Mathematical and Computational Applications

MDPI

Article

Derivative-Free Multiobjective Trust Region Descent Method Using Radial Basis Function Surrogate Models

Manuel Berkemeier [1,*] and Sebastian Peitz [2]

[1] Chair of Applied Mathematics, Faculty for Computer Science, Electrical Engineering and Mathematics, Paderborn University, Warburger Str. 100, 33098 Paderborn, Germany

[2] Department of Computer Science, Faculty for Computer Science, Electrical Engineering and Mathematics, Paderborn University, Warburger Str. 100, 33098 Paderborn, Germany; sebastian.peitz@upb.de

* Correspondence: manuelbb@math.upb.de

Abstract: We present a local trust region descent algorithm for unconstrained and convexly constrained multiobjective optimization problems. It is targeted at heterogeneous and expensive problems, i.e., problems that have at least one objective function that is computationally expensive. Convergence to a Pareto critical point is proven. The method is derivative-free in the sense that derivative information need not be available for the expensive objectives. Instead, a multiobjective trust region approach is used that works similarly to its well-known scalar counterparts and complements multiobjective line-search algorithms. Local surrogate models constructed from evaluation data of the true objective functions are employed to compute possible descent directions. In contrast to existing multiobjective trust region algorithms, these surrogates are not polynomial but carefully constructed radial basis function networks. This has the important advantage that the number of data points needed per iteration scales linearly with the decision space dimension. The local models qualify as *fully linear* and the corresponding general scalar framework is adapted for problems with multiple objectives.

Keywords: multiobjective optimization; trust region methods; multiobjective descent; derivative-free optimization; radial basis functions; fully linear models

check for updates

Citation: Berkemeier, M.; Peitz, S. Derivative-Free Multiobjective Trust Region Descent Method Using Radial Basis Function Surrogate Models. *Math. Comput. Appl.* **2021**, *26*, 31. https://doi.org/10.3390/mca26020031

Academic Editors: Marcela Quiroz, Juan Gabriel Ruiz, Luis Gerardo de la Fraga and Oliver Schütze

Received: 26 February 2021
Accepted: 7 April 2021
Published: 15 April 2021

1. Introduction

Optimization problems arise in a multitude of applications in mathematics, computer science, engineering and the natural sciences. In many real-life scenarios, there are multiple, equally important objectives that need to be optimized. Such problems are then called *Multiobjective Optimization Problems* (MOP). In contrast to the single objective case, an MOP often does not have a single solution but an entire set of optimal trade-offs between the different objectives, which we call *Pareto optimal*. They constitute the *Pareto Set* and their image is the *Pareto Frontier*. The goal in the numerical treatment of an MOP is to either approximate these sets or to find single points within these sets. In applications, the problem can become more difficult when some of the objectives require computationally expensive or time consuming evaluations. For instance, the objectives could depend on a computer simulation or some other *black-box*. It is then of primary interest to reduce the overall number of function evaluations. Consequently, it can become infeasible to approximate derivative information of the true objectives using, e.g., finite differences. This holds true especially if higher order derivatives are required. In this work, optimization methods that do not use the true objective gradients (which nonetheless are assumed to exist) are referred to as *derivative-free*.

There is a variety of methods to deal with MOPs, some of which are also derivative-free or try to constrain the number of expensive function evaluations. A broad overview of different problems and techniques concerning multiobjective optimization can be found, e.g., in [1–4]. One popular approach for calculating Pareto optimal solutions is scalarization, i.e., the transformation of an MOP into a single objective problem, cf. [5] for an overview. Alternatively, classical (single objective) descent algorithms can be adapted for the multiobjective case [6–11]. What is more, the structure of the Pareto Set can be exploited to find multiple solutions [12,13]. There are also methods for non-smooth problems [14,15] and multiobjective direct-search variants [16,17]. Both scalarization and descent techniques may be included in Evolutionary Algorithms (EA) [18–22]. To address computationally expensive objectives or missing derivative information, there are algorithms that use surrogate models (see the surveys [23–25]) or borrow from ideas from scalar trust region methods, e.g., [26].

In single objective optimization, trust region methods are well suited for derivative-free optimization [27,28]. Our work is based on the recent development of multiobjective trust region methods:

- In [29], a trust region method using Newton steps for functions with positive definite Hessians on an open domain is proposed.
- In [30], quadratic Taylor polynomials are used to compute the steepest descent direction which is used in a backtracking manner to find solutions for unconstrained problems.
- In [31], polynomial regression models are used to solve an augmented MOP based on the scalarization in [17]. The algorithm is designed unconstrained bi-objective problems, but the general idea has been formulated for more objectives in [32].
- In [33], quadratic Lagrange polynomials are used and the Pascoletti–Serafini scalarization is employed for the descent step calculation.

Our contribution is the extension of the above-mentioned methods to general fully linear models (and in particular Radial Basis Function (RBF) surrogates as in [34]), which is related to the scalar framework in [35]. Most importantly, this reduces surrogate construction complexity, in terms of objective evaluations per iteration, to linear with respect to the number of decision variables, in contrast to the quadratically increasing number of function evaluations for methods using second degree polynomials. We further prove convergence to critical points when the problem is constrained to a convex and compact set by using an analogous argumentation as in [36]. To this end, we extend the theory in [6] to provide new results concerning the continuity of the solutions of the projected steepest descent direction problem, which is based on the alternative formulation by Fliege and Svaiter [7]. We also show how to keep the convergence properties for constrained problems when the Pascoletti–Serafini scalarization is employed (like in [33]).

The remainder of the paper is structured as follows: Section 2 provides a brief introduction to multiobjective optimality and criticality concepts. In Section 3 the fundamentals of the algorithm are explained. In Section 4 we introduce fully linear surrogate models and describe the construction of suitable polynomial models and RBF models for unconstrained and box-constrained problems. We also formalize the main algorithm in this section. Section 5 deals with the descent step calculation so that a sufficient decrease is achieved in each iteration. Convergence is proven in Section 6 and a few numerical examples for unconstrained and finitely box-constrained problems are shown in Section 7. In Section 7 we also compare the RBF models against linear polynomial models that have the same linear construction complexity. We conclude with a brief discussion in Section 8.

2. Optimality and Criticality in Multiobjective Optimization

We consider the following (real-valued) multiobjective optimization problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbf{f}(\mathbf{x}) := \min_{\mathbf{x}\in\mathcal{X}} \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_k(\mathbf{x}) \end{bmatrix} \in \mathbb{R}^k, \tag{MOP}$$

with a feasible set $\mathcal{X} \subseteq \mathbb{R}^n$ and k objective functions $f_\ell \colon \mathbb{R}^n \to \mathbb{R}$, $\ell = 1, \dots, k$. We further assume (MOP) to be *heterogeneous*. That is, there is a non-empty subset $I_{\text{ex}} \subseteq \{1, \dots, k\}$ of indices so that the gradients of $f_\ell, \ell \in I_{\text{ex}}$, are unknown and cannot be approximated, e.g., via finite differences. The (possibly empty) index set $I_{\text{cheap}} = \{1, \dots, k\} \setminus I_{\text{ex}}$ indicates functions whose gradients are available.

Solutions for (MOP) consist of optimal trade-offs $\mathbf{x}^* \in \mathcal{X}$ between the different objectives and are called non-dominated or Pareto optimal. That is, there is no $\mathbf{x} \in \mathcal{X}$ with $\mathbf{f}(\mathbf{x}) \prec \mathbf{f}(\mathbf{x}^*)$ (i.e., $\mathbf{f}(\mathbf{x}) \leq \mathbf{f}(\mathbf{x}^*)$ and $f_\ell(\mathbf{x}) < f_\ell(\mathbf{x}^*)$ for some index $\ell \in \{1, \dots, k\}$). The subset $\mathcal{P}_{\text{S}} \subseteq \mathcal{X}$ of non-dominated points is then called the *Pareto Set* and its image $\mathcal{P}_{\text{F}} := \mathbf{f}(\mathcal{P}_{\text{S}}) \subseteq \mathbb{R}^k$ is called the *Pareto Frontier*. All concepts can be defined in a local fashion in an analogous way.

Similar to scalar optimization, there is a necessary condition for local optima using the gradients of the objective function. We therefore implicitly assume all objective functions $f_\ell, \ell = 1, \dots, k$, to be continuously differentiable on $\mathcal{X}$. Moreover, the following assumption allows for an easier treatment of tangent cones in the constrained case:

Assumption 1. *Either the problem is unconstrained, i.e.,* $\mathcal{X} = \mathbb{R}^n$ *or the feasible set* $\mathcal{X} \subseteq \mathbb{R}^n$ *is compact and convex. All functions are defined on* $\mathcal{X}$.

The second case is a standard assumption in the MO literature for constrained problems [6,7]. Now let $\boldsymbol{\nabla} f_\ell(\mathbf{x})$ denote the gradient of f_ℓ and $\mathbf{Df}(\mathbf{x}) \in \mathbb{R}^{k \times n}$ the Jacobian of $\mathbf{f}$ at $\mathbf{x} \in \mathcal{X}$.

Definition 1. *We call a vector* $\mathbf{d} \in \mathcal{X} - \mathbf{x}$ *a multi-descent direction for* $\mathbf{f}$ *in* $\mathbf{x}$ *if* $\langle \boldsymbol{\nabla} f_\ell(\mathbf{x}), \mathbf{d} \rangle < 0$ *for all* $\ell \in \{1, \dots, k\}$, *or equivalently if*

$$\max_{\ell=1,\dots,k} \langle \boldsymbol{\nabla} f_\ell(\mathbf{x}^*), \mathbf{d} \rangle < 0 \tag{1}$$

where $\langle \bullet, \bullet \rangle$ *is the standard inner product on* $\mathbb{R}^n$ *and we consider* $\mathcal{X} - \mathbf{x} = \mathcal{X}$ *in the unconstrained case* $\mathcal{X} = \mathbb{R}^n$.

A point $\mathbf{x}^* \in \mathcal{X}$ is called *critical* for (MOP) iff there is no descent direction $\mathbf{d} \in \mathcal{X} - \mathbf{x}^*$ with (1). As all Pareto optimal points are also critical (cf. [6,37] or [2] [Ch. 17]), it is viable to search for optimal points by calculating points from the superset $\mathcal{P}_{\text{crit}} \supseteq \mathcal{P}_{\text{S}}$ of critical points for (MOP). Similar to single objective optimization, using such a first order condition makes sense especially in combination with some global method or when exploring the structure of the critical set. We discuss promising approaches in Section 8. Note, that due the above restrictions, our method is not a general replacement for other methods, e.g., scalarization approaches, but rather an additional tool for situations where those are not applicable.

One intuitive way to approach the critical set is by iteratively performing descent steps. Fliege and Svaiter [7] propose several ways to compute suitable descent directions. The minimizer $\mathbf{d}^*$ of the following problem is known as the multiobjective steepest-descent direction.

$$\min_{\mathbf{d} \in \mathcal{X} - \mathbf{x}} \max_{\ell=1,\dots,k} \langle \boldsymbol{\nabla} f_\ell(\mathbf{x}), \mathbf{d} \rangle \quad \text{s.t.} \quad \|\mathbf{d}\| \leq 1. \tag{P1}$$

Problem (P1) has an equivalent reformulation as

$$\min_{\mathbf{d} \in \mathcal{X} - \mathbf{x}} \beta \qquad \text{s.t.} \quad \|\mathbf{d}\| \leq 1 \quad \text{and} \quad \langle \boldsymbol{\nabla} f_\ell(\mathbf{x}), \mathbf{d} \rangle \leq \beta \; \forall \, \ell = 1, \dots, k, \tag{P2}$$

which is a linear program, if $\mathcal{X}$ is defined by linear constraints and the maximum-norm $\|\bullet\| = \|\bullet\|_\infty$ is used [7]. We thus stick with this choice because it facilitates implementation, but note that other choices are possible (see for example [33]).

Motivated by the next theorem we can use the optimal value of either problem as a measure of criticality, i.e., as a multiobjective pendant for the gradient norm. As is standard in most multiobjective trust region works (cf. [29,30,33]), we flip the sign so that the values are non-negative.

Theorem 1. *For $\mathbf{x} \in \mathcal{X}$ let $\mathbf{d}^*(\mathbf{x})$ be the minimizer of (P1) and $\omega(\mathbf{x})$ be the negative optimal value, that is*

$$\omega(\mathbf{x}) := -\max_{\ell=1,\dots,k} \langle \nabla f_\ell(\mathbf{x}), \mathbf{d}^*(\mathbf{x}) \rangle.$$

Then the following statements hold:

1. *$\omega(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathcal{X}$.*
2. *The function $\omega \colon \mathbb{R}^n \to \mathbb{R}$ is continuous.*
3. *The following statements are equivalent:*
   - *(a) The point $\mathbf{x} \in \mathcal{X}$ is* not *critical.*
   - *(b) $\omega(\mathbf{x}) > 0$.*
   - *(c) $\mathbf{d}^*(\mathbf{x}) \neq \mathbf{0}$.*

   Consequently, the point $\mathbf{x}$ is critical iff $\omega(\mathbf{x}) = 0$.

Proof. For the unconstrained case all statements are proven in [7] (Lemma 3).

The first and the third statement hold true for $\mathcal{X}$ convex and compact by definition. The continuity of ω can be shown similarly as in [6], see Appendix A.1. □

With further conditions on $\mathbf{f}$ and $\mathcal{X}$ the criticality measure $\omega(\mathbf{x})$ is even Lipschitz continuous and subsequently uniformly and Cauchy continuous:

Theorem 2. *If $\nabla f_\ell, \ell = 1, \dots, k$, are Lipschitz continuous and Assumption 1 holds, then the map $\omega(\bullet)$ as defined in Theorem 1 is uniformly continuous.*

Proof. The proof for $\mathcal{X} = \mathbb{R}^n$ is given by Thomann [38]. A proof for the constrained case can be found in Appendix A.1 as to not clutter this introductory section. □

Together with Theorem 1 this hints at $\omega(\bullet)$ being a criticality measure as defined for scalar trust region methods in [36] ([Ch. 8]):

Definition 2. *We call $\pi \colon \mathbb{N}_0 \times \mathbb{R}^n \to \mathbb{R}$, a criticality measure for (MOP) if π is Cauchy continuous with respect to its second argument and if*

$$\lim_{t\to\infty} \pi(t, \mathbf{x}^{(t)}) = 0$$

implies that the sequence $\left\{\mathbf{x}^{(t)}\right\}$ asymptotically approaches a Pareto critical point.

3. Trust Region Ideas

Multiobjective trust region algorithms closely follow the design of scalar approaches (see [36] for an extensive treatment) and provide an alternative to (approximate) linesearch algorithms (e.g., [7]). Consequently, the requirements and convergence proofs in [29,30,33] for the unconstrained multiobjective case are fairly similar to those in [36]. We will reexamine the core concepts to provide a clear understanding and point out the similarities to the single objective case.

The main idea is to iteratively compute multi-descent steps $\mathbf{s}^{(t)}$ in every iteration $t \in \mathbb{N}_0$. We could, for example, use the steepest descent direction given by (P1). This would require knowledge of the true objective gradients, which need not be available for objective functions with indices in I_{ex}. Hence, benevolent surrogate model functions

$$\mathbf{m}^{(t)} \colon \mathbb{R}^n \to \mathbb{R}^k,\ \mathbf{x} \mapsto \mathbf{m}^{(t)}(\mathbf{x}) = \left[m_1^{(t)}(\mathbf{x}), \dots, m_k^{(t)}(\mathbf{x})\right]^T,$$

are employed (at least for the expensive objectives).

The surrogate models are constructed to be sufficiently accurate within a trust region

$$B^{(t)} := B\left(\mathbf{x}^{(t)}; \Delta^{(t)}\right) = \left\{\mathbf{x} \in \mathcal{X} : \left\|\mathbf{x} - \mathbf{x}^{(t)}\right\| \le \Delta^{(t)}\right\}, \quad \text{with } \|\bullet\| = \|\bullet\|_\infty, \tag{2}$$

around the current iterate $\mathbf{x}^{(t)}$. To be precise, the models are made fully linear as described in Section 4.1. This ensures that the model error and the model gradient error are uniformly bounded within the trust region.

The *model* steepest descent direction $\mathbf{d}_{\mathrm{m}}^{(t)}$ can then computed as the optimizer of the surrogate problem

$$\begin{aligned} \omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) &:= -\min_{\mathbf{d}\in\mathcal{X}-\mathbf{x}} \beta \\ &\text{s.t. } \|\mathbf{d}\| \le 1, \text{ and } \langle \boldsymbol{\nabla} m_\ell^{(t)}(\mathbf{x}), \mathbf{d}\rangle \le \beta \quad \forall \ell = 1, \dots, k. \end{aligned} \tag{Pm}$$

Now let $\sigma^{(t)} > 0$ be a step size. The direction $\mathbf{d}_{\mathrm{m}}^{(t)}$ need not be a descent direction for the true objectives $\mathbf{f}$ and the trial point $\mathbf{x}_+^{(t)} = \mathbf{x}^{(t)} + \sigma^{(t)}\mathbf{d}_{\mathrm{m}}^{(t)}$ is only accepted if a measure $\rho^{(t)}$ of improvement and model quality surpasses a positive threshold ν_+. As in [30,33], we scalarize the multiobjective problems by defining

$$\Phi(\mathbf{x}) := \max_{\ell=1,\dots,k} f_\ell(\mathbf{x}), \qquad \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}) := \max_{\ell=1,\dots,k} m_\ell^{(t)}(\mathbf{x}).$$

Whenever $\Phi(\mathbf{x}^{(t)}) - \Phi(\mathbf{x}_+^{(t)}) > 0$, there is a reduction in at least one objective function of $\mathbf{f}$ because of

$$0 < \Phi(\mathbf{x}^{(t)}) - \Phi(\mathbf{x}_+^{(t)}) = f_\ell(\mathbf{x}^{(t)}) - f_q(\mathbf{x}_+^{(t)}) \overset{\text{df.}}{\le} f_\ell(\mathbf{x}^{(t)}) - f_\ell(\mathbf{x}_+^{(t)}),$$

where we denoted by ℓ the (not necessarily unique) maximizing index in $\Phi(\mathbf{x}^{(t)})$ and by q the (neither necessarily unique) maximizing index in $\Phi(\mathbf{x}_+^{(t)})$. (The abbreviation "df." above the inequality symbol stands for "(by) definition" and is used throughout this document when appropriate.) Of course, the same property holds for $\Phi_{\mathrm{m}}^{(t)}(\bullet)$ and $\mathbf{m}^{(t)}$.

Thus, the step size $\sigma^{(t)} > 0$ is chosen so that the step $\mathbf{s}^{(t)} = \sigma^{(t)}\mathbf{d}_{\mathrm{m}}^{(t)}$ satisfies both $\mathbf{x}^{(t)} + \mathbf{s}^{(t)} \in B^{(t)}$ and a "sufficient decrease condition" of the form

$$\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)} + \mathbf{s}^{(t)}) \ge \kappa^{\mathrm{sd}}\omega\left(\mathbf{x}^{(t)}\right) \min\left\{\mathrm{C} \cdot \omega\left(\mathbf{x}^{(t)}\right), 1, \Delta^{(t)}\right\} \ge 0,$$

with constants $\kappa^{\mathrm{sd}} \in (0,1)$ and $\mathrm{C} > 0$, see Section 5. Such a condition is also required in the scalar case [35,36] and essential for the convergence proof in Section 6, where we show $\lim_{t\to\infty} \omega\left(\mathbf{x}^{(t)}\right) = 0$.

