NONLINEAR CONJUGATE GRADIENT METHODS WITH GUARANTEED DESCENT FOR MULTI-OBJECTIVE OPTIMZATION

A PREPRINT

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

Manuel Berkemeier.
Department of Computer Science
Paderborn University, Germany
sebastian.peitz@upb.de

May 31, 2023

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

ABSTRACT

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

Todo list

1 Introduction

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

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

{eqn: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 $\boldsymbol{x} \in \mathbb{R}^N$ with $\boldsymbol{x} \neq \boldsymbol{x}^*$ and $\boldsymbol{f}(\boldsymbol{x}) \prec_{\mathcal{K}} \boldsymbol{f}(\boldsymbol{x}^*) \iff \boldsymbol{f}(\boldsymbol{x}^*) - \boldsymbol{f}(\boldsymbol{x}) \in \operatorname{int}(\mathcal{K})$. In practical applications, one often encounters $\mathcal{K} = \mathbb{R}^K_{\geq 0}$. There is a multitude of methods to solve MOPs.

Elaborate and list these methods, add corresponding references.

Conjugate Gradient Methods

Originally, the conjugate gradient (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 [8]. 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 [8].

Recently, Lucambio Pérez and Prudente [6] have adapted many of the popular nonlinear CG methods to the multi-objective setting. Their directions [3, 6] rely on strong Wolfe conditions being fulfilled. To this end, a suitable step-size algorithm is provided [7]. These multi-objective nonlinear CG methods work well in experiments, but the line-search algorithm might require step-sizes that are undesirably large, and its implementation is more involved than using simple Armijo-like backtracking. In [4], a nonlinear CG method for vector optimization is proposed that works with a simple backtracking algorithm, but it requires estimates of the Lipschitz constants of the objectives.

In contrast, the directions in this work satisfy a *sufficient decrease* condition by construction – independent of the line-search. The directions are adapted (or "translated") from the single-objective setting, albeit there already is a scheme for bi-objective optimization [2].

2 Criticality

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

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

Vice versa, if $\boldsymbol{x}$ is not critical, then there is a *descent direction* $\boldsymbol{v} \in \mathbb{R}^N$, with the defining property

$$\boldsymbol{\nabla f}(\boldsymbol{x}) \cdot \boldsymbol{v} \in -\operatorname{int}(\mathcal{K}).$$

For such a direction, there is some step-size bound $\bar{\sigma} > 0$ such that

$$\boldsymbol{f}(\boldsymbol{x} + \sigma \boldsymbol{v}) \preceq_{\mathcal{K}} \boldsymbol{f}(\boldsymbol{x}) \qquad \forall \sigma \in (0, \bar{\sigma}) \quad \text{(see [5]).}$$

Make the above proper Definition(s)?

Adopting the notation from [5, 6], let $\langle \bullet, \bullet \rangle$ be the usual inner product 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 $C \subset \mathcal{K}^* \setminus \{\boldsymbol{0}\}$ be a compact set generating $\mathcal{K}^*$ as its conical hull: $\operatorname{coni}(C) = \mathcal{K}^*$. Then, the map

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

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

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

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

Theorem 1 (Lemma 3.1 in [5])**.** *Let $\boldsymbol{y}, \boldsymbol{y\prime} \in \mathbb{R}^K$. Then*

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

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

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

then we can infer criticality from the function $\mathfrak{f}$.

thm:criticality_properties

Theorem 2 (Lemma 3.3 in [5]). *Suppose $\boldsymbol{f}$ is continuously differentiable. Consider the following optimization problem:*

{eqn:sd_problem}

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

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

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

Moreover, in [5] it is shown that

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

In the single-objective case, with $\mathcal{K} = \mathbb{R}^K_{\geq 0}$, the solution is $\boldsymbol{\delta} = -\nabla f(\boldsymbol{x})$. The problem in (1) thus generalizes the concept of the steepest descent direction, and we obtain a recipe for "translating" nonlinear CG directions for multiple objectives. Just as in single-objective optimization a sequence of directions $\boldsymbol{d}^{(k)} \in \mathbb{R}^N$ is said to fulfill the sufficient decrease condition if there is a constant $\kappa_{\text{sd}} > 0$ such that $\left\langle -\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)}), \boldsymbol{d}^{(k)} \right\rangle \geq \kappa_{\text{sd}} \left\|\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})\right\|^2$, we qualify them accordingly in the multi-objective case:

Definition 3. The directions $\left\{\boldsymbol{d}^{(k)}\right\}$ are said to have the *sufficient decrease* property iff

{eqn:suff_dec}

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

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

3 Algorithm

The algorithm will be stated in a very generic manner. That is, we do not yet give specific formulas to compute the directions $\left\{\boldsymbol{d}^{(k)}\right\}$, but only assume them to have the sufficient decrease property (2). Additionally, we have to determine step-sizes. In the next subsection, we justify a simple backtracking procedure.

