/j.ejor.2007.05.015.

[38] Kanako Mita, Ellen H. Fukuda, and Nobuo Yamashita. Nonmonotone line searches for unconstrained multiobjective optimization problems. *Journal of Global Optimization*, 75(1):63–90, September 2019. ISSN 0925-5001, 1573-2916. doi:10.1007/s10898-019-00802-0.

[39] Jorge J. Moré, Burton S. Garbow, and Kenneth E. Hillstrom. Testing Unconstrained Optimization Software. *ACM Transactions on Mathematical Software*, 7(1):17–41, March 1981. ISSN 0098-3500, 1557-7295. doi:10.1145/355934.355936.

[40] 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.

[41] 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.

[42] 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.

[43] Carlo Poloni, Giovanni Mosetti, Stefano Contessi, et al. Multi objective optimization by GAs: Application to system and component design. In *Eccomas 96*, pages 1–7. John Wiley & Sons, Ltd, 1996.

[44] Mike Preuss, Boris Naujoks, and Günter Rudolph. Pareto Set and EMOA Behavior for Simple Multimodal Multiobjective Functions. In Thomas Philip Runarsson, Hans-Georg Beyer, Edmund Burke, Juan J. Merelo-Guervós, L. Darrell Whitley, and Xin Yao, editors, *Parallel Problem Solving from Nature - PPSN IX*, volume 4193, pages 513–522. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-38990-3 978-3-540-38991-0. doi:10.1007/11844297_52.

[45] Manuel Valenzuela Rendón and Eduardo Uresti-Charre. A non-generational genetic algorithm for multiobjective optimization. In *Proc. 7th Interational Conference on Genetic Algorithms*, pages 658–665, 1997.

[46] Oliver Schütze, Marco Laumanns, Carlos A. Coello Coello, Michael Dellnitz, and El-Ghazali Talbi. Convergence of stochastic search algorithms to finite size pareto set approximations. *Journal of Global Optimization*, 41(4): 559–577, August 2008. ISSN 0925-5001, 1573-2916. doi:10.1007/s10898-007-9265-7.

[47] M. Sefrioui and J. Perlaux. Nash genetic algorithms: Examples and applications. In *Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No.00TH8512)*, volume 1, pages 509–516 vol.1, July 2000. doi:10.1109/CEC.2000.870339.

[48] Mun-Bo Shim, Myung-Won Suh, Tomonari Furukawa, Genki Yagawa, and Shinobu Yoshimura. Pareto-based continuous evolutionary algorithms for multiobjective optimization. *Engineering Computations*, 19(1):22–48, January 2002. ISSN 0264-4401. doi:10.1108/02644400210413649.

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

[50] 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.

[51] 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.

[52] Ph L Toint. Test problems for partially separable optimization and results for the routine PSPMIN. *The University of Namur, Department of Mathematics, Belgium, Tech. Rep*, 1983.

[53] J. V. Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, December 1928. ISSN 0025-5831, 1432-1807. doi:10.1007/BF01448847.

[54] David Allen Van Veldhuizen. *Multiobjective Evolutionary Algorithms: Classifications, Analyses, and New Innovations*. PhD thesis, Air Force Institute of Technology, USA, 1999.

[55] 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.

[56] 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.