Due to the decrease condition, the denominator in the ratio of actual versus predicted reduction

$$\rho^{(t)} = \begin{cases} \dfrac{\Phi(\mathbf{x}^{(t)}) - \Phi(\mathbf{x}_+^{(t)})}{\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_+^{(t)})} & \text{if } \mathbf{x}^{(t)} \ne \mathbf{x}_+^{(t)}, \\ 0 & \text{if } \mathbf{x}^{(t)} = \mathbf{x}_+^{(t)} \Leftrightarrow \mathbf{s}^{(t)} = \mathbf{0}, \end{cases} \tag{3}$$

is non-negative. A positive $\rho^{(t)}$ implies a decrease in at least one objective f_ℓ, so we accept $\mathbf{x}_+^{(t)}$ as the next iterate if $\rho^{(t)} > \nu_+ > 0$. If $\rho^{(t)}$ is sufficiently large, say $\rho^{(t)} \ge \nu_{++} > \nu_+ > 0$, the next trust region might have a larger radius $\Delta^{(t+1)} \ge \Delta^{(t)}$. If in contrast $\rho < \nu_{++}$, the next trust region radius should be smaller and the surrogates improved.

This encompasses the case $\mathbf{s}^{(t)} = \mathbf{0}$, when the iterate $\mathbf{x}^{(t)}$ is critical for

$$\min_{\mathbf{x}\in B^{(t)}} \mathbf{m}^{(t)}(\mathbf{x}) \in \mathbb{R}^k. \tag{MOPm}$$

Roughly speaking, we suppose that $\mathbf{x}^{(t)}$ is near a critical point for the original problem (MOP) if $\mathbf{m}^{(t)}$ is sufficiently accurate. If we truly are near a critical point, then the trust region radius will approach 0. For further details concerning the acceptance ratio $\rho^{(t)}$, see [33] (Section 2.2).

Remark 1. *We can modify $\rho^{(t)}$ in (3) to obtain a descent in all objectives, i.e., if $\mathbf{x}^{(t)} \neq \mathbf{x}_+^{(t)}$ we test $\rho^{(t)} = \dfrac{f_\ell(\mathbf{x}^{(t)}) - f_\ell(\mathbf{x}_+^{(t)})}{m_\ell^{(t)}(\mathbf{x}^{(t)}) - m_\ell^{(t)}(\mathbf{x}_+^{(t)})} > \nu_+$ for all $\ell = 1, \ldots, k$. This is the* strict *acceptance test.*

4. Surrogate Models and the Final Algorithm

Until now, we have not discussed the actual choice of surrogate models used for $\mathbf{m}^{(t)}$. As is shown in Section 5, the models should be twice continuously differentiable with uniformly bounded hessians. To prove convergence of our algorithm, we have to impose further requirements on the (uniform) approximation qualities of the surrogates $\mathbf{m}^{(t)}$. We can meet these requirements using so-called fully linear models. Moreover, fully linear models intrinsically allow for modifications of the basic trust region method that are aimed at reducing the total number of expensive objective evaluations. Finally, we briefly recapitulate how radial basis functions and multivariate Lagrange polynomials can be made fully linear.

Remark 2. *Although the trust region framework is suitable for general convexly constrained compact sets, we will discuss the construction of fully linear polynomial and RBF models for unconstrained and box-constrained problems only.*

In the constrained case, we treat the constraints as unrelaxable, *that is, we do not allow for evaluations of the true objectives outside $\mathcal{X}$, see the definition of $B^{(t)} \subseteq \mathcal{X}$ in (2). We also ensure to only select training data in $\mathcal{X}$ during the construction of surrogate models.*

To the best of our knowledge there are no construction procedures for the above model types for general (unrelaxable) constraints. A discussion of how some model based algorithms deal with constraints can be found in [28] (Section 7). The issue is also addressed in [27] (Ch. 13) . If the constraints are treated as relaxable, then techniques from [39] (Ch. 15) might be applicable such as merit functions or filter methods, but this is left for future research.

4.1. Fully Linear Models

We start by reciting the abstract definition of full linearity as given in [27,35]:

Definition 3. *Let $\Delta^{\text{ub}} > 0$ be given and let $f\colon \mathbb{R} \to \mathbb{R}$ be a function that is continuously differentiable in an open domain containing $\mathcal{X}$ and has a Lipschitz continuous gradient on $\mathcal{X}$. A set of model functions $\mathcal{M} = \{m\colon \mathbb{R}^n \to \mathbb{R}\} \subseteq C^1(\mathbb{R}^n, \mathbb{R})$ is called a* fully linear *class of models w.r.t. f if the following hold:*

1. *There are positive constants $\epsilon, \dot\epsilon$ and L_m such that for any given $\Delta \in (0, \Delta^{\text{ub}})$ and for any $\mathbf{x} \in \mathcal{X}$ there is a model function $m \in \mathcal{M}$ with Lipschitz continuous gradient and corresponding Lipschitz constant bounded by L_m and such that*
    - *the error between the gradient of the model and the gradient of the function satisfies*

    $$\|\nabla f(\boldsymbol{\xi}) - \nabla m(\boldsymbol{\xi})\| \le \dot\epsilon\Delta, \quad \forall \boldsymbol{\xi} \in B(\mathbf{x}; \Delta),$$

    - *the error between the model and the function satisfies*

    $$|f(\boldsymbol{\xi}) - m(\boldsymbol{\xi})| \le \epsilon\Delta^2, \quad \forall \boldsymbol{\xi} \in B(\mathbf{x}; \Delta).$$

2. *For this class $\mathcal{M}$ there exists "model-improvement" algorithm that, in a finite, uniformly bounded (w.r.t.* $\mathbf{x}$ *and* Δ*) number of steps, can:*
    - *either establish that a given model $m \in \mathcal{M}$ is fully linear on $B(\mathbf{x}; \Delta)$, i.e., it satisfies the error bounds in 1,*
    - *or find a model $\tilde{m}$ that is fully linear on $B(\mathbf{x}; \Delta)$.*

Remark 3. *In the unconstrained case, the requirements in Definition 3 can be relaxed a bit, at least when using the strict acceptance test with $\mathbf{f}(\mathbf{x}^{(T)}) \le \mathbf{f}(\mathbf{x}^{(t)})$ for all $T \ge t \ge 0$. We can then restrict ourselves to the set*

$$\mathcal{X}' := \bigcup_{\mathbf{x} \in L(\mathbf{x}^{(0)})} B\left(\mathbf{x}; \Delta^{\mathrm{ub}}\right), \quad \text{where } L(\mathbf{x}^{(0)}) := \left\{ \mathbf{x} \in \mathbb{R}^n : \mathbf{f}(\mathbf{x}) \le \mathbf{f}(\mathbf{x}^{(0)}) \right\}.$$

For the convergence analysis in Section 6, we further cite [27] ([Lemma 10.25]). The lemma states that a fully linear model is also fully linear in enlarged regions if the error constants are chosen appropriately:

Lemma 1. *For $\mathbf{x} \in \mathcal{X}$ and $\Delta \le \Delta^{\mathrm{ub}}$ consider a function f and a fully-linear model m as in Definition 3 with constants $\epsilon, \dot{\epsilon}, L_m > 0$. Let $L_f > 0$ be a Lipschitz constant of $\boldsymbol{\nabla} f$.*
Assume w.l.o.g. that

$$L_m + L_f \le \epsilon \quad \text{and} \quad \frac{\dot{\epsilon}}{2} \le \epsilon.$$

Then m is fully linear on $B(\mathbf{x}; \tilde{\Delta})$ for any $\tilde{\Delta} \in [\Delta, \Delta^{\mathrm{ub}}]$ with respect to the same constants $\epsilon, \dot{\epsilon}, L_m$.

Finally, we generalize the definition to a *vector* of real valued functions.

Definition 4. *Let $\Delta^{\mathrm{ub}} > 0$ be given and let $\mathbf{f} = [f_1, \dots, f_k]^T$ be a vector of functions satisfying the requirements of Definition 3. Then $\mathbf{m} = [m_1, \dots, m_k]^T$, with $m_\ell \colon \mathbb{R}^n \to \mathbb{R}, \ell \in \{1, \dots, k\}$, belongs to a collection of fully linear classes w.r.t. $\mathbf{f}$ if for each ℓ the function m_ℓ belongs to a fully linear class w.r.t. f_ℓ, with error constants ϵ_ℓ and $\dot{\epsilon}_\ell$.*

The model-improvement algorithm of $\mathbf{m}$ consists in applying the individual improvement algorithms for all indices $\ell \in \{1, \dots, k\}$ and $\mathbf{m}$ is deemed fully linear iff all m_ℓ are fully linear with constants ϵ_ℓ and $\dot{\epsilon}_\ell$.

Definition 4 is stated in a way that allows for different model types for the different objectives. Most importantly, we can use $m_\ell = f_\ell$ and $\nabla m_\ell = \nabla f_\ell$ if the objective is cheap, i.e., $\ell \in I_{\text{cheap}}$, and if f_ℓ not only has Lipschitz gradients but also has a Hessian that is uniformly bounded in terms of its norm. The latter requirement is formalized in Assumption 3 and needed for the convergence analysis.

Algorithm Modifications

With Definitions 3 and 4 we have formalized our assumption that the surrogates become more accurate when we decrease the trust region radius. This motivates the following modifications to the basic procedure:

- "Relaxing" the (finite) surrogate construction process to try for a possible descent even if the surrogates are not fully linear.
- A criticality test depending on $\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)$. If this value is very small at the current iterate, then $\mathbf{x}^{(t)}$ could lie near a Pareto critical point. With the criticality test and Algorithm 1 we ensure that the next model is fully linear and the trust region is not too large. This allows for a more accurate criticality measure and descent step calculation.

- A trust region update that also takes into consideration $\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)$. The radius should be enlarged if we have a large acceptance ratio $\rho^{(t)}$ and the $\Delta^{(t)}$ is small as measured against $\beta\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)$ for a constant $\beta > 0$.

These changes are implemented in Algorithm 2. For more detailed explanations we refer to [27] (Ch. 10).

```
Algorithm 1: Criticality Routine.
    Configuration: A backtracking constant $\alpha \in (0,1)$, $\mu > 0$ from Algorithm 2;
    Input: Current trust region radius $\Delta^{(t)}$, current models $\mathbf{m}^{(t)}$;
    Output: Fully linear models $\mathbf{m}^{(t)}$ and the (possibly shrunken) radius $\Delta^{(t)}$;
    Set $\Delta_0 \leftarrow \Delta^{(t)}$;
    for $j = 1, 2, \ldots$ do
        Set radius: $\Delta^{(t)} \leftarrow \alpha^{j-1}\Delta_0$;
        Make models $\mathbf{m}^{(t)}$ fully linear on $B^{(t)}$ ;        /* can change $\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)$ */
        if $\Delta^{(t)} \leq \mu\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)$ then
            Break;
        end
    end
```

From Algorithm 2 we see that we can classify the iterations based on $\rho^{(t)}$ as in Definition 5.

Definition 5. *For given constants $0 \leq \nu_{+} \leq \nu_{++} < 1, \nu_{++} \neq 0$, we call the iteration with index $t \in \mathbb{N}_0$ of Algorithm 2.*

- . . . successful *if $\rho^{(t)} \geq \nu_{++}$. The set of successful indices is $\mathcal{S} = \{t \in \mathbb{N}_0 : \rho^{(t)} \geq \nu_{++}\} \subseteq \mathbb{N}_0$. The trial point is accepted and the trust region radius can be increased.*
- . . . model-improving *if $\rho^{(t)} < \nu_{++}$ and the models $\mathbf{m}^{(t)} = [m_1^{(t)}, \ldots, m_k^{(t)}]^T$ are not fully linear. In these iterations the trial point is rejected and the trust region radius is not changed.*
- . . . acceptable *if $\nu_{++} > \rho^{(t)} \geq \nu_{+}$ and the models $\mathbf{m}^{(t)}$ are fully linear. If $\nu_{++} = \nu_{+} \in (0,1)$, then there are no acceptable indices. The trial point is accepted but the trust region radius is decreased.*
- . . . inacceptable *otherwise, i.e., if $\rho^{(t)} < \nu_{++}$ and $\mathbf{m}^{(t)}$ are fully linear. The trial point is rejected and the radius decreased.*

Algorithm 2: General Trust Region Method (TRM) for (MOP).

```
Configuration: Criticality parameters $\varepsilon_{\text{crit}} > 0$ and $\mu > \beta > 0$, acceptance
               parameters $1 > \nu_{++} \ge \nu_{+} \ge 0, \nu_{++} \ne 0$, update factors
               $\gamma_{\uparrow} \ge 1 > \gamma_{\downarrow} \ge \gamma_{\Downarrow} > 0$ and $\Delta^{\text{ub}} > 0$;
Input: The initial site $\mathbf{x}^{(0)} \in \mathbb{R}^n$;
for $t = 0, 1, \ldots$ do
    if $t > 0$ and iteration $(t-1)$ was model-improving (cf. Definition 5) then
        Perform at least one improvement step on $\mathbf{m}^{(t-1)}$ and then let
          $\mathbf{m}^{(t)} \leftarrow \mathbf{m}^{(t-1)}$;
    else
        Construct surrogate models $\mathbf{m}^{(t)}$ on $B^{(t)}$;
    end
    /* Criticality Step:                                                    */
    if $\varpi_{\text{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) < \varepsilon_{\text{crit}}$ and ( $\mathbf{m}^{(t)}$ not fully linear or $\Delta^{(t)} > \mu\varpi_{\text{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)$ ) then
        Set $\Delta_{*}^{(t)} \leftarrow \Delta^{(t)}$;
        Call Algorithm 1 so that $\mathbf{m}^{(t)}$ is fully linear on $B^{(t)}$ with
          $\Delta^{(t)} \in \left(0, \mu\varpi_{\text{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)\right]$;
        Then set $\Delta^{(t)} \leftarrow \min\left\{\max\left\{\Delta^{(t)}, \beta\varpi_{\text{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)\right\}, \Delta_{*}^{(t)}\right\}$ ;
    end
    Compute a suitable descent step $\mathbf{s}^{(t)}$;
    Set $\mathbf{x}_{+}^{(t)} \leftarrow \mathbf{x}^{(t)} + \mathbf{s}^{(t)}$, evaluate $\mathbf{f}(\mathbf{x}_{+}^{(t)})$ and compute $\rho^{(t)}$ with (3);
    Perform the following updates:

      $\mathbf{x}^{(t+1)} \leftarrow \begin{cases} \mathbf{x}^{(t)} & \text{if } \rho^{(t)} < \nu_{+} \text{ or } \nu_{+} \le \rho^{(t)} < \nu_{++} \text{ \& } \mathbf{m}^{(t)} \text{ is \textbf{not} fully linear,} \\ \mathbf{x}_{+}^{(t)} & \text{if } \rho^{(t)} \ge \nu_{++} \text{ or } \nu_{+} \le \rho^{(t)} < \nu_{++} \text{ \& } \mathbf{m}^{(t)} \text{ is fully linear,} \end{cases}$
      $\Delta^{(t+1)} \leftarrow \Delta_{+}$, where
        $\Delta_{+} \begin{cases} = \Delta^{(t)} & \text{if } \rho^{(t)} < \nu_{++} \text{ \& } \mathbf{m}^{(t)} \text{ is \textbf{not} fully linear,} \\ \in [\gamma_{\Downarrow}\Delta^{(t)}, \gamma_{\downarrow}\Delta^{(t)}] & \text{if } \rho^{(t)} < \nu_{++} \text{ \& } \mathbf{m}^{(t)} \text{ is fully linear,} \\ \in \left[\Delta^{(t)}, \min\{\gamma_{\uparrow}\Delta^{(t)}, \Delta^{\text{ub}}\}\right] & \text{if } \nu_{++} \le \rho^{(t)} \text{ and } \Delta^{(t)} \ge \beta\omega_{\text{m}}^{(t)}\left(\mathbf{x}^{(t)}\right), \\ = \min\{\gamma_{\uparrow}\Delta^{(t)}, \Delta^{\text{ub}}\} & \text{if } \nu_{++} \le \rho^{(t)} \text{ and } \Delta^{(t)} < \beta\omega_{\text{m}}^{(t)}\left(\mathbf{x}^{(t)}\right). \end{cases}$
end
```

4.2. Fully Linear Lagrange Polynomials

Quadratic Taylor polynomial models are used very frequently. As explained in [27] we can alternatively use multivariate interpolating Lagrange polynomial models when derivative information is not available. We will consider first and second degree Lagrange models. Even though the latter require $\mathcal{O}(n^2)$ function evaluations they are still cheaper than second degree finite difference models. For this reason, these models are also used in [33,38].

To construct an interpolating polynomial model we have to provide p data sites, where p is the dimension of the space Π_n^d of real-valued n-variate polynomials with degree d. For $d = 1$ we have $p = n + 1$ and for $d = 2$ it is $p = \dfrac{(n+1)(n+2)}{2}$. If $n \ge 2$, the *Mairhuber–Curtis* theorem [40] applies and the data sites must form a so-called *poised* set in $\mathcal{X}$. The set $\Xi = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_p\} \subset \mathbb{R}^n$ is poised if for any basis $\{\psi_i\}_i$ of Π_n^d the matrix $\mathbf{M}_\psi := \left[\psi_i(\boldsymbol{\xi}_j)\right]_{1 \le i,j \le p}$ is non-singular. Then for any function $f \colon \mathbb{R}^n \to \mathbb{R}$ there is a unique interpolating polynomial $m(\mathbf{x}) = \sum_{i=1}^{p} \lambda_i \psi_i(\mathbf{x})$ with $m(\boldsymbol{\xi}_j) = f(\boldsymbol{\xi}_j)$ for all $j = 1, \ldots, p$.

Given a poised set Ξ the associated Lagrange basis $\{l_i\}_i$ of Π_n^d is defined by $l_i(\boldsymbol{\xi}_{\mathbf{j}}) = \delta_{i,j}$. The model coefficients then simply are the data values, i.e., $\lambda_i = f(\boldsymbol{\xi}_i)$.

Same as in [38], we implement Algorithm 6.2 from [27] to ensure poisedness. It selects training sites Ξ from the current (slightly enlarged) trust region of radius $\theta_1 \Delta^{(t)}, \theta_1 \geq 1$, and calculates the associated lagrange basis. We can then separately evaluate the true objectives f_ℓ on Ξ to easily build the surrogates $m_\ell^{(t)}$, $\ell \in \{1, \ldots, k\}$. Our implementation always includes $\boldsymbol{\xi}_1 = \mathbf{x}^{(t)}$ and tries to select points from a database of prior evaluations first.