3.1 (Modified) Armijo Stepsize

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

{eqn:amidoDinarchy}

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

Let $\mathtt{r} \in (0,1)$. The step-size $\sigma > 0$ satisfies the modified Armijo condition if

{eqn:armijo_strict}

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

There is a suitable step-size $\sigma > 0$.

Proof. Suppose there was not:

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

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

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

Rearranging and dividing by $\sigma > 0$ gives

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

$$\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x})\boldsymbol{d}\rangle > -\mathtt{r}\sigma \|\boldsymbol{d}\|^2 - \left\langle \boldsymbol{w}, \frac{\boldsymbol{R}(\sigma)}{\sigma}\right\rangle \tag{5}$$

{eqn:disembranglingJonty}

where, by definition of the total differential,

$$\frac{\boldsymbol{R}(\sigma)}{\sigma} \to \boldsymbol{0} \quad \text{as } \sigma \to 0 \text{ from above.}$$

Because $\boldsymbol{d}$ is a descent direction, the value on the left-hand side (LHS) in (5) is constant and strictly negative, while the right-hand side (RHS) goes to zero. A contradiction! □

A step-size satisfying (4) 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 $\mathtt{r} \in (0,1)$ be constants and $\sigma_0^{(k)}$ an initial step-size bounded below by the constant $\mathtt{M} > 0$. Take

{eqn:backtracking_stepsize}

$$\sigma_{(k)} = \max_{j \in \mathbb{N}_0} \mathtt{b}^j \sigma_0^{(k)} \quad \text{such that (4) holds.} \tag{6}$$

3.2 Algorithm

We are now in a position to state the complete algorithm in Algorithm 1.

Algorithm 1: Algorithm with Generic Descent Direction

algo:main_algo

Data: $N \in \mathbb{N}, K \in \mathbb{N}, \boldsymbol{f} \colon \mathbb{R}^N \to \mathbb{R}^K, \boldsymbol{x}^{(0)} \in \mathbb{R}^N, \mathtt{r} \in (0,1), \mathtt{b} \in (0,1), \kappa_{\text{sd}} > 0, \sigma_0^{(k)} \geq \mathtt{M} > 0$.
Result: A critical point $\boldsymbol{x}^{(k)}$ or a sequence $\{\boldsymbol{x}^{(k)}\}_{k \in \mathbb{N}_0}$, containing a critical accumulation point.
for $k \in \mathbb{N}_0$ **do**
  if $\boldsymbol{x}^{(k)}$ *is critical* **then** STOP;
  Compute a direction $\boldsymbol{d}^{(k)}$ satisfying (2);
  Compute a step-size $\sigma_{(k)}$ by backtracking like in (6);
  Set $\boldsymbol{x}^{(k+1)} \leftarrow \boldsymbol{x}^{(k)} + \sigma_{(k)}\boldsymbol{d}^{(k)}$;
end

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

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

$$\mathcal{F} = \left\{ \boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{f}(\boldsymbol{x}) \preceq_{\mathcal{K}} \boldsymbol{f}(\boldsymbol{x}^{(0)}) \right\}.$$

Furthermore, $\boldsymbol{f}$ is continuously differentiable in an open set containing $\mathcal{F}$ and its Jacobian $\boldsymbol{\nabla f}$ is Lipschitz continuous with constant $\mathtt{L}_{\boldsymbol{f}} > 0$.

ass:monotonic_seq_bounded

Assumption 2. Every non-increasing sequence in $\boldsymbol{f}(\mathcal{F})$,

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

is bounded below by some $\boldsymbol{y}_{\text{lb}} \in \mathbb{R}^K$ as per $\boldsymbol{y}_{\text{lb}} \preceq_{\mathcal{K}} \boldsymbol{y}^{(k)}$ for all k. In terms of φ, this means $\varphi(\boldsymbol{y}_{\text{lb}}) - \varphi(\boldsymbol{y}^{(k)}) \leq \varphi(\boldsymbol{y}_{\text{lb}} - \boldsymbol{y}^{(k)}) \leq 0$ for all k.

thm:stepNorm_zero

Lemma 4. *Consider an algorithmic sequence $\left\{(\boldsymbol{x}^{(k)}, \boldsymbol{d}^{(k)}, \sigma_{(k)})\right\}_k$. Suppose Assumption 2 holds. Then*

$$\lim_{k \to \infty} \sigma^2 \|\boldsymbol{d}_k\|^2 = \lim_{k \to \infty} \sigma^2 \|\boldsymbol{d}_k\|^2 = 0.$$