We employ an additional algorithm (Algorithm 6.3 in [27]) to ensure that the set Ξ is even Λ-*poised*, see [27] ([Definition 3.6]). The procedure is still finite and ensures the models are actually *fully linear*. The quality of the surrogate models can be improved by choosing a small algorithm parameter $\Lambda > 1$. Our implementation tries again to recycle points from a database. Different to before, interpolation at $\mathbf{x}^{(t)}$ can no longer be guaranteed. This second step can also be omitted first and then used as a model-improvement step in a subsequent iteration.

4.3. Fully Linear Radial Basis Function Models

The main drawback of quadratic Lagrange models is that we still need $\mathcal{O}(n^2)$ function evaluations in each iteration of Algorithm 2. A possible fix is to use under-determined regression polynomials instead [27,31,41]. Motivated by the findings in [34] we chose so-called Radial Basis Function (RBF) models as an alternative. RBF are well-known for their approximation capabilities on irregular data [40]. In our implementation they have the form

$$m(\mathbf{x}) = \sum_{i=1}^{N} c_i \varphi(\|\mathbf{x} - \boldsymbol{\xi}_i\|_2) + \pi(\mathbf{x}), \quad \text{with } \pi = \sum_{j=1}^{n+1} \lambda_j \psi_j \in \Pi_n^1 \text{ and } N \geq n+1, \tag{4}$$

which conforms to the construction by Wild et al. [34]. Here, φ is a function from a domain containing $\mathbb{R}_{\geq 0}$ to $\mathbb{R}$. For a fixed φ the mapping $\varphi(\|\bullet\|)$ from $\mathbb{R}^n \to \mathbb{R}$ is radially symmetric with respect to its argument and the mapping $(\mathbf{x}, \boldsymbol{\xi}) \mapsto \varphi(\|\mathbf{x} - \boldsymbol{\xi}\|_2)$ is called a *kernel*.

We will describe the procedure only briefly and refer to [34,42] and the dissertation [41] for more details. To conform to the algorithmic framework the models must have Hessians of uniformly bounded norm. Additionally, we want them to be twice differentiable due to the following, very general result:

Theorem 3 (Th 4.1 in [41])**.** *Suppose that f and m are continuously differentiable in an open domain containing $B^{(t)}$ and that ∇f and ∇m are Lipschitz in $B^{(t)}$. Further suppose that m interpolates f on a Λ-poised set $\Xi = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{n+1}\}$ (for a fixed $\Lambda < \infty$). Then m is fully linear for f as in Definition 3.*

The Λ-poised set is determined using pivotal algorithms from [34,41] in an enlarged trust region of radius $\theta_1 \Delta^{(t)}, \theta_1 \geq 1$. If we restrict ourselves to functions φ that are conditionally positive definite (c.p.d.—see [34] for the definition) of order $D \leq 2$, then for any $f \colon \mathbb{R}^n \to \mathbb{R}$ an interpolating model m of form (4) is uniquely determined by solving a linear equation system. If further φ is either twice continuously differentiable on an open domain containing $[0, \infty)$ with $\varphi'(0) = 0$, then m from (4) is twice continuously differentiable and has Lipschitz gradients exactly if its Hessian stays bounded. This is the case for all φ we consider (see Table 1). The hessian norm is determined by the magnitudes of the coefficients c_i and by $|\varphi'(r)/r|$ and $|\varphi''(r)|$.

Table 1. Some radial functions $\varphi\colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ that are c.p.d. of order $D \leq 2$, cf. [34].

Name	$\varphi(r)$	c.p.d. order D
Cubic	r^3	2
Multiquadric	$-\sqrt{1+(\alpha r)^2}, \alpha > 0$	1
Gaussian	$\exp(-(\alpha r)^2), \alpha > 0$	0

If there are exactly $N = n + 1$ points from a poised set Ξ, then the coefficients c_i vanish and the model (4) is a linear polynomial. The values $|\varphi'(r)/r|$ and $|\varphi''(r)|$ are bounded because of $r \in [0, \Delta^{\text{ub}}]$ and $\varphi'(0) = 0$. To exploit the nonlinear modeling capabilities of RBF and perform exploration, there is a procedure in [34] to select additional (database) points from within a region of maximum radius $\theta_2\Delta^{\text{ub}}$, $\theta_2 \geq \theta_1 \geq 1$, so that the values $|c_i|$ stay bounded. Modifications for box constraints can be found in [41] ([Sec. 6.3.1]) and [43].

Table 1 shows the RBF we are using and of which order they are. Both the Gaussian and the Multiquadric allow for fine-tuning with a shape parameter $\alpha > 0$. This can potentially improve the conditioning of the interpolation system.

Figure 1b illustrates the effect of the shape parameter. As can be seen, the radial functions become narrower for larger shape parameters. Hence, we do not only use a constant shape parameter $\alpha = 1$ like [34] do, but we also use an α that is (within lower and upper bounds) inversely proportional to $\Delta^{(t)}$.

Figure 1a shows interpolation of a nonlinear function by a surrogate based on the Multiquadric with a linear tail.

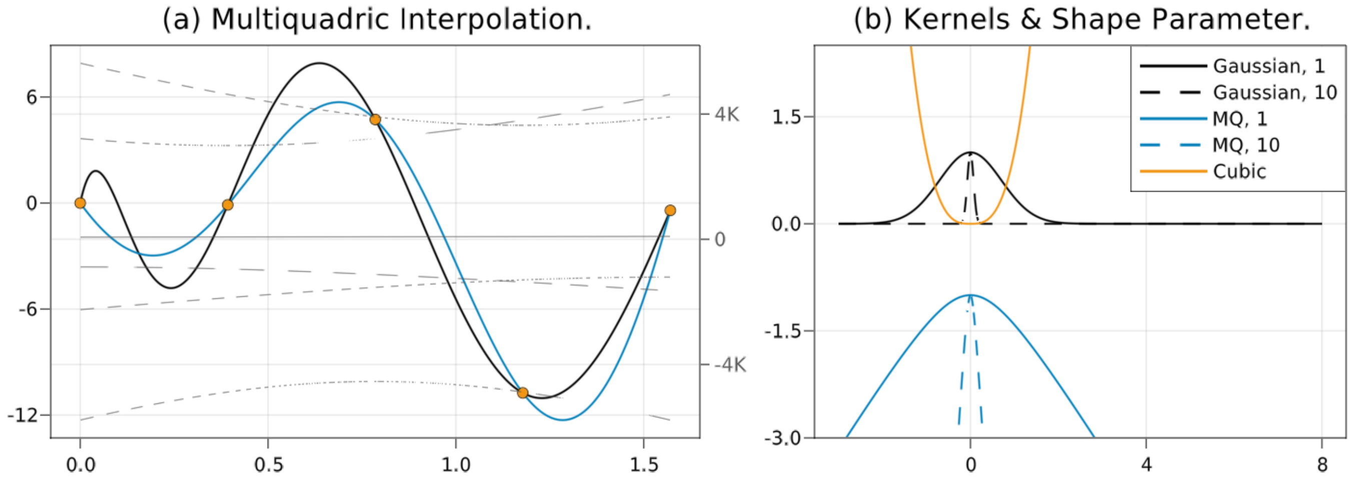

Figure 1. (**a**) Interpolation of a nonlinear function (black) by a Multiquadric surrogate (blue) based on 5 discrete training points (orange). Dashed lines show the kernels and the polynomial tail. (**b**) Different kernels in 1D with varying shape parameter (1 or 10), see also Table 1.

5. Descent Steps

In this section we introduce some possible steps $\mathbf{s}^{(t)}$ to use in Algorithm 2. We begin by defining the best step along the steepest descent direction as given by (Pm). Subsequently, backtracking variants are defined that use a multiobjective variant of Armijo's rule.

5.1. Pareto–Cauchy Step

Both the *Pareto–Cauchy point* as well as a backtracking variant, the *modified Pareto–Cauchy point*, are points along the descent direction $\mathbf{d}_{\text{m}}^{(t)}$ within $B^{(t)}$ so that a sufficient decrease measured by $\Phi_{\text{m}}^{(t)}(\bullet)$ and $\omega_{\text{m}}^{(t)}(\bullet)$ is achieved. Under mild assumptions we can then derive a decrease in terms of $\omega(\bullet)$.

Definition 6. *For $t \in \mathbb{N}_0$ let $\mathbf{d}_{\mathrm{m}}^{(t)}$ be a minimizer for* (Pm)*. The best attainable trial point $\mathbf{x}_{\mathrm{PC}}^{(t)}$ along $\mathbf{d}_{\mathrm{m}}^{(t)}$ is called the* Pareto–Cauchy point *and given by*

$$\begin{aligned} \mathbf{x}_{\mathrm{PC}}^{(t)} &:= \mathbf{x}^{(t)} + \sigma^{(t)} \cdot \mathbf{d}_{\mathrm{m}}^{(t)}, \\ \sigma^{(t)} &= \arg\min_{0 \le \sigma} \Phi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)} + \sigma \cdot \mathbf{d}_{\mathrm{m}}^{(t)}\right) \quad \text{s.t. } \mathbf{x}_{\mathrm{PC}}^{(t)} \in B^{(t)}. \end{aligned} \tag{5}$$

Let $\sigma^{(t)}$ be the minimizer in (5)*. We call $\mathbf{s}_{\mathrm{PC}}^{(t)} := \sigma^{(t)} \mathbf{d}_{\mathrm{m}}^{(t)}$ the* Pareto–Cauchy step.

If we make the following standard assumption, then the Pareto–Cauchy point allows for a lower bound on the improvement in terms of $\Phi_{\mathrm{m}}^{(t)}$.

Assumption 2. *For all $t \in \mathbb{N}_0$ the surrogates $\mathbf{m}^{(t)}(\mathbf{x}) = [m_1^{(t)}(\mathbf{x}), \dots, m_k^{(t)}(\mathbf{x})]^T$ are twice continuously differentiable on an open set containing $\mathcal{X}$. Denote by $\boldsymbol{H} m_\ell^{(t)}(\mathbf{x})$ the Hessian of $m_\ell^{(t)}$ for $\ell = 1, \dots, k$.*

Theorem 4. *If Assumptions 1 and 2 are satisfied, then for any iterate $\mathbf{x}^{(t)}$ the Pareto–Cauchy point $\mathbf{x}_{\mathrm{PC}}^{(t)}$ satisfies*

$$\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_{\mathrm{PC}}^{(t)}) \ge \frac{1}{2} \omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \cdot \min\left\{ \frac{\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)}{\mathrm{c} H_{\mathrm{m}}^{(t)}}, \Delta^{(t)}, 1 \right\}, \tag{6}$$

where

$$H_{\mathrm{m}}^{(t)} = \max_{\ell = 1, \dots, k} \max_{\mathbf{x} \in B^{(t)}} \left\| \boldsymbol{H} m_\ell^{(t)}(\mathbf{x}) \right\|_F \tag{7}$$

and the constant $\mathrm{c} > 0$ relates the trust region norm $\|\bullet\|$ to the Euclidean norm $\|\bullet\|_2$ via

$$\|\mathbf{x}\|_2 \le \sqrt{\mathrm{c}} \|\mathbf{x}\| \qquad \forall \mathbf{x} \in \mathbb{R}^n. \tag{8}$$

If $\|\bullet\| = \|\bullet\|_\infty$ is used, then c can be chosen as $\mathrm{c} = k$. The proof for Theorem 4 is provided after the next auxiliary lemma.

Lemma 2. *Under Assumptions 1 and 2, let $\mathbf{d}$ be a non-increasing direction at $\mathbf{x}^{(t)} \in \mathbb{R}^n$ for $\mathbf{m}^{(t)}$, i.e.,*

$$\left\langle \boldsymbol{\nabla} m_\ell^{(t)}(\mathbf{x}^{(t)}), \mathbf{d} \right\rangle \le 0 \qquad \forall \ell = 1, \dots, k.$$

Let $q \in \{1, \dots, k\}$ be any objective index and $\bar{\sigma} \ge \min\left\{\Delta^{(t)}, \|\mathbf{d}\|\right\}$. Then it holds that

$$m_q^{(t)}(\mathbf{x}^{(t)}) - \min_{0 < \sigma < \bar{\sigma}} m_q^{(t)}\left(\mathbf{x}^{(t)} + \sigma \frac{\mathbf{d}}{\|\mathbf{d}\|}\right) \ge \frac{w}{2} \min\left\{ \frac{w}{\|\mathbf{d}\|^2 \mathrm{c} H_{\mathrm{m}}^{(t)}}, \frac{\Delta^{(t)}}{\|\mathbf{d}\|}, 1 \right\},$$

where we have used the shorthand notation

$$w = -\max_{\ell = 1, \dots, k} \left\langle \boldsymbol{\nabla} m_\ell^{(t)}(\mathbf{x}^{(t)}), \mathbf{d} \right\rangle \ge 0.$$

Lemma 2 states that a minimizer along any non-increasing direction $\mathbf{d}$ achieves a minimum reduction w.r.t. $\Phi_{\mathrm{m}}^{(t)}$. Similar results can be found in in [30] or [33]. But since we do not use polynomial surrogates $\mathbf{m}^{(t)}$, we have to employ the multivariate version of Taylor's theorem to make the proof work. We can do this because according to Assumption 2, the functions $m_q^{(t)}, q \in \{1, \dots, k\}$ are twice continuously differentiable in an open domain containing $\mathcal{X}$. Moreover, Assumption 1 ensures that the function is defined on the line from χ to $\mathbf{x}$. As shown in [44] (Ch. 3) a first degree expansion at $\mathbf{x} \in B(\chi, \Delta)$ around $\chi \in \mathcal{X}$ then leads to

$$m_q^{(t)}(\mathbf{x}) = m_q(\chi) + \nabla m_q^{(t)}(\chi)^T \mathbf{h} + \frac{1}{2}\mathbf{h}^T \boldsymbol{H} m_q^{(t)}(\xi_q)\mathbf{h}, \qquad \text{with } \mathbf{h} = (\mathbf{x} - \chi),$$
$$\text{for some } \xi_q \in \{\mathbf{x} + \theta(\chi - \mathbf{x}) : \theta \in [0,1]\}, \text{ for all } q = 1, \dots, k. \tag{9}$$

Proof of Lemma 2. Let the requirements of Lemma 2 hold and let $\mathbf{d}$ be a non-increasing direction for $\mathbf{m}^{(t)}$. Then:

$$m_q^{(t)}(\mathbf{x}^{(t)}) - \min_{0<\sigma<\bar{\sigma}} m_q^{(t)}\left(\mathbf{x}^{(t)} + \sigma\frac{\mathbf{d}}{\|\mathbf{d}\|}\right) = \max_{0\leq\sigma\leq\bar{\sigma}}\left\{m_q^{(t)}(\mathbf{x}^{(t)}) - m_q^{(t)}\left(\mathbf{x}^{(t)} + \sigma\frac{\mathbf{d}}{\|\mathbf{d}\|}\right)\right\}$$
$$\overset{(9)}{=} \max_{0\leq\sigma\leq\bar{\sigma}}\left\{m_q^{(t)}(\mathbf{x}^{(t)}) - \left(m_q^{(t)}(\mathbf{x}^{(t)}) + \frac{\sigma}{\|\mathbf{d}\|}\langle\nabla m_q^{(t)}(\mathbf{x}^{(t)}), \mathbf{d}\rangle + \frac{\sigma^2}{2\|\mathbf{d}\|^2}\langle\mathbf{d}, \boldsymbol{H} m_q^{(t)}(\xi_q)\mathbf{d}\rangle\right)\right\}$$
$$\geq \max_{0\leq\sigma\leq\bar{\sigma}}\left\{-\frac{\sigma}{\|\mathbf{d}\|}\max_{j=1,\dots,k}\langle\nabla m_j^{(t)}(\mathbf{x}^{(t)}), \mathbf{d}\rangle - \frac{\sigma^2}{2\|\mathbf{d}\|^2}\langle\mathbf{d}, \boldsymbol{H} m_q^{(t)}(\xi_q)\mathbf{d}\rangle\right\}.$$

We use the shorthand $w = -\max_j\langle\nabla m_j^{(t)}(\mathbf{x}^{(t)}), \mathbf{d}\rangle$ and the Cauchy–Schwartz inequality to get

$$\dots \geq \max_{0\leq\sigma\leq\bar{\sigma}}\left\{\frac{\sigma}{\|\mathbf{d}\|}w - \frac{\sigma^2}{2\|\mathbf{d}\|^2}\|\mathbf{d}\|_2^2\left\|\boldsymbol{H} m_q^{(t)}(\xi)\right\|_F\right\} \overset{(8),(7)}{\geq} \max_{0\leq\sigma\leq\bar{\sigma}}\left\{\frac{\sigma}{\|\mathbf{d}\|}w - \frac{\sigma^2}{2}\mathrm{c}H_{\mathrm{m}}^{(t)}\right\}.$$

The RHS is concave and we can thus easily determine the global maximizer σ^*. Similar to [30] (Lemma 4.1) we find

$$m_q^{(t)}(\mathbf{x}^{(t)}) - \min_{0<\sigma<\bar{\sigma}} m_q^{(t)}\left(\mathbf{x}^{(t)} + \sigma\frac{\mathbf{d}}{\|\mathbf{d}\|}\right) \geq \frac{w}{2}\min\left\{\frac{w}{\|\mathbf{d}\|^2\mathrm{c}H_{\mathrm{m}}^{(t)}}, \frac{\Delta^{(t)}}{\|\mathbf{d}\|}, 1\right\},$$

where we have additionally used $\bar{\sigma} \geq \min\{\Delta^{(t)}, 1\}$. □

Proof of Theorem 4. If $\mathbf{x}^{(t)}$ is Pareto critical for (MOPm), then $\mathbf{d}_{\mathrm{m}}^{(t)} = \mathbf{0}$ and $\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) = 0$ and the inequality holds trivially.

Else, let the indices $\ell, q \in \{1, \dots, k\}$ be such that

$$\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_{\mathrm{PC}}^{(t)}) = m_\ell^{(t)}(\mathbf{x}^{(t)}) - m_q^{(t)}(\mathbf{x}_{\mathrm{PC}}^{(t)}) \geq m_q(\mathbf{x}^{(t)}) - m_q(\mathbf{x}_{\mathrm{PC}}^{(t)})$$

and define

$$\bar{\sigma} := \begin{cases}\min\left\{\Delta^{(t)}, \left\|\mathbf{d}_{\mathrm{m}}^{(t)}\right\|\right\} & \text{if } \left\|\mathbf{d}_{\mathrm{m}}^{(t)}\right\| < 1 \text{ or } \Delta^{(t)} \leq 1,\\ \Delta^{(t)} & \text{else.}\end{cases} \tag{10}$$

Then clearly $\bar{\sigma} \geq \min\left\{\Delta^{(t)}, \left\|\mathbf{d}_{\mathrm{m}}^{(t)}\right\|\right\}$ and for the Pareto–Cauchy point we have

$$m_q^{(t)}\left(\mathbf{x}_{\mathrm{PC}}^{(t)}\right) = \min_{0\leq\sigma\leq\bar{\sigma}} m_q\left(\mathbf{x}^{(t)} + \frac{\sigma}{\left\|\mathbf{d}_{\mathrm{m}}^{(t)}\right\|}\mathbf{d}_{\mathrm{m}}^{(t)}\right).$$

From Lemma 2 and $\left\|\mathbf{d}_{\mathrm{m}}^{(t)}\right\|$ the bound (6) immediately follows. □

Remark 4. *Some authors define the Pareto–Cauchy point as the actual minimizer* $\mathbf{x}_{\min}^{(t)}$ *of* $\Phi_{\mathrm{m}}^{(t)}$ *within the current trust region (instead of the minimizer along the steepest descent direction). For this true minimizer the same bound* (6) *holds. This is due to*

$$\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_{\min}^{(t)}) = m_\ell(\mathbf{x}^{(t)}) - \min_{\mathbf{x}\in B^{(t)}} m_q(\mathbf{x}) \geq m_q(\mathbf{x}^{(t)}) - m_q(\mathbf{x}_{\mathrm{PC}}^{(t)}).$$

5.2. Modified Pareto–Cauchy Point via Backtracking

A common approach in trust region methods is to find an approximate solution to (5) within the current trust region. Usually a backtracking procedure similar to Armijo's inexact line-search is used for the Pareto–Cauchy subproblem, see [36] (Section 6.3) and [30]. Doing so, we can still guarantee a sufficient decrease.

Before we actually define the backtracking step along $\mathbf{d}_{\mathrm{m}}^{(t)}$, we derive a more general lemma. It illustrates that backtracking along any suitable direction is well-defined.

Lemma 3. *Suppose Assumptions 1 and 2 hold. For* $\mathbf{x}^{(t)} \in \mathbb{R}^n$, *let* $\mathbf{d}$ *be a descent direction for* $\mathbf{m}^{(t)}$ *and let* $q \in \{1,\ldots,k\}$ *be any objective index and* $\bar\sigma > 0$. *Then, for any fixed constants* $\mathrm{a},\mathrm{b} \in (0,1)$ *there is an integer* $j \in \mathbb{N}_0$ *such that*

$$\Psi\left(\mathbf{x}^{(t)} + \frac{\mathrm{b}^j\bar\sigma}{\|\mathbf{d}\|}\mathbf{d}\right) \leq \Psi(\mathbf{x}^{(t)}) - \frac{\mathrm{a}\mathrm{b}^j\bar\sigma}{\|\mathbf{d}\|}w \tag{11}$$

where, again, we have used the shorthand notation $w = -\max_{\ell=1,\ldots,k}\left\langle \boldsymbol{\nabla} m_\ell^{(t)}(\mathbf{x}^{(t)}), \mathbf{d}\right\rangle > 0$ *and* Ψ *is either some specific model,* $\Psi = m_\ell$, *or the maximum value,* $\Psi = \Phi_{\mathrm{m}}^{(t)}$.
Moreover, if we define the step $\mathbf{s}^{(t)} = \frac{\mathrm{b}^j\bar\sigma}{\|\mathbf{d}\|}\mathbf{d}$ *for the* ***smallest*** $j \in \mathbb{N}_0$ *satisfying* (11), *then there is a constant* $\kappa_{\mathrm{m}}^{\mathrm{sd}} \in (0,1)$ *such that*

$$\Psi(\mathbf{x}^{(t)}) - \Psi\left(\mathbf{x}^{(t)} + \mathbf{s}^{(t)}\right) \geq \kappa_{\mathrm{m}}^{\mathrm{sd}} w \min\left\{\frac{w}{\|\mathbf{d}\|^2 \mathrm{c} H_{\mathrm{m}}^{(t)}}, \frac{\bar\sigma}{\|\mathbf{d}\|}\right\}. \tag{12}$$

Proof. The first part can be derived from the fact that $\mathbf{d}$ is a descent direction, see e.g., [6]. However, we will use the approach from [30] to also derive the bound (12). With Taylor's Theorem we obtain

$$\begin{aligned}
&\Psi\left(\mathbf{x}^{(t)} + \frac{\mathrm{b}^j\bar\sigma}{\|\mathbf{d}\|}\mathbf{d}\right) = m_\ell\left(\mathbf{x}^{(t)} + \frac{\mathrm{b}^j\bar\sigma}{\|\mathbf{d}\|}\mathbf{d}\right) \qquad \text{(for some } \ell \in \{1,\ldots,k\}\text{)}\\
&= m_\ell^{(t)}(\mathbf{x}^{(t)}) + \frac{\mathrm{b}^j\bar\sigma}{\|\mathbf{d}\|}\langle \boldsymbol{\nabla} m_\ell^{(t)}(\mathbf{x}^{(t)}), \mathbf{d}\rangle + \frac{(\mathrm{b}^j\bar\sigma)^2}{2\|\mathbf{d}\|^2}\langle \mathbf{d}, \boldsymbol{H} m_\ell^{(t)}(\boldsymbol{\xi}_\ell)\mathbf{d}\rangle\\
&\leq \Psi(\mathbf{x}^{(t)}) + \max_{q=1,\ldots,k}\frac{\mathrm{b}^j\bar\sigma}{\|\mathbf{d}\|}\langle \boldsymbol{\nabla} m_q^{(t)}(\mathbf{x}^{(t)}), \mathbf{d}\rangle + \max_{q=1,\ldots,k}\frac{(\mathrm{b}^j\bar\sigma)^2}{2\|\mathbf{d}\|^2}\langle \mathbf{d}, \boldsymbol{H} m_q^{(t)}(\boldsymbol{\xi}_q)\mathbf{d}\rangle\\
&\overset{(\mathrm{Pm}),(7)}{\leq} \Psi(\mathbf{x}^{(t)}) - \frac{\mathrm{b}^j\bar\sigma}{\|\mathbf{d}\|}w + \frac{(\mathrm{b}^j\bar\sigma)^2}{2}\mathrm{c}H_{\mathrm{m}}^{(t)}.
\end{aligned} \tag{13}$$

In the last line, we have additionally used the Cauchy–Schwartz inequality. For a constructive proof, suppose now that (11) is violated for some $j \in \mathbb{N}_0$, i.e.,

$$\Psi\left(\mathbf{x}^{(t)} + \frac{\mathrm{b}^j\bar\sigma}{\|\mathbf{d}\|}\mathbf{d}\right) > \Psi(\mathbf{x}^{(t)}) - \frac{\mathrm{a}\mathrm{b}^j\bar\sigma}{\|\mathbf{d}\|}w.$$

Plugging in (13) for the LHS and substracting $\Psi(\mathbf{x}^{(t)})$ then leads to

$$\mathrm{b}^j > \frac{2(1-\mathrm{a})w}{\|\mathbf{d}\|\bar\sigma \mathrm{c} H_{\mathrm{m}}^{(t)}},$$

where the right hand side is positive and completely independent of j. Since $\mathrm{b} \in (0,1)$, there must be a $j^* \in \mathbb{N}_0, j^* > j$, for which $\mathrm{b}^{j^*} \leq \dfrac{2(1-\mathrm{a})w}{\|\mathbf{d}\|\bar{\sigma}\mathrm{c}H_{\mathrm{m}}^{(t)}}$ so that (11) must also be fulfilled for this b^{j^*}.

Analogous to the proof of [30] ([Lemma 4.2]) we can now derive the constant $\kappa_{\mathrm{m}}^{\mathrm{sd}}$ from (12) as $\kappa_{\mathrm{m}}^{\mathrm{sd}} = \min\{2\mathrm{b}(1-\mathrm{a}), \mathrm{a}\}$.
□

Lemma 3 applies naturally to the step along $\mathbf{d}_{\mathrm{m}}^{(t)}$:

Definition 7. *For* $\mathbf{x}^{(t)} \in B^{(t)}$ *let* $\mathbf{d}_{\mathrm{m}}^{(t)}$ *be a solution to* (Pm) *and define the* modified Pareto–Cauchy step *as*

$$\tilde{\mathbf{s}}_{\mathrm{PC}}^{(t)} := \mathrm{b}^j\bar{\sigma}\frac{\mathbf{d}_{\mathrm{m}}^{(t)}}{\left\|\mathbf{d}_{\mathrm{m}}^{(t)}\right\|},$$

where again $\bar{\sigma}$ *as in* (10) *and* $j \in \mathbb{N}_0$ *is the smallest integer that satisfies*

$$\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)} + \tilde{\mathbf{s}}_{\mathrm{PC}}^{(t)}) \leq \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \frac{\mathrm{a}\mathrm{b}^j\bar{\sigma}}{\left\|\mathbf{d}_{\mathrm{m}}^{(t)}\right\|}\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \tag{14}$$

for predefined constants $\mathrm{a}, \mathrm{b} \in (0,1)$.

The definition of $\bar{\sigma}$ ensures, that $\mathbf{x}^{(t)} + \tilde{\mathbf{s}}_{\mathrm{PC}}^{(t)}$ is contained in the current trust region $B^{(t)}$. Furthermore, these steps provide a sufficient decrease very similar to (6):

Corollary 1. *Suppose Assumptions 1 and 2 hold. For the step* $\tilde{\mathbf{s}}_{\mathrm{PC}}^{(t)}$ *the following statements are true:*

1. *A* $j \in \mathbb{N}_0$ *as in* (14) *exists.*
2. *There is a constant* $\kappa_{\mathrm{m}}^{\mathrm{sd}} \in (0,1)$ *such that the modified Pareto–Cauchy step* $\tilde{\mathbf{s}}_{\mathrm{PC}}^{(t)}$ *satisfies*

$$\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)} + \tilde{\mathbf{s}}_{\mathrm{PC}}^{(t)}) \geq \kappa_{\mathrm{m}}^{\mathrm{sd}}\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)\min\left\{\frac{\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)}{\mathrm{c}H_{\mathrm{m}}^{(t)}}, \Delta^{(t)}, 1\right\}.$$

Proof. If $\mathbf{x}^{(t)}$ is critical, then the bound is trivial. Otherwise, the existence of a j satisfying (14) follows from Lemma 3 for $\Psi = \Phi_{\mathrm{m}}^{(t)}$. The lower bound on the decrease follows immediately from $\bar{\sigma} \geq \min\left\{\left\|\mathbf{d}_{\mathrm{m}}^{(t)}\right\|, \Delta^{(t)}\right\}$. □

From Lemma 3 it follows that the backtracking condition (14) can be modified to explicitly require a decrease in *every* objective:

Definition 8. *Let* $j \in \mathbb{N}_0$ *the smallest integer satisfying*

$$\min_{\ell=1,\dots,k}\left\{m_{\ell}^{(t)}(\mathbf{x}^{(t)}) - m_{\ell}^{(t)}\left(\mathbf{x}^{(t)} + \mathrm{b}^j\bar{\sigma}\frac{\mathbf{d}_{\mathrm{m}}^{(t)}}{\left\|\mathbf{d}_{\mathrm{m}}^{(t)}\right\|}\right)\right\} \geq \frac{\mathrm{a}\mathrm{b}^j\bar{\sigma}}{\left\|\mathbf{d}_{\mathrm{m}}^{(t)}\right\|}\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right).$$

We define the strict *modified Pareto–Cauchy point as* $\hat{\mathbf{x}}_{\mathrm{PC}}^{(t)} = \mathbf{x}^{(t)} + \hat{\mathbf{s}}_{\mathrm{PC}}^{(t)}$ *and the corresponding step as* $\hat{\mathbf{s}}_{\mathrm{PC}}^{(t)} = \mathrm{b}^j\bar{\sigma}\dfrac{\mathbf{d}_{\mathrm{m}}^{(t)}}{\left\|\mathbf{d}_{\mathrm{m}}^{(t)}\right\|}$.

Corollary 2. *Suppose Assumptions 1 and 2 hold.*

1. *The strict modified Pareto–Cauchy point exists, the backtracking is finite.*
2. *There is a constant* $\kappa_{\mathrm{m}}^{\mathrm{sd}} \in (0,1)$ *such that*

$$\min_{\ell=1,\dots,k}\left\{ m_\ell^{(t)}(\mathbf{x}^{(t)}) - m_\ell^{(t)}\left(\hat{\mathbf{x}}_{\mathrm{PC}}^{(t)}\right)\right\} \geq \kappa_{\mathrm{m}}^{\mathrm{sd}} \omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \min\left\{ \frac{\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)}{\mathrm{c}H_{\mathrm{m}}^{(t)}}, \Delta^{(t)}, 1\right\}. \tag{15}$$

Remark 5. *In the preceding subsections, we have shown descent steps along the model steepest descent direction. Similar to the single objective case we do not necessarily have to use the steepest descent direction and different step calculation methods are viable. For instance, Thomann and Eichfelder [33] use the well-known Pascoletti–Serafini scalarization to treat the subproblem* (MOPm)*. We refer to their work and Appendix B to see how this method can be related to the steepest descent direction.*

5.3. Sufficient Decrease for the Original Problem

In the previous subsections, we have shown how to compute steps $\mathbf{s}^{(t)}$ to achieve a sufficient decrease in terms of $\Phi_{\mathrm{m}}^{(t)}$ and $\omega_{\mathrm{m}}^{(t)}(\bullet)$. For a descent step $\mathbf{s}^{(t)}$ the bound is of the form

$$\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)} + \mathbf{s}^{(t)}) \geq \kappa_{\mathrm{m}}^{\mathrm{sd}} \omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \min\left\{ \frac{\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)}{\mathrm{c}H_{\mathrm{m}}^{(t)}}, \Delta^{(t)}, 1\right\}, \quad \kappa_{\mathrm{m}}^{\mathrm{sd}} \in (0,1), \tag{16}$$

and thereby very similar to the bounds for the scalar projected gradient trust region method [36]. By introducing a slightly modified version of $\omega_{\mathrm{m}}^{(t)}(\bullet)$, we can transform (16) into the bound used in [30,33].

Lemma 4. *If* $\pi(t, \mathbf{x}^{(t)})$ *is a criticality measure for some multiobjective problem, then* $\tilde{\pi}(t, \mathbf{x}^{(t)}) = \min\left\{1, \pi(t, \mathbf{x}^{(t)})\right\}$ *is also a criticality measure for the same problem.*

Proof. We have $0 \leq \tilde{\pi}(t, \mathbf{x}^{(t)}) \leq \pi(t, \mathbf{x}^{(t)})$. Thus, $\tilde{\pi} \to 0$ whenever $\pi \to 0$. The minimum of uniformly continuous functions is again uniformly continuous. □

We next make another standard assumption on the class of surrogate models.

Assumption 3. *The norm of all model hessians is uniformly bounded above on* $\mathcal{X}$*, i.e., there is a positive constant* H_{m} *such that*

$$\left\| \boldsymbol{H} m_\ell^{(t)}(\mathbf{x}) \right\|_F \leq \mathrm{H}_{\mathrm{m}} \qquad \forall \ell = 1, \dots, k, \forall \mathbf{x} \in B^{(t)}, \ \forall t \in \mathbb{N}_0.$$

W.l.o.g., we assume

$$\mathrm{H}_{\mathrm{m}} \cdot \mathrm{c} > 1, \quad \textit{with } \mathrm{c} \textit{ as in } (8). \tag{17}$$

Remark 6. *From this assumption it follows that the model gradients are then Lipschitz as well. Together with Theorem 2, we then know that* $\omega_{\mathrm{m}}^{(t)}(\bullet)$ *is a criticality measure for* (MOPm)*.*

Motivated by the previous remark, we will from now on refer to the following functions

$$\varpi(\mathbf{x}) := \min\{\omega(\mathbf{x}), 1\} \text{ and } \varpi_{\mathrm{m}}^{(t)}(\mathbf{x}) := \min\left\{\omega_{\mathrm{m}}^{(t)}(\mathbf{x}), 1\right\} \forall t = 0, 1, \dots \tag{18}$$

We can thereby derive the sufficient decrease condition in "standard form":

Corollary 3. *Under Assumption 3, suppose that for* $\mathbf{x}^{(t)}$ *and some descent step* $\mathbf{s}^{(t)}$ *the bound (16) holds. For the criticality measure* $\varpi_{\mathrm{m}}^{(t)}(\bullet)$ *it follows that*

$$\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)} + \mathbf{s}^{(t)}) \geq \kappa_{\mathrm{m}}^{\mathrm{sd}} \varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \min\left\{ \frac{\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)}{\mathrm{cH_m}}, \Delta^{(t)} \right\}. \tag{19}$$

Proof. $\varpi_{\mathrm{m}}^{(t)}(\bullet)$ is a criticality measure due to Assumption 3 and Lemma 4. Further, from (18) and (17) it follows that

$$\frac{\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)}{\mathrm{cH_m}} \leq \frac{1}{\mathrm{cH_m}} \leq 1$$

and if we plug this into (16) we obtain (19). □

To relate the RHS of (19) to the criticality $\omega(\bullet)$ of the original problem, we require another assumption.

Assumption 4. *There is a constant* $\kappa_\omega > 0$ *such that*

$$\left| \omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \omega\left(\mathbf{x}^{(t)}\right) \right| \leq \kappa_\omega \omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right).$$

This assumption is also made by Thomann and Eichfelder [33] and can easily be justified by using fully linear surrogate models and a bounded trust region radius in combination with a criticality test, see Lemma 7.

Assumption 4 can be used to formulate the next two lemmata relating the model criticality and the true criticality. They are proven in Appendix A.2. From these lemmata and Corollary 3 the final result, Corollary 4, easily follows.

Lemma 5. *If Assumption 4 holds, then it holds for* $\varpi_{\mathrm{m}}^{(t)}(\bullet)$ *and* $\varpi(\bullet)$ *from (18) that*

$$\left| \varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \varpi\left(\mathbf{x}^{(t)}\right) \right| \leq \kappa_\omega \varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right).$$

Lemma 6. *From Assumption 4 it follows that*

$$\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \geq \frac{1}{\kappa_\omega + 1} \varpi\left(\mathbf{x}^{(t)}\right) \quad \textit{with } (\kappa_\omega + 1)^{-1} \in (0, 1).$$

Corollary 4. *Suppose that Assumptions 3 and 4 hold and that* $\mathbf{x}^{(t)}$ *and* $\mathbf{s}^{(t)}$ *satisfy (19). Then*

$$\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)} + \mathbf{s}^{(t)}) \geq \kappa^{\mathrm{sd}} \varpi\left(\mathbf{x}^{(t)}\right) \min\left\{ \frac{\varpi\left(\mathbf{x}^{(t)}\right)}{\mathrm{cH_m}}, \Delta^{(t)} \right\}, \tag{20}$$

where $\kappa^{\mathrm{sd}} = \frac{\kappa_{\mathrm{m}}^{\mathrm{sd}}}{1+\kappa_\omega} \in (0, 1)$.

6. Convergence

6.1. Preliminary Assumptions and Definitions

To prove convergence of Algorithm 2 we first have to make sure that at least one of the objectives is bounded from below. This is a weaker requirement than the standard assumption that all objectives are bounded from below:

Assumption 5. *The maximum* $\max_{\ell=1,\dots,k} f_\ell(\mathbf{x})$ *of all objective functions is bounded from below on* $\mathcal{X}$.