Proof. The Armijo condition (4) holds. With Theorem 1 it follows that

$$\varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(k)}+\sigma_{(k)}\boldsymbol{d}^{(k)})\right)-\varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(k)})\right)\le\varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(k)}+\sigma_{(k)}\boldsymbol{d}^{(k)})-\boldsymbol{f}(\boldsymbol{x}^{(k)})\right)\le\varphi\left(-\mathtt{r}\sigma^2\left\|\boldsymbol{d}^{(k)}\right\|^2\boldsymbol{e}\right)$$

$$\Leftrightarrow\quad\varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(k)})\right)-\varphi\left(\boldsymbol{f}(\boldsymbol{x}^{(k)}+\sigma_{(k)}\boldsymbol{d}^{(k)})\right)\ge-\varphi\left(-\mathtt{r}\sigma_{(k)}^2\left\|\boldsymbol{d}^{(k)}\right\|^2\boldsymbol{e}\right)=\mathtt{r}\sigma_{(k)}^2\left\|\boldsymbol{d}^{(k)}\right\|^2\underbrace{\inf_{\boldsymbol{w}\in C}\langle\boldsymbol{w},\boldsymbol{e}\rangle}_{\ge\mathtt{c}_{\boldsymbol{e}}>0}.$$

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

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

Due to Assumption 2, the RHS simplifies:

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

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

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

□

thm:armijoZoutendijk

Theorem 5. *Suppose Assumptions 1 and 2 hold and that the step-sizes $\sigma_{(k)}$ in Algorithm 1 satisfy Eq.* (4)*. Then following Zoutendijk-like condition follows:*

$$\sum_{k\in\mathbb{N}_0}\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2}=\sum_{k\in\mathbb{N}_0}\frac{\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})^2}{\left\|\boldsymbol{d}^{(k)}\right\|^2}<\infty \tag{7}$$

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

First, suppose $\sigma_{(k)}\ne\sigma_0^{(k)}$. Due to the backtracking procedure, the Armijo condition must be violated for $\sigma_{(k)}\mathtt{b}^{-1}>\sigma_{(k)}$. I.e., there is $\boldsymbol{w}\in C$ such that

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

It follows, that

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

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

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

where in the last line we have used the Cauchy-Schwarz inequality and the Lipschitz continuity of $\boldsymbol{\nabla f}$ by Assumption 1. Because C is compact, $\|\boldsymbol{w}\|$ is bounded and there is a constant $\mathtt{c}>0$ such that

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

Plugging this into the Armijo condition for $\sigma_{(k)}$ gives

$$\varphi(\boldsymbol{f}(\boldsymbol{x}^{(k)}))-\varphi\left(\boldsymbol{f}\left(\boldsymbol{x}^{(k)}\right)\right)\ge\mathtt{rcc}_{\boldsymbol{e}}\frac{\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k)})^2}{\left\|\boldsymbol{d}^{(k)}\right\|^2}\ge\mathtt{rcc}_{\boldsymbol{e}}\kappa_{\mathrm{sd}}\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2}. \tag{8}$$

{eqn:untuningCreatrixes}

Now, suppose $\sigma_{(k)} = \sigma_0^{(k)}$. By definition of $\boldsymbol{\delta}^{(k)}$ as the minimizer of (1), we have

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

Because $\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\kappa_{\mathrm{sd}}^{-1}\boldsymbol{d}^{(k)}) \leq \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})$, it must hold that $\left\|\boldsymbol{\delta}^{(k)}\right\|^2 \leq \frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{\kappa_{\mathrm{sd}}^2}$. Thus,

$$\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \leq \frac{1}{\kappa_{\mathrm{sd}}^4}\left\|\boldsymbol{d}^{(k)}\right\|^2 \leq \frac{1}{\kappa_{\mathrm{sd}}^4 \mathtt{r}\left(\sigma_0^{(k)}\right)^2}\left(\varphi(\boldsymbol{f}(\boldsymbol{x}^{(k)})) - \varphi\left(\boldsymbol{f}\left(\boldsymbol{x}^{(k)}\right)\right)\right)$$

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

$$\bar{\mathtt{c}}\frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2} \leq \varphi(\boldsymbol{f}(\boldsymbol{x}^{(k)})) - \varphi\left(\boldsymbol{f}\left(\boldsymbol{x}^{(k)}\right)\right) \qquad \text{\{eqn:skiascopyNoncoital\}} \quad (9)$$

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

Similarly to above, with Assumption 2 we can deduce convergence from (8) and (9):

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

□

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

Corollary 6. *Suppose, the criticality is uniformly bounded from below via*

$$\left\|\boldsymbol{d}^{(k)}\right\| \geq \varepsilon_{\mathrm{crit}} > 0 \qquad \forall k \in \mathbb{N}_0. \qquad \text{\{eqn:critBoundedBelow\}} \quad (10)$$

If the directions $\left\{\boldsymbol{d}^{(k)}\right\}$ *and step-sizes are chosen so that Theorem 5 applies,* ***and*** *it additionally holds that*

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

then $\liminf_{k\to\infty}\left\|\boldsymbol{\delta}^{(k)}\right\| = 0$. *That is, there is a subsequence of iterates* $\{\boldsymbol{x}^{(k)}\}$ *converging to a Pareto-critical point.*

We might refer to a subsequence converging to a Pareto-critical point as a *critical* sequence later on. If the directions $\left\{\boldsymbol{d}^{(k)}\right\}$ remain bounded, then this constitutes a special case: `thm:dirsBoundedConvergence`

Corollary 7. *Suppose that* (10) *holds and that Theorem 5 applies. If there is a constant* $\mathtt{C}_{\boldsymbol{d}} > 0$ *with*

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

then there is a critical subsequence.