To be able to use $\varpi(\bullet)$ as a criticality measure and to refer to fully linear models, we further require:

Assumption 6. *The objective $\mathbf{f}\colon \mathbb{R}^n \to \mathbb{R}^k$ is continuously differentiable in an open domain containing $\mathcal{X}$ and has a Lipschitz continuous gradient on $\mathcal{X}$.*

We summarize the assumptions on the surrogates as follows:

Assumption 7. *The vector of surrogate model functions $m_1^{(t)}, \dots, m_k^{(t)}$ belongs to a collection of fully linear classes as in Definition 4: For each objective objective index $\ell = 1, \dots, k$ there are error constants ϵ_ℓ so that $\dot{\epsilon}_\ell$ and $m_\ell^{(t)}$ can be made to satisfy the bounds in Definition 3.*

For the subsequent analysis we define component-wise maximum constants as

$$\epsilon := \max_{\ell=1,\dots,k} \epsilon_\ell, \quad \dot{\epsilon} := \max_{\ell=1,\dots,k} \dot{\epsilon}_\ell. \tag{21}$$

We also wish for the descent steps to fulfill a sufficient decrease condition for the surrogate criticality measure as discussed in Section 5.

Assumption 8. *For all $t \in \mathbb{N}_0$ the descent steps $\mathbf{s}^{(t)}$ are assumed to fulfill both $\mathbf{x}^{(t)} + \mathbf{s}^{(t)} \in B^{(t)}$ and (19).*

Finally, to avoid a cluttered notation when dealing with subsequences we define the following shorthand notations:

$$\varpi_{\mathrm{m}}^{(t)} := \varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right), \; \varpi^{(t)} := \varpi\left(\mathbf{x}^{(t)}\right) \quad \forall t \in \mathbb{N}_0.$$

6.2. Convergence Proof

In the following we prove convergence of Algorithm 2 to Pareto critical points. We account for the case that no criticality test is used, i.e., $\varepsilon_{\text{crit}} = 0$. We then require all surrogates to be fully linear in each iteration and need Assumption 4. The proof is an adapted version of the scalar case in [35].

It is also similar to the proofs for the multiobjective algorithms in [30,33]. However, in both cases, no criticality test is employed, there is no distinction between successful and acceptable iterations ($\nu_+ = \nu_{++}$) and interpolation at $\mathbf{x}^{(t)}$ by the surrogates is required. We indicate notable differences when appropriate.

We start with two results concerning the criticality test in Algorithm 2.

Lemma 7. *For each iteration $t \in \mathbb{N}_0$ Assumption 4 is fulfilled if the model $\mathbf{m}^{(t)}$ is fully-linear and the criticality test was performed and—if applicable—Algorithm 1 has finished.*

Proof. Let $\ell, q \in \{1, \dots, k\}$ and $\mathbf{d}_\ell, \mathbf{d}_q \in \mathcal{X} - \mathbf{x}^{(t)}$ be solutions of (P1) and (Pm), respectively, such that

$$\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) = -\langle \nabla m_\ell^{(t)}(\mathbf{x}^{(t)}), \mathbf{d}_\ell \rangle, \; \omega\left(\mathbf{x}^{(t)}\right) = -\langle \nabla f_q(\mathbf{x}^{(t)}), \mathbf{d}_q \rangle.$$

If $\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \geq \omega\left(\mathbf{x}^{(t)}\right)$, then, using Cauchy–Schwartz and $\|\mathbf{d}_\ell\| \leq 1$,

$$\begin{aligned}\left|\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \omega\left(\mathbf{x}^{(t)}\right)\right| &= \langle \nabla f_q(\mathbf{x}^{(t)}), \mathbf{d}_q \rangle - \langle \nabla m_\ell^{(t)}(\mathbf{x}^{(t)}), \mathbf{d}_\ell \rangle \\ &\overset{\text{df.}}{\leq} \langle \nabla f_q(\mathbf{x}^{(t)}), \mathbf{d}_\ell \rangle - \langle \nabla m_q^{(t)}(\mathbf{x}^{(t)}), \mathbf{d}_\ell \rangle \\ &\leq \left\| \nabla f_q(\mathbf{x}^{(t)}) - \nabla m_q^{(t)}(\mathbf{x}^{(t)}) \right\|_2,\end{aligned}$$

and if $\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) < \omega\left(\mathbf{x}^{(t)}\right)$, we obtain

$$\left|\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \omega\left(\mathbf{x}^{(t)}\right)\right| \le \left\|\boldsymbol{\nabla} m_{\ell}^{(t)}(\mathbf{x}^{(t)}) - \boldsymbol{\nabla} f_{\ell}(\mathbf{x}^{(t)})\right\|_2.$$

Because $\mathbf{m}^{(t)}$ is fully linear, it follows that

$$\left|\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \omega\left(\mathbf{x}^{(t)}\right)\right| \le \sqrt{\mathrm{c}}\dot{\epsilon}\Delta^{(t)}, \qquad \text{with } \dot{\epsilon} \text{ from (21).}$$

If we just left Algorithm 1, then the model is fully linear for $\Delta^{(t)}$ due to Lemma 1 and we have $\Delta^{(t)} \le \mu\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \le \mu\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)$. If we otherwise did not enter Algorithm 1 in the first place, it must hold that $\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \ge \varepsilon_{\mathrm{crit}}$ and

$$\Delta^{(t)} \le \Delta^{\mathrm{ub}} = \frac{\Delta^{\mathrm{ub}}}{\varepsilon_{\mathrm{crit}}}\varepsilon_{\mathrm{crit}} \le \frac{\Delta^{\mathrm{ub}}}{\varepsilon_{\mathrm{crit}}}\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)$$

and thus

$$\left|\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \omega\left(\mathbf{x}^{(t)}\right)\right| \le \kappa_{\omega}\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right), \quad \kappa_{\omega} = \sqrt{\mathrm{c}}\dot{\epsilon}\max\left\{\mu, \varepsilon_{\mathrm{crit}}^{-1}\Delta^{\mathrm{ub}}\right\} > 0.$$

□

In the subsequent analysis, we require mainly steps with fully linear models to achieve sufficient decrease for the true problem. Due to Lemma 7, we can dispose of Assumption 4 by using the criticality routine:

Assumption 9. *Either $\varepsilon_{\mathrm{crit}} > 0$ or Assumption 4 holds.*

We have also implicitly shown the following property of the criticality measures.

Corollary 5. *If $\mathbf{m}^{(t)}$ is fully linear for $\mathbf{f}$ with $\dot{\epsilon} > 0$ as in (21) then*

$$\left|\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \varpi\left(\mathbf{x}^{(t)}\right)\right| \le \left|\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \omega\left(\mathbf{x}^{(t)}\right)\right| \le \sqrt{\mathrm{c}}\dot{\epsilon}\Delta^{(t)}.$$

Lemma 8. *If $\mathbf{x}^{(t)}$ is not critical for the true problem (MOP), i.e., $\varpi\left(\mathbf{x}^{(t)}\right) \neq 0$, then Algorithm 1 will terminate after a finite number of iterations.*

Proof. At the start of Algorithm 1, we know that $\mathbf{m}^{(t)}$ is not fully linear or $\Delta^{(t)} > \mu\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)$. For clarity, we denote the first model by $\mathbf{m}_0^{(t)}$ and define $\Delta_0 = \Delta^{(t)}$. We then ensure that the model is made fully linear on $\Delta_1^{(t)} = \Delta_0$ and denote this fully linear model by $\mathbf{m}_1^{(t)}$. If afterwards $\Delta_1^{(t)} \le \mu\varpi_{\mathrm{m}_1}^{(t)}\left(\mathbf{x}^{(t)}\right)$, then Algorithm 1 terminates.

Otherwise, the process is repeated: the radius is multiplied by $\alpha \in (0,1)$ so that in the j-th iteration we have $\Delta_j^{(t)} = \alpha^{j-1}\Delta_0$ and $\mathbf{m}_j^{(t)}$ is made fully linear on $\Delta_j^{(t)}$ until

$$\Delta_j^{(t)} = \alpha^{j-1}\Delta_0 \le \mu\varpi_{\mathrm{m}_j}^{(t)}\left(\mathbf{x}^{(t)}\right).$$

The only way for Algorithm 1 to loop infinitely is

$$\varpi_{\mathrm{m}_j}^{(t)}\left(\mathbf{x}^{(t)}\right) < \frac{\alpha^{j-1}\Delta_0}{\mu} \qquad \forall j \in \mathbb{N}. \tag{22}$$

Because $\mathbf{m}_j^{(t)}$ is fully linear on $\alpha^{j-1}\Delta_0$, we know from Corollary 5 that

$$\left|\varpi_{\mathrm{m}_j}^{(t)}\left(\mathbf{x}^{(t)}\right) - \varpi\left(\mathbf{x}^{(t)}\right)\right| \le \sqrt{\mathrm{c}}\dot{\epsilon}\alpha^{j-1}\Delta_0 \qquad \forall j \in \mathbb{N}.$$

Using the triangle inequality together with (22) gives us

$$\left|\varpi\left(\mathbf{x}^{(t)}\right)\right| \le \left|\varpi_{\mathrm{m}_j}^{(t)}\left(\mathbf{x}^{(t)}\right) - \varpi\left(\mathbf{x}^{(t)}\right)\right| + \left|\varpi_{\mathrm{m}_j}^{(t)}\left(\mathbf{x}^{(t)}\right)\right| \le \left(\mu^{-1} + \sqrt{\mathrm{c}}\epsilon\right)\alpha^{j-1}\Delta_0 \quad \forall j \in \mathbb{N}.$$

As $\alpha \in (0,1)$, this implies $\varpi\left(\mathbf{x}^{(t)}\right) = 0$ and $\mathbf{x}^{(t)}$ is hence critical. □

We next state another auxiliary lemma that we need for the convergence proof.

Lemma 9. *Suppose Assumptions 6 and 7 hold. For the iterate* $\mathbf{x}^{(t)}$ *let* $\mathbf{s}^{(t)} \in \mathbb{R}^n$ *be a any step with* $\mathbf{x}_+^{(t)} = \mathbf{x}^{(t)} + \mathbf{s}^{(t)} \in B^{(t)}$*. If* $\mathbf{m}^{(t)}$ *is fully linear on* $B^{(t)}$ *then it holds that*

$$\left|\Phi(\mathbf{x}_+^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_+^{(t)})\right| \le \epsilon\left(\Delta^{(t)}\right)^2.$$

Proof. The proof follows from the definition of Φ and $\Phi_{\mathrm{m}}^{(t)}$ and the full linearity of $\mathbf{m}^{(t)}$. It can be found in [33] (Lemma 4.16). □

Convergence of Algorithm 2 is proven by showing that in certain situations, the iteration must be acceptable or successful as defined in Definition 5. This is done indirectly and relies on the next two lemmata. They use the preceding result to show that in a (hypothetical) situation where no Pareto critical point is approached, the trust region radius must be bounded from below.

Lemma 10. *Suppose Assumptions 1, 3 and 6 to 8 hold. If* $\mathbf{x}^{(t)}$ *is not Pareto critical for* (MOPm) *and* $\mathbf{m}^{(t)}$ *is fully linear on* $B^{(t)}$ *and*

$$\Delta^{(t)} \le \frac{\kappa_{\mathrm{m}}^{\mathrm{sd}}(1-\nu_{++})\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)}{2\lambda}, \quad \textit{where } \lambda = \max\{\epsilon, \mathrm{cH_m}\} \textit{ and } \kappa_{\mathrm{m}}^{\mathrm{sd}} \textit{ as in (19)},$$

then the iteration is successful, that is, $t \in \mathcal{S}$ *and* $\Delta^{t+1} \ge \Delta^{(t)}$.

Proof. The proof is very similar to [35] (Lemma 5.3) and [33] (Lemma 4.17). In contrast to the latter, we use the surrogate problem and do not require interpolation at $\mathbf{x}^{(t)}$:

By definition we have $\kappa_{\mathrm{m}}^{\mathrm{sd}}(1-\nu_{++}) < 1$ and hence it follows from Assumptions 4 and 8 and Corollary 3 that

$$\begin{aligned}\Delta^{(t)} &\le \frac{\kappa_{\mathrm{m}}^{\mathrm{sd}}(1-\nu_{++})\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)}{2\lambda}\\ &\le \frac{\varpi_{\mathrm{m}}^{(t)}}{2\lambda} \le \frac{\varpi_{\mathrm{m}}^{(t)}}{2\mathrm{cH_m}} \le \frac{\varpi_{\mathrm{m}}^{(t)}}{\mathrm{cH_m}}.\end{aligned} \tag{23}$$

With Assumption 8 we can plug this into (19) and obtain

$$\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_+^{(t)}) \ge \kappa_{\mathrm{m}}^{\mathrm{sd}}\varpi_{\mathrm{m}}^{(t)} \min\left\{\frac{\varpi_{\mathrm{m}}^{(t)}}{\mathrm{cH_m}}, \Delta^{(t)}\right\} \ge \kappa_{\mathrm{m}}^{\mathrm{sd}}\varpi_{\mathrm{m}}^{(t)}\Delta^{(t)}. \tag{24}$$

Due to Assumption 7 we can take Definition (3) and estimate

$$\left|\rho^{(t)} - 1\right| = \left| \frac{\Phi(\mathbf{x}^{(t)}) - \Phi(\mathbf{x}_+^{(t)}) - (\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_+^{(t)})}{\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_+^{(t)})} \right|$$
$$\leq \frac{\left|\Phi(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)})\right| + \left|\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_+^{(t)}) - \Phi(\mathbf{x}_+^{(t)})\right|}{\left|\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_+^{(t)})\right|}$$
$$\overset{\text{Lemma 9, (24)}}{\leq} \frac{2\epsilon\left(\Delta^{(t)}\right)^2}{\kappa_{\mathrm{m}}^{\mathrm{sd}}\varpi_{\mathrm{m}}^{(t)}\Delta^{(t)}} \leq \frac{2\lambda\Delta^{(t)}}{\kappa_{\mathrm{m}}^{\mathrm{sd}}\varpi_{\mathrm{m}}^{(t)}} \overset{(23)}{\leq} 1 - \nu_{++}.$$

Therefore $\rho^{(t)} \geq \nu_{++}$ and the iteration t using step $\mathbf{s}^{(t)}$ is successful. □

The same statement can be made for the true problem and $\varpi(\bullet)$:

Corollary 6. *Suppose Assumptions 1, 3 and 6 to 9 hold. If $\mathbf{x}^{(t)}$ is not Pareto critical for* (MOP) *and $\mathbf{m}^{(t)}$ is fully linear on $B^{(t)}$ and*

$$\Delta^{(t)} \leq \frac{\kappa^{\mathrm{sd}}(1 - \nu_{++})\varpi\left(\mathbf{x}^{(t)}\right)}{2\lambda}, \quad \textit{where } \lambda = \max\{\epsilon, \mathrm{cH_m}\}, \kappa_{\mathrm{m}}^{\mathrm{sd}} \textit{ as in } (20),$$

then the iteration is successful, that is $t \in \mathcal{S}$ and $\Delta^{t+1} \geq \Delta^{(t)}$.

Proof. The proof works exactly the same as for Lemma 10. But due to Assumption 9 we can use Lemma 7 and employ the sufficient decrease condition (20) for $\varpi(\bullet)$ instead. □

As in [35] (Lemma 5.4) and [33] (Lemma 4.18), it is now easy to show that when no Pareto critical point of (MOPm) is approached the trust region radius must be bounded:

Lemma 11. *Suppose Assumptions 1, 3 and 6 to 8 hold and that there exists a constant $\varpi_{\mathrm{m}}^{\mathrm{lb}} > 0$ such that $\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \geq \varpi_{\mathrm{m}}^{\mathrm{lb}}$ for all t. Then there is a constant $\Delta^{\mathrm{lb}} > 0$ with*

$$\Delta^{(t)} \geq \Delta^{\mathrm{lb}} \quad \textit{for all } t \in \mathbb{N}_0.$$

Proof. We first investigate the criticality step and assume $\varepsilon_{\mathrm{crit}} > \varpi_{\mathrm{m}}^{(t)} \geq \varpi_{\mathrm{m}}^{\mathrm{lb}}$. After we finish the criticality loop, we get radius $\Delta^{(t)}$ so that $\Delta^{(t)} \geq \min\{\Delta_*^{(t)}, \beta\varpi_{\mathrm{m}}^{(t)}\}$ and therefore $\Delta^{(t)} \geq \min\{\beta\varpi_{\mathrm{m}}^{\mathrm{lb}}, \Delta_*^{(t)}\}$ for all t.

Outside the criticality step, we know from Lemma 10 that whenever $\Delta^{(t)}$ falls below

$$\tilde{\Delta} := \frac{\kappa_{\mathrm{m}}^{\mathrm{sd}}(1 - \nu_{++})\varpi_{\mathrm{m}}^{\mathrm{lb}}}{2\lambda},$$

iteration t must be either model-improving or successful and hence $\Delta^{(t+1)} \geq \Delta^{(t)}$ and the radius cannot decrease until $\Delta^{(k)} > \tilde{\Delta}$ for some $k > t$. Because $\gamma_{\downarrow\downarrow} \in (0,1)$ is the severest possible shrinking factor in Algorithm 2, we therefore know that $\Delta^{(t)}$ can never be actively shrunken to a value below $\gamma_{\downarrow\downarrow}\tilde{\Delta}$.

Combining both bounds on $\Delta^{(t)}$ results in

$$\Delta^{(t)} \geq \Delta^{\mathrm{lb}} := \min\{\beta\varpi_{\mathrm{m}}^{\mathrm{lb}}, \gamma_{\downarrow\downarrow}\tilde{\Delta}, \Delta_*^{(0)}\} \qquad \forall t \in \mathbb{N}_0,$$

where we have again used the fact that $\Delta_*^{(t)}$ cannot be reduced further if it is less than or equal to $\tilde{\Delta}$ due to the update mechanism in Algorithm 2. □

We can now state the first convergence result:

Theorem 5. *Suppose that Assumptions 1, 3 and 6 to 8 hold. If Algorithm 2 has only a finite number $0 \le |\mathcal{S}| < \infty$ of successful iterations $\mathcal{S} = \{t \in \mathbb{N}_0 : \rho^{(t)} \ge \nu_{++}\}$ then*

$$\lim_{t\to\infty} \varpi\left(\mathbf{x}^{(t)}\right) = 0.$$

Proof. If the criticality loop runs infinitely, then the result follows from Lemma 8.

Otherwise, let t_0 any index larger than the last successful index (or $t_0 \ge 0$ if $\mathcal{S} = \emptyset$). All $t \ge t_0$ then must be model-improving, acceptable or inacceptable. In all cases, the trust region radius $\Delta^{(t)}$ is never increased. Due to Assumption 7, the number of successive model-improvement steps is bounded above by $\mathrm{M} \in \mathbb{N}$. Hence, $\Delta^{(t)}$ is decreased by a factor of $\gamma \in [\gamma_{\downarrow\downarrow}, \gamma_{\downarrow}] \subseteq (0,1)$ at least once every M iterations. Thus,

$$\sum_{t>t_0}^{\infty} \Delta^{(t)} \le N \sum_{i=1}^{\infty} \gamma_{\downarrow}^{i} \Delta^{(t_0)} = \frac{N\gamma_{\downarrow}}{1-\gamma_{\downarrow}} \Delta^{(t_0)},$$

and $\Delta^{(t)}$ **must go to zero** for $t \to \infty$.