4 Specific Directions with Guaranteed Descent

4.1 Two Flavors of Fletcher-Reeves

We will show two fun variants derived from the single-objective scheme in [11]. The directions $\boldsymbol{d}^{(k)}$ are inspired by the classical Fletcher-Reeves (FR) recipe and given by

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

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

FR Restart Variant

Unfortunately, simply replacing $-\boldsymbol{g}^{(k)}$ with the multi-objective steepest descent direction $\boldsymbol{\delta}^{(k)}$ from (1) and modifying the product in $\theta_{(k)}$ as per

$$\tilde{\beta}_{(k)} = \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} = \frac{\mathfrak{f}^{(k)}_{\boldsymbol{x}}(\boldsymbol{\delta}^{(k)})}{\mathfrak{f}^{(k-1)}_{\boldsymbol{x}}(\boldsymbol{\delta}^{(k-1)})}, \quad \tilde{\theta}_{(k)} = \frac{\mathfrak{f}^{(k-1)}_{\boldsymbol{x}}(\boldsymbol{d}^{(k-1)}) - \mathfrak{f}^{(k)}_{\boldsymbol{x}}(\boldsymbol{d}^{(k-1)})}{\mathfrak{f}^{(k-1)}_{\boldsymbol{x}}(\boldsymbol{\delta}^{(k-1)})} \qquad \text{\{eqn:hydrosulphurousTontiners\}} \tag{11}$$

does not suffice to show the sufficient decrease property (2). $\tilde{\theta}_{(k)}$ might become negative. This can be avoided with standard Wolfe conditions. Or with restarts, as in the following definition:

$$\boldsymbol{d}^{(k)} = \begin{cases} \boldsymbol{\delta}^{(k)} & \text{if } k = 0, \\ \theta_{(k)}\boldsymbol{\delta}^{(k)} + \beta_{(k)}\boldsymbol{d}^{(k-1)}, & \text{if } k \geq 1, \end{cases} \quad (\theta_{(k)}, \beta_{(k)}) = \begin{cases} (\tilde{\theta}_{(k)}, \tilde{\beta}_{(k)}) & \text{if } \tilde{\theta}_{(k)} < 0, \\ (1, 0) & \text{else.} \end{cases} \qquad \text{\{eqn:frMO1\}} \tag{FR MOI}$$

Theorem 8. *The directions in* (FR MOI) *have the sufficient decrease property* (2) *with* $\kappa_{\mathrm{sd}} = 1$.

Proof. For $k = 0$ and $\tilde{\theta}_{(k)}$ there is nothing to show. Assume $k \geq 1$ and $\tilde{\theta}_{(k)} \geq 0$. Let $\boldsymbol{w} \in C$.

$$\begin{aligned} \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k)} \right\rangle &= \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) \left(\theta_{(k)}\boldsymbol{\delta}^{(k)} + \beta_{(k)}\boldsymbol{d}^{(k-1)}\right) \right\rangle \\ &= \theta_{(k)} \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{\delta}^{(k)} \right\rangle + \beta_{(k)} \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k-1)} \right\rangle \\ &\leq \theta_{(k)}\mathfrak{f}^{(k)}_{\boldsymbol{x}}(\boldsymbol{\delta}^{(k)}) + \beta_{(k)}\mathfrak{f}^{(k)}_{\boldsymbol{x}}(\boldsymbol{d}^{(k-1)}) \\ &\overset{(11)}{=} \mathfrak{f}^{(k)}_{\boldsymbol{x}}(\boldsymbol{\delta}^{(k)}) \frac{\mathfrak{f}^{(k-1)}_{\boldsymbol{x}}(\boldsymbol{d}^{(k-1)}) \cancel{- \mathfrak{f}^{(k)}_{\boldsymbol{x}}(\boldsymbol{d}^{(k-1)}) + \mathfrak{f}^{(k)}_{\boldsymbol{x}}(\boldsymbol{d}^{(k-1)})}}{\mathfrak{f}^{(k-1)}_{\boldsymbol{x}}(\boldsymbol{\delta}^{(k-1)})}. \end{aligned}$$

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

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

□

thm:frMO1convergence

Theorem 9. *Suppose Assumptions 1 and 2 hold and that the criticality* $\left\|\boldsymbol{\delta}^{(k)}\right\|$ *is bounded below like in* (10). *Then the Algorithm with modified Armijo-stepsizes according to* (4) *and with directions defined by* (FR MOI) *generates a critical subsequence.*

Proof. Denote by $\mathcal{P} \subseteq \mathbb{N}_0$ those iteration indices for which $\tilde{\theta}_{(k)} \geq 0$ and by $\mathcal{N} \subseteq \mathbb{N}_0$ the indices with $\tilde{\theta}_{(k)} < 0$.