Clearly, for any $\tau \ge t_0$, the iterates (and trust region centers) $\mathbf{x}^{(\tau)}$ and $\mathbf{x}^{(t_0)}$ cannot be further apart than the sum of all subsequent trust region radii, i.e.,

$$\left\|\mathbf{x}^{(\tau)} - \mathbf{x}^{(t_0)}\right\| \le \sum_{t\ge t_0}^{\infty} \Delta^{(t)} \le \frac{N\gamma_{\downarrow}}{1-\gamma_{\downarrow}} \Delta^{(t_0)}.$$

The RHS goes to zero as we let t_0 go to infinity and so must the norm on the LHS, i.e.,

$$\lim_{t_0\to\infty} \left\|\mathbf{x}^{(\tau)} - \mathbf{x}^{(t_0)}\right\| = 0. \tag{25}$$

Now let $\tau = \tau(t_0) \ge t_0$ be the first iteration index so that $\mathbf{m}^{(\tau)}$ is fully linear. Then

$$\left|\varpi_{\mathrm{m}}^{(t_0)}\right| \le \left|\varpi^{(t_0)} - \varpi^{(\tau)}\right| + \left|\varpi^{(\tau)} - \varpi_{\mathrm{m}}^{(\tau)}\right| + \left|\varpi_{\mathrm{m}}^{(\tau)}\right|$$

and for the terms on the right and for $t_0 \to \infty$, we find:

- Because of Assumptions 1 and 6 and Theorem 2 $\varpi(\bullet)$ is Cauchy-continuous and with (25) the first term goes to zero.
- Due to Corollary 5 the second term is in $\mathcal{O}(\Delta^{(\tau)})$ and goes to zero.
- Suppose the third term does not go to zero as well, i.e., $\{\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(\tau)}\right)\}$ is bounded below by a positive constant. Due to Assumptions 1 and 7 the iterates $x^{(\tau)}$ are not Pareto critical for (MOPm) and because of $\Delta^{(\tau)} \to 0$ and Lemma 10 there would be a successful iteration, a contradiction. Thus the third term must go to zero as well.

  We conclude that the left side, $\varpi\left(\mathbf{x}^{(t_0)}\right)$, goes to zero as well for $t_0 \to \infty$. □

We now address the case of infinitely many successful iterations, first for the surrogate measure $\varpi_{\mathrm{m}}^{(t)}(\bullet)$ and then for $\varpi(\bullet)$. We show that the criticality measures are not bounded away from zero.

We start with the observation that in any case the trust region radius converges to zero:

Lemma 12. *If Assumptions 1, 3 and 6 to 8 hold, then the subsequence of trust region radii generated by Algorithm 2 goes to zero, i.e.,* $\lim_{t\to\infty} \Delta^{(t)} = 0$.

Proof. We have shown in the proof of Theorem 5 that this is the case for finitely many successful iterations.

Suppose there are infinitely many successful iterations. Take any successful index $t \in \mathcal{S}$. Then $\rho^{(t)} \geq \nu_{++}$ and from Assumption 8 it follows for $\mathbf{x}^{(t+1)} = \mathbf{x}_+^{(t)} = \mathbf{x}^{(t)} + \mathbf{s}^{(t)}$ that

$$\Phi(\mathbf{x}^{(t)}) - \Phi(\mathbf{x}_+^{(t)}) \geq \nu_{++}\left(\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_+^{(t)})\right) \overset{(19)}{\geq} \nu_{++}\kappa_{\mathrm{m}}^{\mathrm{sd}}\varpi_{\mathrm{m}}^{(t)} \min\left\{\frac{\varpi_{\mathrm{m}}^{(t)}}{\mathrm{cH_m}}, \Delta^{(t)}\right\}.$$

The criticality step ensures that $\varpi_{\mathrm{m}}^{(t)} \geq \min\left\{\varepsilon_{\mathrm{crit}}, \frac{\Delta^{(t)}}{\mu}\right\}$ so that

$$\Phi(\mathbf{x}^{(t)}) - \Phi(\mathbf{x}_+^{(t)}) \geq \nu_{++}\kappa_{\mathrm{m}}^{\mathrm{sd}} \min\left\{\varepsilon_{\mathrm{crit}}, \frac{\Delta^{(t)}}{\mu}\right\} \min\left\{\frac{\Delta^{(t)}}{\mu \mathrm{cH_m}}, \Delta^{(t)}\right\} \geq 0. \tag{26}$$

Now the right hand side has to go to zero: Suppose it was bounded below by a positive constant $\varepsilon > 0$. We could then compute a lower bound on the improvement from the first iteration with index 0 up to $t+1$ by summation

$$\Phi(\mathbf{x}^{(0)}) - \Phi(\mathbf{x}^{(t+1)}) \geq \sum_{\tau \in \mathcal{S}_t} \Phi(\mathbf{x}^{(\tau)}) - \Phi(\mathbf{x}^{(\tau+1)}) \geq |\mathcal{S}_t|\varepsilon$$

where $\mathcal{S}_t = \mathcal{S} \cap \{0, \ldots, t\}$ are all successful indices with a maximum index of t. Because $\mathcal{S}$ is unbounded, the right side diverges for $t \to \infty$ and so must the left side in contradiction to Φ being bounded below by Assumption 5. From (26) we see that this implies $\Delta^{(t)} \to 0$ for $t \in \mathcal{S}, t \to \infty$.

Now consider any sequence $\mathcal{T} \subseteq N$ of indices that are not necessarily successful, i.e., $|\mathcal{T} \setminus \mathcal{S}| \geq 0$. The radius is only ever increased in successful iterations and at most by a factor of $\gamma_{\uparrow}$. Since $\mathcal{S}$ is unbounded, there is for any $\tau \in \mathcal{T}$ a largest $t_\tau \in \mathcal{S}$ with $t_\tau \leq \tau$. Then $\Delta^{(\tau)} \leq \gamma_{\uparrow}\Delta^{(t_\tau)}$ and because of $\Delta^{(t_\tau)} \to 0$ it follows that

$$\lim_{\substack{\tau \in \mathcal{T}, \\ \tau \to \infty}} \Delta^{(\tau)} = 0,$$

which concludes the proof. □

Lemma 13. *Suppose Assumptions 1, 3 and 5 to 8 hold. For the iterates produced by Algorithm 2 it holds that*

$$\liminf_{t \to \infty} \varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) = 0.$$

Proof. For a contradiction, suppose that $\liminf_{t\to\infty} \varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \neq 0$. Then there is a constant $\varpi_{\mathrm{m}}^{\mathrm{lb}} > 0$ with $\varpi_{\mathrm{m}}^{(t)} \geq \varpi_{\mathrm{m}}^{\mathrm{lb}}$ for all $t \in \mathbb{N}_0$. According to Lemma 11, there exists a constant $\Delta^{\mathrm{lb}} > 0$ with $\Delta^{(t)} \geq \Delta^{\mathrm{lb}}$ for all t. This contradicts Lemma 12. □

The next result allows us to transfer the result to $\varpi(\bullet)$.

Lemma 14. *Suppose Assumptions 1, 6 and 7 hold. For any subsequence $\{t_i\}_{i\in\mathbb{N}} \subseteq \mathbb{N}_0$ of iteration indices of Algorithm 2 with*

$$\lim_{i \to \infty} \varpi_{\mathrm{m}}^{(t_i)}\left(\mathbf{x}^{(t_i)}\right) = 0, \tag{27}$$

it also holds that

$$\lim_{i \to \infty} \varpi\left(\mathbf{x}^{(t_i)}\right) = 0. \tag{28}$$

Proof. By (27), $\varpi_{\mathrm{m}}^{(t_i)} < \varepsilon_{\mathrm{crit}}$ for sufficiently large i. If $\mathbf{x}^{(t_i)}$ is critical for (MOP), then the result follows from Lemma 8. Otherwise, $\mathbf{m}^{(t_i)}$ is fully linear on $B\left(\mathbf{x}^{(t_i)}; \Delta^{(t_i)}\right)$ for some $\Delta^{(t_i)} \leq \mu\varpi_{\mathrm{m}}^{(t_i)}$. From Corollary 5 it follows that

$$\left|\varpi_{\mathrm{m}}^{(t_i)} - \varpi^{(t_i)}\right| \leq \sqrt{c}\dot{\epsilon}\Delta^{(t_i)} \leq \sqrt{c}\dot{\epsilon}\mu\varpi_{\mathrm{m}}^{(t_i)}.$$

The triangle inequality yields

$$\varpi^{(t_i)} \leq \left|\varpi^{(t_i)} - \varpi_{\mathrm{m}}^{(t_i)}\right| + \varpi_{\mathrm{m}}^{(t_i)} \leq (\sqrt{c}\dot{\epsilon}\mu + 1)\varpi_{\mathrm{m}}^{(t_i)}$$

for sufficiently large i and (27) then implies (28). □

The next global convergence result immediately follows from Theorem 5 and Lemmas 13 and 14:

Theorem 6. *Suppose Assumptions 1, 3 and 5 to 8 hold. Then* $\liminf_{t\to\infty} \varpi\left(\mathbf{x}^{(t)}\right) = 0$.

This shows that if the iterates are bounded, then there is a subsequence of iterates in $\mathbb{R}^n$ approximating a Pareto critical point. We next show that *all* limit points of a sequence generated by Algorithm 2 are Pareto critical.

Theorem 7. *Suppose Assumptions 1 and 3 to 8 hold. Then* $\lim_{t\to\infty} \varpi\left(\mathbf{x}^{(t)}\right) = 0$.

Proof. We have already proven the result for finitely many successful iterations, see Theorem 5. We thus suppose that $\mathcal{S}$ is unbounded.

For the purpose of establishing a contradiction, suppose that there exists a sequence $\{t_j\}_{j\in\mathbb{N}}$ of indices that are successful or acceptable with

$$\varpi^{\left(t_j\right)} \geq 2\varepsilon > 0 \quad \text{for some } \varepsilon > 0 \text{ and all } j. \tag{29}$$

We can ignore model-improving and inacceptable iterations: During those the iterate does not change, and we find a larger acceptable or successful index with the same criticality value.

From Theorem 6 we obtain that for every such t_j, there exists a first index $\tau_j > t_j$ such that $\varpi\left(\mathbf{x}^{(\tau_j)}\right) < \varepsilon$. We thus find another subsequence indexed by $\{\tau_j\}$ such that

$$\varpi^{(t)} \geq \varepsilon \text{ for } t_j \leq t < \tau_j \text{ and } \varpi^{\left(\tau_j\right)} < \varepsilon. \tag{30}$$

Using (29) and (30), it also follows from a triangle inequality that

$$\left|\varpi^{\left(t_j\right)} - \varpi^{\left(\tau_j\right)}\right| \geq \varpi^{\left(t_j\right)} - \varpi^{\left(\tau_j\right)} > 2\varepsilon - \varepsilon = \varepsilon \qquad \forall j \in \mathbb{N}. \tag{31}$$

With $\{t_j\}$ and $\{\tau_j\}$ as in (30), define the following subset set of indices

$$\mathcal{T} = \left\{t \in \mathbb{N}_0 : \exists j \in \mathbb{N} \text{ such that } t_j \leq t < \tau_j\right\}.$$

By (30) we have $\varpi^{(t)} \geq \varepsilon$ for $t \in \mathcal{T}$, and due to Lemma 14, we also know that then $\varpi_{\mathrm{m}}^{(t)}$ cannot go to zero neither, i.e., there is some $\varepsilon_{\mathrm{m}} > 0$ such that

$$\varpi_{\mathrm{m}}^{(t)} \geq \varepsilon_{\mathrm{m}} > 0 \qquad \forall t \in \mathcal{T}.$$

From Lemma 12 we know that $\Delta^{(t)} \xrightarrow{t\to\infty} 0$ so that by Corollary 6, any sufficiently large $t \in \mathcal{T}$ must be either successful or model-improving (if $\mathbf{m}^{(t)}$ is not fully linear). For $t \in \mathcal{T} \cap \mathcal{S}$, it follows from Assumption 8 that

$$\Phi(\mathbf{x}^{(t)}) - \Phi(\mathbf{x}^{(t+1)}) \geq \nu_{++}\left(\Phi_{\mathrm{m}}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}(\mathbf{x}^{(t+1)})\right) \geq \nu_{++}\kappa_{\mathrm{m}}^{\mathrm{sd}}\varepsilon_{\mathrm{m}} \min\left\{\frac{\varepsilon_{\mathrm{m}}}{\mathrm{cH_m}}, \Delta^{(t)}\right\} \geq 0.$$

If $t \in \mathcal{T} \cap \mathcal{S}$ is sufficiently large, we have $\Delta^{(t)} \leq \dfrac{\varepsilon_{\mathrm{m}}}{\mathrm{cH_m}}$ and

$$\Delta^{(t)} \leq \frac{1}{\nu_{++}\kappa_{\mathrm{m}}^{\mathrm{sd}}\varepsilon_{\mathrm{m}}}\Big(\Phi(\mathbf{x}^{(t)}) - \Phi(\mathbf{x}^{(t+1)})\Big).$$

Since the iteration is either successful or model-improving for sufficiently large $t \in \mathcal{T}$, and since $\mathbf{x}^{(t)} = \mathbf{x}^{(t+1)}$ for a model-improving iteration, we deduce from the previous inequality that

$$\left\|\mathbf{x}^{(t_j)} - \mathbf{x}^{(\tau_j)}\right\| \leq \sum_{\substack{t=t_j,\\ t\in\mathcal{T}\cap\mathcal{S}}}^{\tau_j-1} \left\|\mathbf{x}^{(t)} - \mathbf{x}^{(t+1)}\right\| \leq \sum_{\substack{t=t_j,\\ t\in\mathcal{T}\cap\mathcal{S}}}^{\tau_j-1} \Delta^{(t)} \leq \frac{1}{\nu_{++}\kappa_{\mathrm{m}}^{\mathrm{sd}}\varepsilon_{\mathrm{m}}}\Big(\Phi(\mathbf{x}^{(t_j)}) - \Phi(\mathbf{x}^{(\tau_j)})\Big)$$

for $j \in \mathbb{N}$ sufficiently large. The sequence $\left\{\Phi(\mathbf{x}^{(t)})\right\}_{t\in\mathbb{N}_0}$ is bounded below (Assumption 5) and monotonically decreasing by construction. Hence, the RHS above must converge to zero for $j \to \infty$. This implies $\lim_{j\to\infty}\left\|\mathbf{x}^{(t_j)} - \mathbf{x}^{(\tau_j)}\right\| = 0$.

Because of Assumptions 1 and 6, $\varpi(\bullet)$ is uniformly continuous so that then

$$\lim_{j\to\infty} \varpi\left(\mathbf{x}^{(t_j)}\right) - \varpi\left(\mathbf{x}^{(\tau_j)}\right) = 0,$$

which is a contradiction to (31). Thus, no subsequence of acceptable or successful indices as in (29) can exist.

□

7. Numerical Examples

In this section we provide some more details on the actual implementation of Algorithm 2 and present the results of various experiments. We compare different surrogate model types with regard to their efficacy (in terms of expensive objective evaluations) and their ability to find Pareto critical points.

7.1. Implementation Details

We implemented the framework in the Julia language (the code is available under https://github.com/manuelbb-upb/Morbit.jl, accessed on 15 April 2021) and used the surrogate construction algorithms from Sections 4.2 and 4.3. Concerning the RBF models, the algorithms are thus the same as in [41]. The `OSQP` solver [45] is used to solve (Pm). For non-linear problems we use the `NLopt.jl` [46] package. More specifically we use the `MMA` algorithm [47] in conjunction with `DynamicPolynomials.jl` [48] to construct the Lagrange polynomials. The Pascoletti–Serafini subproblems is solved using the population based `ISRES` method [49] with `MMA` for polishing. The derivatives of cheap objective functions are obtained by means of automatic differentiation [50] and Taylor models use `FiniteDiff.jl`.

In accordance with Algorithm 2, we perform the shrinking trust region update via

$$\Delta^{(t+1)} \leftarrow \begin{cases} \gamma_{\downarrow\downarrow}\Delta^{(t)} & \text{if } \rho^{(t)} < \nu_{+}, \\ \gamma_{\downarrow}\Delta^{(t)} & \text{if } \rho^{(t)} < \nu_{++}. \end{cases}$$

Note that for box-constrained problems we internally scale the feasible set to the unit hypercube $[0,1]^n$ and all radii are measured with regard to this scaled domain.

For **stopping**, we use a disjunction of different criteria:

- We have an upper bound $\mathrm{N_{it.}} \in \mathbb{N}$ on the maximum number of iterations and an upper bound $\mathrm{N_{exp.}} \in \mathbb{N}$ on the number of expensive objective evaluations.

- The surrogate criticality naturally allows for a stopping test and due to Lemma 11 the trust region radius can also be used (see also [33] [Sec. 5]). We combine this with a relative tolerance test and stop if

$$\Delta^{(t)} \leq \Delta_{\min} \text{ OR } \left(\Delta^{(t)} \leq \Delta_{\text{crit}} \text{ AND } \omega\left(\mathbf{x}^{(t)}\right) \leq \omega_{\min}\right).$$

- At a truly critical point the criticality loop Algorithm 1 runs infinitely. We stop after a maximum number $\text{N}_{\text{loops}} \in \mathbb{N}_0$ of iterations.
- We also employ the common relative stopping criteria

$$\left\|\mathbf{x}^{(t)} - \mathbf{x}^{(t+1)}\right\|_\infty \leq \delta_x \left\|\mathbf{x}^{(t)}\right\|_\infty \quad \text{and}$$
$$\left\|\mathbf{f}(\mathbf{x}^{(t)}) - \mathbf{f}(\mathbf{x}^{(t+1)})\right\|_\infty \leq \delta_f \left\|\mathbf{f}(\mathbf{x}^{(t)})\right\|_\infty$$

to provoke early stopping.

7.2. A First Example

We ran our method on a multitude of academic test problems with a varying number of decision variables n and objective functions k. We were able to approximate Pareto critical points in both cases, if we treat the problems as heterogeneous and if we declare them as expensive. We benchmarked RBF against polynomial models, because in [33] it was shown that a trust region method using second degree Lagrange polynomials outperforms commercial solvers on scalarized problems. Most often, RBF surrogates outperform other model types with regard to the number of expensive function evaluations.