The case $\mathcal{P} = \emptyset$ reduces to $\boldsymbol{d}^{(k)} = \boldsymbol{\delta}^{(k)}$ for all k and

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

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

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

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

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

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

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

resulting in

$$\left\|\boldsymbol{d}^{(k)}\right\|^2 = \beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2 - \theta_{(k)}^2 \left\|\boldsymbol{\delta}^{(k)}\right\|^2 + 2\theta_{(k)} \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle. \quad \text{\{eqn:ethylsBabyproofed\}} \tag{12}$$

As shown in [5], the steepest descent direction is a weighted sum of gradients. More precisely, let $\tilde{C} = \operatorname{conv}(C)$, then

$$\boldsymbol{\delta}^{(k)} = \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})^{\mathsf{T}} \boldsymbol{w}^* \quad \text{with} \quad \boldsymbol{w}^* = \arg\min_{\boldsymbol{w} \in \tilde{C}} \left\| \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})^{\mathsf{T}} \boldsymbol{w} \right\|^2 .$$

Thus,

$$\begin{aligned}
\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle &= - \left\langle \boldsymbol{d}^{(k)}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})^{\mathsf{T}} \boldsymbol{w}^* \right\rangle \\
&\geq \min_{\boldsymbol{w} \in C} - \left\langle \boldsymbol{d}^{(k)}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) \boldsymbol{w} \right\rangle = - \max_{\boldsymbol{w} \in C} \left\langle \boldsymbol{d}^{(k)}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) \boldsymbol{w}^{\mathsf{T}} \right\rangle = -\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k)}) \\
&\geq -\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)}) \geq 0,
\end{aligned}$$

where we have used the fact that it does not matter whether we calculate $\mathfrak{f}_{\boldsymbol{x}}^{(k)}$ over C or $\tilde{C}$, and lastly the sufficient decrease property of $\boldsymbol{d}^{(k)}$. It follows that

$$\frac{2\theta_{(k)} \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} = \frac{2\theta_{(k)} \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle}{(-\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)}))(-\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)}))} \leq \frac{2\theta_{(k)} \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle}{\left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle (-\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)}))} = \frac{2\theta_{(k)}}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}.$$

Combining this with (12) gives

$$\begin{aligned}
\frac{\left\|\boldsymbol{d}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} &= \frac{\beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} - \frac{\theta_{(k)}^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} + \frac{2\theta_{(k)} \left\langle \boldsymbol{\delta}^{(k)}, \boldsymbol{d}^{(k)} \right\rangle}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} \\
&\leq \frac{\beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} - \frac{\theta_{(k)}^2 - 2\theta_{(k)}}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} \\
&= \frac{\beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} - \frac{(\theta_{(k)} - 1)^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} + \frac{1}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} \\
&\leq \frac{\beta_{(k)}^2 \left\|\boldsymbol{d}^{(k-1)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k)}\right\|^4} + \frac{1}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}. \qquad \text{\{eqn:recalcitratedLinearly\}}
\end{aligned} \tag{13}$$

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

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

Recursion gives

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

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

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

This concludes the proof, as Corollary 6 applies. □

FR MaxiMin Variant

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

$$\begin{aligned}
\theta_{(k)}(\boldsymbol{w}) &:= \frac{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) \boldsymbol{d}^{(k-1)} \right\rangle - \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k-1)}) \boldsymbol{d}^{(k-1)} \right\rangle}{-\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})} \qquad \text{and} \qquad \text{\{eqn:characteriserAbout\}} \\
\beta_{(k)} &:= \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^2}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} = \frac{-\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})}{-\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})} = \frac{-\max_{\boldsymbol{w} \in C} \left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)}) \boldsymbol{\delta}^{(k)} \right\rangle}{-\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})}.
\end{aligned} \tag{14}$$

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

{eqn:doorsAttention}

$$\max_{\boldsymbol{v}\in C}\min_{\boldsymbol{w}\in C}\left\langle \boldsymbol{v}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\left(\theta_{(k)}(\boldsymbol{w})\boldsymbol{\delta}^{(k)} + \beta_{(k)}\boldsymbol{d}^{(k-1)}\right)\right\rangle. \tag{15}$$

The directions follow as

{eqn:frMO2}

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

Theorem 10. *The directions in* (FR MOII) *have the sufficient decrease property* (2) *with* $\kappa_{\mathrm{sd}} = 1$.