This is illustrated in Figure 2. It shows two runs of Algorithm 2 on the non-convex problem (T6), taken from [38]:

$$\min_{\mathbf{x}\in\mathcal{X}} \begin{bmatrix} x_1 + \ln(x_1) + x_2^2, \\ x_1^2 + x_2^4 \end{bmatrix}, \mathcal{X} = [\varepsilon, 30] \times [0, 30] \subseteq \mathbb{R}^2, \varepsilon = 10^{-12}. \tag{T6}$$

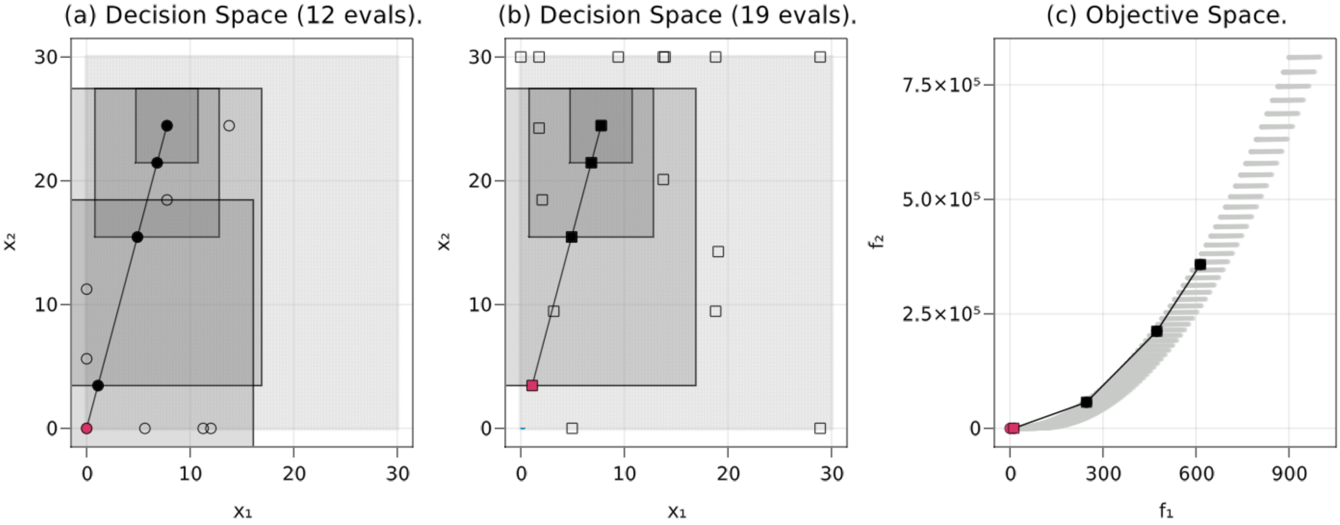

Figure 2. Two runs with maximum number of expensive evaluations set to 20 (soft limit). Test points are light-gray, the iterates are black, final iterate is red, white markers show other points where the objectives are evaluated. The successive trust regions are also shown. (**a**) Using Radial Basis Function (RBF) surrogate models we converge to the optimum using only 12 expensive evaluations. (**b**) Quadratic Lagrange models do not reach the optimum using 19 evaluations. (**c**) Iterations and test points in the objective space.

The first objective function is treated as expensive while the second is cheap. In contrast to most other MOPs, there is only one solution and this Pareto optimal point is $[\varepsilon, 0]^T$. When we set a very restrictive limit of $\mathrm{N}_{\mathrm{exp.}} = 20$ then we run out of budget with second degree Lagrange surrogates before we reach the optimum, see Figure 2b. As evident in Figure 2a, surrogates based on (cubic) RBF do require significantly less training data. For the RBF models the algorithm stopped after two critical loops and the model refinement during these loops is made clear by the samples on the problem boundary converging to zero. The complete set of relevant parameters for the test runs is given in Table 2. We used a strict acceptance test and the strict Pareto–Cauchy step.

Table 2. Parameters for Figure 2, radii relative to $[0,1]^n$.

Param.	$\varepsilon_{\mathrm{crit}}$	$\mathrm{N}_{\mathrm{exp.}}$	$\mathrm{N}_{\mathrm{loops}}$	μ	β	Δ^{ub}	Δ_{min}	$\Delta^{(0)}$	ν_{+}	ν_{++}	$\gamma_{\downdownarrows}$	$\gamma_{\downarrow}$	$\gamma_{\uparrow}$
Value	10^{-3}	20	2	2×10^3	10^3	0.5	10^{-3}	0.1	0.1	0.4	0.51	0.75	2

7.3. Benchmarks on Scalable Test-Problems

To assess the performance with a growing number of decision variables n, we performed tests on scalable problems of the ZDT and DTLZ family [51,52]. Figure 3 shows results for the bi-objective problems ZDT1-ZDT3 and for the k-objective problems DTLZ1 and DTLZ6 (we used $k = \max\{2, n-4\}$ objectives). All problems are box constrained. Twelve feasible starting points (from the Halton sequence) were generated for each problem setting, i.e., for each combination of n, a test problem and a descent method. The acceptance test and the backtracking were strict.

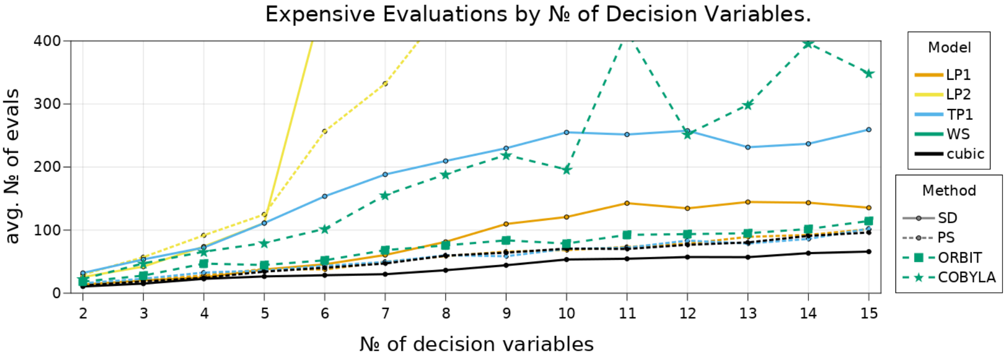

Figure 3. Average number of expensive objective evaluations by number of decision variables n, surrogate type and descent method. “SD” refers to steepest descent and “PS” to Pascoletti–Serafini. “LP1” (orange) are linear Lagrange models, “LP2” (yellow) quadratic Lagrange models, “TP1” (blue) are linear Taylor polynomials based on finite differences and “cubic” (black) refers to cubic RBF models. Additionally the results for weighted sum runs are shown in green, using the `COBYLA` solver and a single objective variant of the trust region framework, `ORBIT`.

In all cases the first objective was considered cheap and all other objectives expensive. First and second degree Lagrange models were compared against linear Taylor models and (cubic) RBF surrogates. The Lagrange models were built using a Λ-poised set, with $\Lambda = 1.5$. In the case of quadratic models we used a precomputed set of points for $n \geq 6$. The Taylor models used finite differences and points outside of box constraints were simply projected back onto the boundary. The RBF models were allowed to include up to $(n+1)(n+2)/2$ training points from the database if $n \leq 10$ and else the maximum number of points was $2n+1$. Points were first selected from a box of radius $\theta_1 \Delta^{(t)}$ with $\theta_1 = 2$ and then from a box of radius $\theta_2 \Delta^{\mathrm{ub}}$ with $\theta_2 = 2$. All other parameters differing from the parameters in Table 2 are listed in Table 3. The stopping parameters were chosen so as to exit early and save evaluations.

Table 3. Parameters for Figure 3, radii relative to $[0,1]^n$.

Parameter	$\varepsilon_{\mathrm{crit}}$	$\mathrm{N}_{\mathrm{it.}}$	$\mathrm{N}_{\mathrm{exp.}}$	$\mathrm{N}_{\mathrm{loops}}$	Δ_{crit}	ω_{min}	Δ_{min}	δ_x	δ_f	ν_+	ν_{++}
Value	10^{-2}	100	$n \times 10^3$	3	10^{-2}	10^{-3}	10^{-6}	10^{-3}	10^{-3}	0	0.1

As expected, the second degree Lagrange polynomials require the most objective evaluations and the quadratic dependence on n is clearly visible in Figure 3, and the quadratic growth of the dark-blue line continues for $n \geq 8$. On average, the linear Lagrange models perform better than the linear Taylor polynomials when using the steepest descent steps—also in accordance with our expectations, because only $n+1$ points are needed for each model (versus $2n$ points). Most models—even the linear ones—profit from using the Pascoletti–Serafini subproblems (see Appendix B) over the steepest descent steps. By far the least evaluations (on average) are needed for the RBF models: The black line consistently stays below all other data points. Note, that the RBF models likely appear to perform slightly better with the steepest descent steps because of the early stopping. In other experiments we noticed that RBF models with Pascoletti–Serafini steps can save evaluations when more precise solutions are required.

For comparison, we also used the weighted sum approach with the single objective $\sum_\ell f_\ell$ on each problem instance. We tested both the derivative-free `COBYLA` solver (described in [53] and implemented by `NLopt.jl`) and the trust region method using steepest descent and cubic RBF models, i.e., our own implementation of `ORBIT` [34]. Both solvers were restricted to the same number of maximum function evaluations. In fact, `ORBIT` was configured with the exact same parameters as in Table 3 and the relative stopping tolerances for `COBYLA` were $\delta_x = \delta_f = 10^{-2}$. Although, `COBYLA` also uses linear models it requires significantly more evaluations than most other algorithms. The results of the `ORBIT` scalarization are more comparable to that of the multiobjective runs.

7.3.1. Solution Quality

Figure 4 illustrates that not only do RBF perform better on average, but also overall. With regard to the final solution criticality, there are a few outliers mostly due to DTLZ1 (see also Figure 5). However, in most cases the solution criticality is acceptable, except for the linear Lagrange models. Moreover, Figure 5 shows that a good percentage of problem instances is solved with RBF, especially when compared to the other linear models. Note, that in cases where the true objectives are not differentiable at the final iterate, ω was set to 0 because the selected problems are non-differentiable only in Pareto optimal points. In Figure 5 it also becomes apparent that the bi-objective DTLZ1 instances were the most challenging for all algorithms. DTLZ1 has many local minima and it is likely to exit early near such a local minimum due to repeated unsuccessful iterations. Likewise, ZDT3 is “flat” towards the true Pareto Front so that it becomes hard to make progress there.

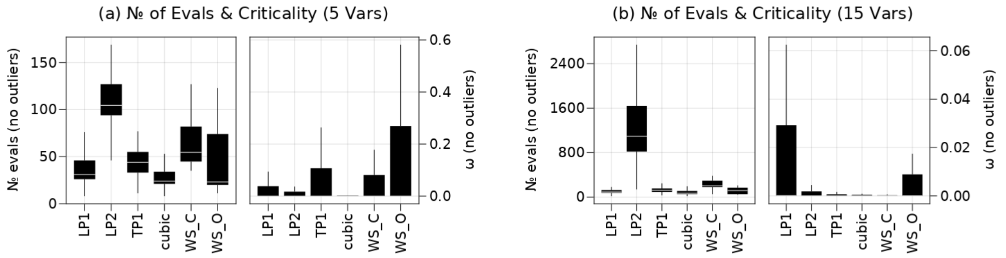

Figure 4. Box-plots of the number of evaluations and the solution criticality for $n = 5$ and $n = 15$ for the runs from Figure 3. Outliers are not shown. “WS_C” and “WS_O” refer to the weighted sum approach using `COBYLA` and `ORBIT`, respectively.

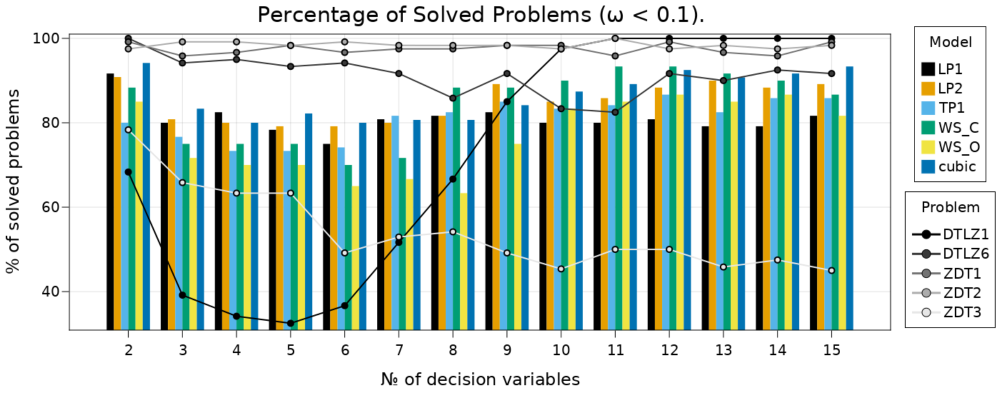

Figure 5. Each group of bars shows the percentage of solved problem instances, i.e., test runs were the final solution criticality has a value below 0.1. From left to right, the bars correspond to the Trust Region Method (TRM) using linear Lagrange polynomials, the TRM with quadratic Lagrange polynomials, TRM with linear Taylor polynomials, weighted sum with `COBYLA`, weighted sum with `ORBIT` and TRM with cubic RBF. Per model and n-value there were 60 runs.

Besides criticality, another metric of interest is the spread of solutions for different starting points. Figure 6 shows the final iterates when the algorithm is applied to the bi-objective problems ZDT1 and ZDT2 for 10 different starting points. Additionally, the problems are solved using the weighted sum approach with the derivative-free `COBYLA` solver. For each starting point the optimizers were allowed 30 objective evaluations and no data were re-used between runs.

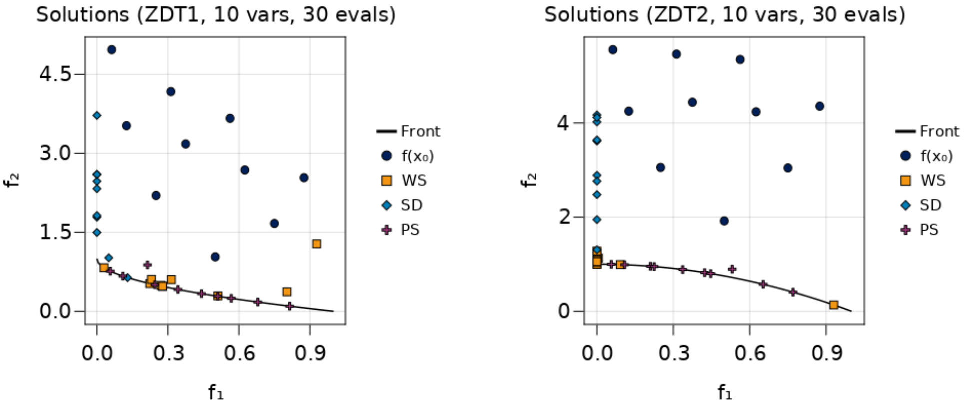

Figure 6. Final iterates in objective space for the bi-objective problems ZDT1 and ZDT2 in 10 variables. The weighted sum method (WS) is compared against the trust region method using steepest descent (DS) and the Pascoletti–Serafini (PS) method.

As can bee seen, for these problems, the trust region method readily reaches the critical set using only 30 evaluations. Here, the steepest descent direction tends to generate solutions on the problem boundary when applied in such a global manner—with relatively large trust region radii ($\Delta^{(0)} = 0.1$ and $\Delta^{\text{ub}} = 0.5$). Nonetheless, the method remains applicable for local refinement of approximate solutions, e.g., after a coarse search for good starting points using global methods or as a corrector in continuation frameworks. The Pascoletti–Serafini step can be employed with different reference points/directions to provide a better covering than both the steepest descent steps and the weighted sum approach. For Figure 6, the points $\{[0, -10i], i = 1, \ldots, 10\}$ were used. The weighted sum approach (with fixed weights) tends to produce clustered solutions. Especially for the non-convex problem ZDT2 only the boundary points of the true Pareto Front are reached, as expected [1].

7.3.2. RBF Comparison

Furthermore, we compared the RBF kernels from Table 1. In [34], the cubic kernel performs best on single-objective problems while the Gaussian does worst. As can be seen in Figure 7 this holds for multiple objective functions, too: The Gaussian and the Multiquadric require more function evaluations than the Cubic, especially in higher dimensions. If, however, we use a very simple adaptive strategy to fine-tune the shape parameter, then both kernels can finish significantly faster. In both cases, the shape parameter was set to $\alpha = 20/\Delta^{(t)}$ in each iteration. Nevertheless, the cubic function appears to be a good choice in general.

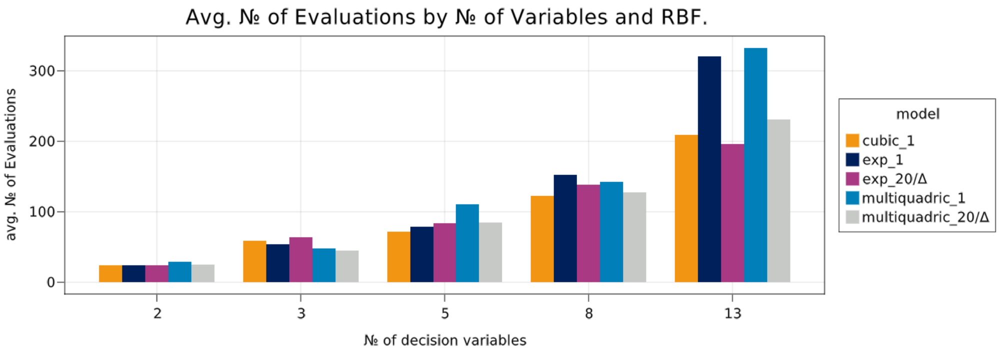

Figure 7. Each group of bars shows the influence of a adaptive shape radius on the performance of different RBF models (tested on ZDT3) for different decision space dimensions. From left to right the bars correspond to the cubic RBF, the Gaussian—with constant shape factor 1 and with adaptive shape factor $20/\Delta^{(t)}$—and the Multiquadric—with shape factors 1 and $20/\Delta^{(t)}$.

8. Conclusions

We have developed a trust region framework for heterogeneous and expensive multiobjective optimization problems. It is based on similar work [29–31,33] and our main contributions are the integration of constraints and of radial basis function surrogates. Subsequently, our method is is provably convergent to first order critical points for unconstrained problems and when the feasible set is convex and compact, while requiring significantly less expensive function evaluations due to a linear scaling of model construction complexity with respect to the number of decision variables.

For future work, several modifications and extensions can likely be transferred from the single-objective to the multiobjective case. For examples, the trust region update can be made step-size-dependent (rather than to depend $\rho^{(t)}$ alone) to allow for a more precise model refinement, see [36] ([Ch. 10]). We have also experimented with the nonlinear CG method [9] for a multiobjective Steihaug–Toint step [36] ([Ch. 7]) and early results look promising.

Going forward, we would like to apply our algorithm to a real world application, similar to what has been done in [54]. Moreover, it would be desirable to obtain not just one but multiple Pareto critical solutions. Because the Pascoletti–Serafini scalarization is still compatible with constraints, the iterations can be guided in image space by providing different global reference points. Furthermore, it is straightforward to use RBF with the heuristic methods from [55] for heterogeneous problems. We believe that it should also be possible to propagate multiple solutions and combine the TRM method with non-dominance testing as has been done [31] and in [56]. One can think of other globalization strategies as well: RBF models have been used in multiobjective Stochastic Search algorithms [57] and trust region ideas have been included into population based strategies [26]. It will thus be interesting to see whether the theoretical convergence properties can be maintained within these contexts by employing a careful trust-region management. Finally, re-using the data sampled near the final iterate within a continuation framework like in [58] is a promising next step.