Proof. The property trivially holds for $k = 0$. Let $k \geq 1$. By construction (FR MOII), it holds that

$$\begin{aligned}
\mathfrak{f}^{(k)}_{\boldsymbol{x}}(\boldsymbol{d}^{(k)}) &= \left\langle \boldsymbol{v}^*, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k)}\right\rangle\\
&= \left\langle \boldsymbol{v}^*, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\left(\theta_{(k)}(\boldsymbol{w}^*)\boldsymbol{\delta}^{(k)} + \beta_{(k)}\boldsymbol{d}^{(k-1)}\right)\right\rangle\\
&= \theta_{(k)}(\boldsymbol{w}^*)\left\langle \boldsymbol{v}^*, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{\delta}^{(k)}\right\rangle + \beta_{(k)}\left\langle \boldsymbol{v}^*, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k-1)}\right\rangle\\
&\overset{(15)}{\leq} \theta_{(k)}(\boldsymbol{v}^*)\left\langle \boldsymbol{v}^*, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{\delta}^{(k)}\right\rangle + \beta_{(k)}\left\langle \boldsymbol{v}^*, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k-1)}\right\rangle\\
&\overset{(*)}{\leq} \frac{-\left\langle \boldsymbol{v}^*, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{\delta}^{(k)}\right\rangle\left\langle \boldsymbol{v}^*, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k-1)})\boldsymbol{d}^{(k-1)}\right\rangle}{-\mathfrak{f}^{(k-1)}_{\boldsymbol{x}}(\boldsymbol{\delta}^{(k-1)})}\\
&\overset{df}{\leq} \mathfrak{f}^{(k)}_{\boldsymbol{x}}(\boldsymbol{\delta}^{(k)})\frac{\mathfrak{f}^{(k-1)}_{\boldsymbol{x}}(\boldsymbol{d}^{(k-1)})}{\mathfrak{f}^{(k-1)}_{\boldsymbol{x}}(\boldsymbol{\delta}^{(k-1)})}.
\end{aligned}$$

For the bound $(*)$ we have plugged in the definition (14) and used

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

Finally, sufficient decrease follows by induction because of $\frac{\mathfrak{f}^{(k-1)}_{\boldsymbol{x}}(\boldsymbol{d}^{(k-1)})}{\mathfrak{f}^{(k-1)}_{\boldsymbol{x}}(\boldsymbol{\delta}^{(k-1)})} \leq 1$. □

Theorem 11. *Suppose Assumptions 1 and 2 hold and that the criticality* $\left\|\boldsymbol{\delta}^{(k)}\right\|$ *is bounded below like in* (10). *Then the Algorithm with modified Armijo-stepsizes according to* (4) *and with directions defined by* (FR MOII) *generates a critical subsequence.*

Proof. The proof is the same as that of Theorem 9 for the case $|\mathcal{P}| = \infty$ and $|\mathcal{N}| < \infty$. To see this, note that the bound (13) still holds and $\beta_{(k)}$ is the same. Thus, we get divergence of $\sum_k \frac{\left\|\boldsymbol{\delta}^{(k)}\right\|^4}{\left\|\boldsymbol{d}^{(k)}\right\|^2}$ in contradiction to Corollary 6. □

4.2 Three-Term Polak-Ribière-Polyak Scheme

Zhang et al. [10] define the following three-term Polak-Ribière-Polyak (PRP) directions for single-objective optimization:

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

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

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

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

Unfortunately, it is not sufficient to simply replace $-\boldsymbol{g}^{(k)}$ with $\boldsymbol{\delta}^{(k)}$, the minimizer in (1), to obtain a multi-objective scheme. But the following MaxiMin directions work.

For $k \geq 1$ and $\boldsymbol{w} \in C$, define $\boldsymbol{y}^{(k)} = \boldsymbol{\delta}^{(k-1)} - \boldsymbol{\delta}^{(k)}$ and

{eqn:factorialsPreemergent}

$$\beta_{(k)}(\boldsymbol{w}) = \frac{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{y}^{(k)}\right\rangle}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \quad \text{and} \quad \theta_{(k)}(\boldsymbol{w}) = \frac{\left\langle \boldsymbol{w}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k-1)}\right\rangle}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2}. \tag{16}$$

Then, let

$$\boldsymbol{d}^{(k)} = \begin{cases} \boldsymbol{\delta}^{(k)} & \text{if } k = 0 \\ \boldsymbol{\delta}^{(k)} + \beta_{(k)}(\boldsymbol{w}^*)\boldsymbol{d}^{(k-1)} - \theta_{(k)}(\boldsymbol{w}^*)\boldsymbol{y}^{(k)}, & \text{if } k \geq 1. \end{cases} \quad \text{{eqn:prpMO1}} \quad \text{(PRP MOI)}$$

where $(\boldsymbol{v}^*, \boldsymbol{w}^*) \in C \times C$ solve

$$\varphi\left(\min_{\boldsymbol{w}\in C} \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{d}^{(k)}(\boldsymbol{w})\right) = \max_{\boldsymbol{v}\in C}\min_{\boldsymbol{w}\in C}\left\langle \boldsymbol{v}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\left(\boldsymbol{\delta}^{(k)} + \beta_{(k)}(\boldsymbol{w})\boldsymbol{d}^{(k-1)} - \theta(\boldsymbol{w})\boldsymbol{y}^{(k)}\right)\right\rangle. \quad (17)$$

Theorem 12. *The directions in* (PRP MOI) *have the sufficient decrease property* (2) *with* $\kappa_{\text{sd}} = 1$.

Proof. The case $k = 0$ is trivial. Let $k \geq 1$. Then, by definition of $(\boldsymbol{v}^*, \boldsymbol{w}^*)$,

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

Plugging in (16) makes the trailing terms vanish:

$$\mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{d}^{(k)}) \leq \left\langle \boldsymbol{v}^*, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{\delta}^{(k)}\right\rangle \leq \max_{\boldsymbol{v}*\in C}\left\langle \boldsymbol{v}, \boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\boldsymbol{\delta}^{(k)}\right\rangle = \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)}) \leq 0.$$

□

Proving convergence for these specific directions requires stricter assumptions than provided by Assumption 2. ass:sublevelset_bounded

Assumption 3. The sublevel set

$$\mathcal{F} = \left\{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{f}(\boldsymbol{x}) \preceq_{\mathcal{K}} \boldsymbol{f}(\boldsymbol{x}^{(0)})\right\} \subseteq \mathbb{R}^N$$

is bounded. thm:prpMO1convergence

Theorem 13. *Suppose that Assumptions 1 and 3 hold. If the step-sizes $\sigma_{(k)}$ satisfy the modified Armijo-condition* (4), *then the Algorithm with directions* (PRP MOI) *has a subsequence converging to a critical point.*

Proof. For a proof by contradiction, assume that the criticality $\left\|\boldsymbol{\delta}^{(k)}\right\|$ is bounded below like in (10). Because of Assumptions 1 and 3, the norm of the steepest descent is also uniformly bounded above by a constant $\mathtt{C}_{\boldsymbol{\delta}} > 0$. If we can show the same for $\boldsymbol{d}^{(k)}$, that concludes the proof because of Corollary 7.

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

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \left\|\boldsymbol{\delta}^{(k)}\right\| + \frac{2\left\|\boldsymbol{w}^*\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\right\|\left\|\boldsymbol{\delta}^{(k-1)} - \boldsymbol{\delta}^{(k)}\right\|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2}\left\|\boldsymbol{d}^{(k-1)}\right\|. \quad \text{{eqn:hungeringlyTouches}} \quad (18)$$

Because of the assumptions and [9], we know the steepest descent direction to be $\mathtt{H}$-Hölder continuous:

Hölder-continuity is show in [9] for the case $\mathcal{K} = \mathbb{R}^N$. Should work for other cones as well, but better check!

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

Furthermore, the Jacobians must be bounded and because C is compact, there is some constant $\mathtt{C}_C > 0$ with $\left\|\boldsymbol{w}^*\boldsymbol{\nabla f}(\boldsymbol{x}^{(k)})\right\| \leq \mathtt{C}_C$ for all k. Thus, (18) leads to

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \mathtt{C}_{\boldsymbol{\delta}} + \frac{2\mathtt{C}_C\mathtt{H}\left\|\sigma^{(k-1)}\boldsymbol{d}^{(k-1)}\right\|^{\frac{1}{2}}}{\varepsilon_{\text{crit}}^2}\left\|\boldsymbol{d}^{(k-1)}\right\|. \quad (19)$$

With the Armijo condition, Lemma 4 is applicable and $\left\|\sigma^{(k-1)}\boldsymbol{d}^{(k-1)}\right\|^{\frac{1}{2}}$ vanishes. There must be $\mathtt{r} \in (0,1)$ and $k_0 \in \mathbb{N}_0$ with

$$\left\|\boldsymbol{d}^{(k)}\right\| \leq \mathtt{C}_{\boldsymbol{\delta}} + \mathtt{r}\left\|\boldsymbol{d}^{(k-1)}\right\| \forall k \geq k_0. \quad \text{{eqn:astoopSnick}} \quad (20)$$

Repeated application of (20) leads to a geometric sum:

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

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

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

□

4.3 Cone-Projection PRP Scheme

To ensure sufficient decrease, Cheng [1] simply project the residual term in a standard two-term PRP scheme onto the null space of $\boldsymbol{\nabla} f(\boldsymbol{x}^{(k)})$.

With multiple gradients, there usually is no single orthogonal space for all of them. However, the (convex) cone of non-ascent directions

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

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

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

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

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

where

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

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

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

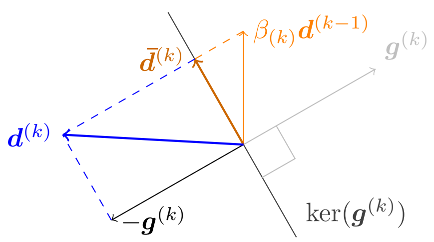

Figure 1: Projection Scheme in the single-objective setting. The standard residual term is projected onto the plane orthogonal to the gradient, so that the fincal CG direction (blue) has sufficient decrease by construction.

To go to the multi-objective setting, use the PRP coefficient as in [6]. (If we used a coefficient similar to (16) we would again need Assumption 3 in the convergence proof.)

$$\beta_{(k)} = \frac{\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k)}) - \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})}{-\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k-1)})}, \quad k \geq 1. \qquad \texttt{\{eqn:tarnishesBandeaux\}} \tag{21}$$

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

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

{eqn:prpMOProj}

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

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

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

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

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

Theorem 14. *The directions in* (PRP MOII) *have the sufficient decrease property* (2) *with* $\kappa_{\text{sd}} = 1$.

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

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

□

Theorem 15. *Suppose Assumptions 1 and 2 hold and that the criticality* $\left\|\boldsymbol{\delta}^{(k)}\right\|$ *is bounded below like in* (10)*. Then the Algorithm with modified Armijo-stepsizes according to* (4) *and with directions defined by* (PRP MOII) *generates a critical subsequence.*

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

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

Let $k \geq 1$. We find that

$$\left\|\boldsymbol{d}^{(k)}\right\| = \left\|\boldsymbol{\delta}^{(k)} + \bar{\boldsymbol{d}}^{(k)}\right\| \leq \left\|\boldsymbol{\delta}^{(k)}\right\| + \left\|\bar{\boldsymbol{d}}^{(k)}\right\| \leq \left\|\boldsymbol{\delta}^{(k)}\right\| + |\beta_{(k)}|\left\|\boldsymbol{d}^{(k-1)}\right\|. \tag{22}$$

{eqn:stanchionsBotanizing}

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

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

where the existence of $\mathtt{C}_C > 0$ follows from the compactness of $\mathtt{C}_C$. We find the same bound for the case $\mathfrak{f}_{\boldsymbol{x}}^{(k-1)}(\boldsymbol{\delta}^{(k)}) < \mathfrak{f}_{\boldsymbol{x}}^{(k)}(\boldsymbol{\delta}^{(k)})$. Looking at the definition (21), we see that there must be a constant $\mathtt{C}_\beta > 0$ with

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

Combining this with (22) results in

$$\begin{aligned}\frac{\left\|\boldsymbol{d}^{(k)}\right\|}{\left\|\boldsymbol{\delta}^{(k)}\right\|^2} &\leq \frac{1}{\left\|\boldsymbol{\delta}^{(k)}\right\|} + \frac{\mathtt{C}_\beta\sigma_{(k-1)}\left\|\boldsymbol{d}^{(k-1)}\right\|}{\left\|\boldsymbol{\delta}^{(k)}\right\|}\frac{\left\|\boldsymbol{d}^{(k-1)}\right\|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2} \\ &\overset{(10)}{\leq} \frac{1}{\varepsilon_{\text{crit}}} + \frac{\mathtt{C}_\beta\sigma_{(k-1)}\left\|\boldsymbol{d}^{(k-1)}\right\|}{\varepsilon_{\text{crit}}}\frac{\left\|\boldsymbol{d}^{(k-1)}\right\|}{\left\|\boldsymbol{\delta}^{(k-1)}\right\|^2}.\end{aligned} \tag{23}$$

{eqn:pardiSubvocalisation}

Due to Lemma 4, the step-length goes to zero and there is a k_0 such that $\frac{\mathtt{c}_\beta \sigma_{(k-1)} \left\| \boldsymbol{d}^{(k-1)} \right\|}{\varepsilon_{\mathrm{crit}}} < \mathtt{r}$ for some $r \in (0,1)$ and all $k > k_0$. Analogous to the proof of Theorem 13, we can recurse (23) to deduce that $\frac{\left\| \boldsymbol{d}^{(k)} \right\|}{\left\| \boldsymbol{\delta}^{(k)} \right\|^2}$ is uniformly bounded above for all $k \in \mathbb{N}_0$, e.g., by $\mathtt{A} > 0$. Then

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

in contradiction to Theorem 5! Existence of a critical sequence follows by Corollary 6. □

Fix broken bibtex entries...

References

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.