Supplementary Materials: Our Julia implementation of the solver is available online at https://github.com/manuelbb-upb/Morbit.jl accessed on 15 April 2021.

Author Contributions: Conceptualization, M.B. and S.P.; methodology, M.B.; software, M.B.; validation, M.B. and S.P.; formal analysis, M.B. and S.P.; investigation, M.B.; writing—original draft preparation, M.B.; writing—review and editing, S.P.; visualization, M.B.; supervision, S.P.; All authors have read and agreed to the published version of the manuscript.

Funding: This research has been funded by the European Union and the German Federal State of North Rhine-Westphalia within the EFRE.NRW project "SET CPS".

EUROPEAN UNION
Investing in our Future
European Regional
Development Fund

EFRE.NRW
Investitionen in Wachstum
und Beschäftigung

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Miscellaneous Proofs

Appendix A.1. Continuity of the Constrained Optimal Value

In this subsection we show the continuity of $\omega(\mathbf{x})$ in the constrained case, where $\omega(\mathbf{x})$ is the negative optimal value of (P1), i.e.,

$$\begin{aligned} \omega(\mathbf{x}) := -\min_{\mathbf{d}\in\mathcal{X}-\mathbf{x}} \max_{\ell=1,\dots,k} \langle \nabla f_\ell(\mathbf{x}), \mathbf{d}\rangle, \\ \text{s.t. } \|\mathbf{d}\| \le 1. \end{aligned}$$

The proof of the continuity of $\omega(\mathbf{x})$, as stated in Theorem 1, follows the reasoning from [6], where continuity is shown for a related constrained descent direction program.

Proof of Item 2 in Theorem 1. Let the requirements of Item 1 be fulfilled, i.e., let $\mathbf{f}$ be continuously differentiable and let $\mathcal{X} \subset \mathbb{R}^n$ be convex and compact. Further, let $\mathbf{x}$ be a point in $\mathcal{X}$ and denote the minimizing direction in (P1) by $\mathbf{d}(\mathbf{x})$ and the optimal value by $\theta(\mathbf{x})$. We show that $\theta(\mathbf{x})$ is continuous, by which $\omega(\mathbf{x}) = -\theta(\mathbf{x})$ is continuous as well.

First, note the following properties of the maximum function:

1. $\mathbf{u} \mapsto \max_\ell u_\ell$ is positively homogenous and hence

$$\max_\ell(\langle \nabla f_\ell(\mathbf{x}), \mathbf{d}_1\rangle + \langle \nabla f_\ell(\mathbf{x}), \mathbf{d}_2\rangle) \le \max_\ell \langle \nabla f_\ell(\mathbf{x}), \mathbf{d}_1\rangle + \max_\ell \langle \nabla f_\ell(\mathbf{x}), \mathbf{d}_2\rangle.$$

2. $\mathbf{u} \mapsto \max_\ell u_\ell$ is Lipschitz with constant 1 so that

$$\left| \max_\ell \langle \nabla f_\ell(\mathbf{x}_1), \mathbf{d}_1\rangle - \max_\ell \langle \nabla f_\ell(\mathbf{x}_2), \mathbf{d}_2\rangle \right| \le \|\mathbf{Df}(\mathbf{x}_1)\mathbf{d}_1 - \mathbf{Df}(\mathbf{x}_2)\mathbf{d}_2\|,$$

for both the maximum and the Euclidean norm.

Now let $\{\mathbf{x}^{(t)}\} \subseteq \mathcal{X}$ be a sequence with $\mathbf{x}^{(t)} \to \mathbf{x}$. Due to the constraints, we have that $\mathbf{d}(\mathbf{x}) \in \mathcal{X} - \mathbf{x}$ and thereby $\mathbf{d}(\mathbf{x}) + \mathbf{x} - \mathbf{x}^{(t)} \in \mathcal{X} - \mathbf{x}^{(t)}$. Let

$$(0, 1] \ni \sigma^{(t)} := \begin{cases} \min\left\{1, \dfrac{1}{\left\|\mathbf{d}(\mathbf{x}) + \mathbf{x} - \mathbf{x}^{(t)}\right\|}\right\} & \text{if } \mathbf{d}(\mathbf{x}) \ne \mathbf{x}^{(t)} - \mathbf{x}, \\ 1 & \text{else.} \end{cases}$$

Then $\sigma^{(t)}\left(\mathbf{d}(\mathbf{x}) + \mathbf{x} - \mathbf{x}^{(t)}\right)$ is feasible for (P1) at $\mathbf{x}^{(t)}$:

- $\sigma^{(t)}\left(\mathbf{d}(\mathbf{x})+\mathbf{x}-\mathbf{x}^{(t)}\right) \in \mathcal{X}-\mathbf{x}^{(t)}$ because $\mathcal{X}-\mathbf{x}^{(t)}$ is convex and $\mathbf{0}$, $\left(\mathbf{d}(\mathbf{x})+\mathbf{x}-\mathbf{x}^{(t)}\right) \in \mathcal{X}-\mathbf{x}^{(t)}$ as well as $\sigma^{(t)} \in (0,1]$.
- $\left\|\sigma^{(t)}\left(\mathbf{d}(\mathbf{x})+\mathbf{x}-\mathbf{x}^{(t)}\right)\right\| \leq 1$ by the definition of $\sigma^{(t)}$.

  By the definition of (P1) it follows that

$$\begin{aligned}
&\max_\ell\langle\nabla f_\ell(\mathbf{x}^{(t)}),\mathbf{d}(\mathbf{x}^{(t)})\rangle \leq \sigma^{(t)}\max_\ell\langle\nabla f_\ell(\mathbf{x}^{(t)}),\mathbf{d}(\mathbf{x})+\mathbf{x}-\mathbf{x}^{(t)}\rangle\\
&\text{and by the maximum property 1}\\
&\max_\ell\langle\nabla f_\ell(\mathbf{x}^{(t)}),\mathbf{d}(\mathbf{x}^{(t)})\rangle \leq \sigma^{(t)}\max_\ell\langle\nabla f_\ell(\mathbf{x}^{(t)}),\mathbf{d}(\mathbf{x})\rangle+\sigma^{(t)}\max_\ell\langle\nabla f_\ell(\mathbf{x}^{(t)}),\mathbf{x}-\mathbf{x}^{(t)}\rangle.
\end{aligned} \tag{A1}$$

We make the following observations:

- Because of $\left\|\mathbf{d}(\mathbf{x})+\mathbf{x}-\mathbf{x}^{(t)}\right\| \xrightarrow{t\to\infty} \|\mathbf{d}(\mathbf{x})\| \leq 1$, it follows that $\sigma^{(t)} \xrightarrow{t\to\infty} 1$.
- Because all objective gradients are continuous, it holds for all $\ell \in \{1,\dots,k\}$ that $\nabla f_\ell(\mathbf{x}^{(t)}) \to \nabla f_\ell(\mathbf{x})$ and because $\mathbf{u} \mapsto \max_\ell u_\ell$ is continuous as well, it then follows that
$$\max_\ell\langle\nabla f_\ell(\mathbf{x}^{(t)}),\mathbf{d}(\mathbf{x})\rangle \to \max_\ell\langle\nabla f_\ell(\mathbf{x}),\mathbf{d}(\mathbf{x})\rangle \quad \text{for } t\to\infty.$$
- The last term on the RHS of (A1) vanishes for $t\to\infty$.

  By taking the limit superior on (A1), we then find that

$$\limsup_{t\to\infty}\theta(\mathbf{x}^{(t)}) = \limsup_{t\to\infty}\max_\ell\langle\nabla f_\ell(\mathbf{x}^{(t)}),\mathbf{d}(\mathbf{x}^{(t)})\rangle \leq \max_\ell\langle\nabla f_\ell(\mathbf{x}),\mathbf{d}(\mathbf{x})\rangle = \theta(\mathbf{x}) \tag{A2}$$

Vice versa, we know that because of $\mathbf{d}(\mathbf{x}^{(t)}) \in \mathcal{X}-\mathbf{x}^{(t)}$, it holds that $\mathbf{d}(\mathbf{x}^{(t)})+\mathbf{x}^{(t)}-\mathbf{x} \in \mathcal{X}-\mathbf{x}$ and as above we find that

$$\max_\ell\langle\nabla f_\ell(\mathbf{x}),\mathbf{d}(\mathbf{x})\rangle \leq \lambda^{(t)}\max_\ell\langle\nabla f_\ell(\mathbf{x}),\mathbf{d}(\mathbf{x}^{(t)})\rangle+\lambda^{(t)}\max_\ell\langle\nabla f_\ell(\mathbf{x}),\mathbf{x}^{(t)}-\mathbf{x}\rangle \tag{A3}$$

with

$$\lambda^{(t)} := \begin{cases}\min\left\{1,\dfrac{1}{\left\|\mathbf{d}(\mathbf{x})+\mathbf{x}^{(t)}-\mathbf{x}\right\|}\right\} & \text{if } \mathbf{d}(\mathbf{x}) \neq \mathbf{x}^{(t)}-\mathbf{x},\\ 1 & \text{else.}\end{cases}$$

Again, the last term of (A3) vanishes in the limit so that by using the properties of the maximum function and the continuity of ∇f_ℓ, as well as $\lambda^{(t)} \xrightarrow{t\to\infty} 1$, in taking the limit inferior on (A3) we find that

$$\begin{aligned}
&\theta(\mathbf{x}) = \max_\ell\langle\nabla f_\ell(\mathbf{x}),\mathbf{d}(\mathbf{x})\rangle \leq \liminf_{t\to\infty}\max_\ell\langle\nabla f_\ell(\mathbf{x}),\mathbf{d}(\mathbf{x}^{(t)})\rangle\\
&\leq \liminf_{t\to\infty}\left[\left(\max_\ell\langle\nabla f_\ell(\mathbf{x}),\mathbf{d}(\mathbf{x}^{(t)})\rangle-\max_\ell\langle\nabla f_\ell(\mathbf{x}^{(t)}),\mathbf{d}(\mathbf{x}^{(t)})\rangle\right)+\max_\ell\langle\nabla f_\ell(\mathbf{x}^{(t)}),\mathbf{d}(\mathbf{x}^{(t)})\rangle\right]\\
&\leq \liminf_{t\to\infty}\left[\left\|\mathbf{Df}(\mathbf{x})-\mathbf{Df}(\mathbf{x}^{(t)})\right\|\left\|\mathbf{d}(\mathbf{x}^{(t)})\right\|+\max_\ell\langle\nabla f_\ell(\mathbf{x}^{(t)}),\mathbf{d}(\mathbf{x}^{(t)})\rangle\right]\\
&\leq \liminf_{t\to\infty}\max_\ell\langle\nabla f_\ell(\mathbf{x}^{(t)}),\mathbf{d}(\mathbf{x}^{(t)})\rangle = \liminf_{t\to\infty}\theta(\mathbf{x}^{(t)}).
\end{aligned} \tag{A4}$$

Combining (A2) and (A4) shows that $\theta(\mathbf{x}^{(t)}) \xrightarrow{t\to\infty} \theta(\mathbf{x})$. □

Theorem 2 claims that $\omega(\mathbf{x})$ is uniformly continuous, provided the objective gradients are Lipschitz. The implied Cauchy continuity is an important property in the convergence proof of the algorithm.

Proof of Theorem 2. We will consider the constrained case only, when $\mathcal{X}$ is convex and compact and show uniform continuity a fortiori by proving that $\omega(\bullet)$ is Lipschitz. Let the objective gradients be Lipschitz continuous. Then $\mathbf{Df}$ is Lipschitz as well with constant $L > 0$. Let $\mathbf{x},\mathbf{y} \in \mathcal{X}$ with $\mathbf{x} \neq \mathbf{y}$ (the other case is trivial) and let again $\mathbf{d}(\mathbf{x}),\mathbf{d}(\mathbf{y})$ be the respective optimizers.

Suppose w.l.o.g. that

$$\left|\max_\ell \langle \nabla f_\ell(\mathbf{x}), \mathbf{d}(\mathbf{x})\rangle - \max_\ell \langle \nabla f_\ell(\mathbf{y}), \mathbf{d}(\mathbf{y})\rangle\right| = \max_\ell \langle \nabla f_\ell(\mathbf{x}), \mathbf{d}(\mathbf{x})\rangle - \max_\ell \langle \nabla f_\ell(\mathbf{y}), \mathbf{d}(\mathbf{y})\rangle$$

If we define

$$(0,1] \ni \sigma := \begin{cases} \min\left\{1, \frac{1}{\|\mathbf{d}(\mathbf{y})+\mathbf{y}-\mathbf{x}\|}\right\} & \text{if } \mathbf{d}(\mathbf{y}) \neq \mathbf{x}-\mathbf{y}, \\ 1 & \text{else,} \end{cases}$$

then again $\sigma(\mathbf{d}(\mathbf{y}) + \mathbf{y} - \mathbf{x})$ is feasible for (P1) at $\mathbf{y}$. Thus,

$$\begin{aligned} &\max_\ell \langle \nabla f_\ell(\mathbf{x}), \mathbf{d}(\mathbf{x})\rangle - \max_\ell \langle \nabla f_\ell(\mathbf{y}), \mathbf{d}(\mathbf{y})\rangle \\ &\qquad \overset{\text{df.}}{\le} \max_\ell \langle \nabla f_\ell(\mathbf{x}), \sigma(\mathbf{d}(\mathbf{y}) + \mathbf{y} - \mathbf{x})\rangle - \max_\ell \langle \nabla f_\ell(\mathbf{y}), \mathbf{d}(\mathbf{y})\rangle \\ &\qquad \le \|\sigma \mathbf{Df}(\mathbf{x})(\mathbf{d}(\mathbf{y}) + \mathbf{y} - \mathbf{x}) - \mathbf{Df}(\mathbf{y})\mathbf{d}(\mathbf{y})\| \\ &\qquad \overset{\sigma\le 1}{\le} \|\sigma \mathbf{Df}(\mathbf{x}) - \mathbf{Df}(\mathbf{y})\| \|\mathbf{d}(\mathbf{y})\| + \|\mathbf{Df}(\mathbf{x})\| \|\mathbf{x}-\mathbf{y}\|, \end{aligned} \tag{A5}$$

where we have again used the maximum property 2 for the second inequality. We now investigate the first term on the RHS. Using $\|\mathbf{d}(\mathbf{y})\| \le 1$ and adding a zero, we find

$$\begin{aligned} \|\sigma \mathbf{Df}(\mathbf{x}) - \mathbf{Df}(\mathbf{y})\| \|\mathbf{d}(\mathbf{y})\| &\le \|\mathbf{Df}(\mathbf{x}) - \mathbf{Df}(\mathbf{y}) - (1-\sigma)\mathbf{Df}(\mathbf{x})\| \\ &\le L\|\mathbf{x}-\mathbf{y}\| + (1-\sigma)\|\mathbf{Df}(\mathbf{x})\|. \end{aligned} \tag{A6}$$

Furthermore, $\|\mathbf{d}(\mathbf{y}) + \mathbf{y} - \mathbf{x}\| \le 1 + \|\mathbf{y} - \mathbf{x}\|$ implies $1/(1 + \|\mathbf{y}-\mathbf{x}\|) \le \sigma$ and

$$1 - \sigma \le 1 - \frac{1}{1 + \|\mathbf{y}-\mathbf{x}\|} = \frac{\|\mathbf{y}-\mathbf{x}\|}{1 + \|\mathbf{y}-\mathbf{x}\|} \le \|\mathbf{y}-\mathbf{x}\|.$$

We use this inequality and plug (A6) into (A5) to obtain

$$\begin{aligned} \max_\ell \langle \nabla f_\ell(\mathbf{x}), \mathbf{d}(\mathbf{x})\rangle - \max_\ell \langle \nabla f_\ell(\mathbf{y}), \mathbf{d}(\mathbf{y})\rangle &\le L\|\mathbf{x}-\mathbf{y}\| + 2\|\mathbf{Df}(\mathbf{x})\| \|\mathbf{x}-\mathbf{y}\| \\ &\le (L + 2D)\|\mathbf{x}-\mathbf{y}\|, \end{aligned}$$

with $D = \max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{Df}(\mathbf{x})\|$ which is well-defined because $\mathcal{X}$ is compact and $\|\mathbf{Df}(\bullet)\|$ is continuous. □

Appendix A.2. Modified Criticality Measures

Proof of Lemma 5. There are two cases to consider:

- If $\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \ge \omega\left(\mathbf{x}^{(t)}\right)$ then

  $$\left|\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \omega\left(\mathbf{x}^{(t)}\right)\right| = \omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \omega\left(\mathbf{x}^{(t)}\right) \le \kappa_\omega \omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right).$$

  Now

  $$\left|\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \varpi\left(\mathbf{x}^{(t)}\right)\right| \in \left\{ \begin{array}{r} \omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \omega\left(\mathbf{x}^{(t)}\right) \\ 1 - \omega\left(\mathbf{x}^{(t)}\right) \le \omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) - \omega\left(\mathbf{x}^{(t)}\right) \\ 1 - 1 = 0 \end{array} \right\} \le \kappa_\omega \omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right).$$

- The case $\omega\left(\mathbf{x}^{(t)}\right) < \omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)$ can be shown similarly.

□

Proof of Lemma 6. Use Lemma 5 and then investigate the two possible cases:

- If $\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \geq \varpi\left(\mathbf{x}^{(t)}\right)$, then the first inequality follows because of $1 \geq 1/(1+\kappa_{\omega})$.
- If $\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) < \varpi\left(\mathbf{x}^{(t)}\right)$, then $\varpi\left(\mathbf{x}^{(t)}\right) - \varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \leq \kappa_{\omega}\varpi_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)$, and again the first inequality follows.

□

Appendix B. Pascoletti–Serafini Step

One example of an alternative descent step $\mathbf{s}^{(t)} \in \mathbb{R}^n$ is given in [33]. Thomann and Eichfelder [33] leverage the Pascoletti–Serafini scalarization to define local subproblems that guide the iterates towards the (local) model ideal point. To be precise, it is shown that the trial point $\mathbf{x}_+^{(t)}$ can be computed as the solution to

$$\min_{\tau\in\mathbb{R},\mathbf{x}\in B^{(t)}} \tau \quad \text{s.t. } \mathbf{m}^{(t)}(\mathbf{x}^{(t)}) + \tau\mathbf{r}^{(t)} - \mathbf{m}^{(t)}(\mathbf{x}) \geq \mathbf{0}, \tag{A7}$$

where $\mathbf{r}^{(t)} = \mathbf{m}^{(t)}(\mathbf{x}^{(t)}) - \mathbf{i}_{\mathrm{m}}^{(t)} \in \mathbb{R}^k_{\geq 0}$ is the direction vector pointing from the local model ideal point

$$\mathbf{i}_{\mathrm{m}}^{(t)} = \left[i_1^{(t)}, \dots, i_k^{(t)}\right]^T, \text{ with } i_\ell^{(t)} = \min_{\mathbf{x}\in\mathcal{X}} m_\ell^{(t)}(\mathbf{x}) \text{ for } \ell = 1, \dots, k, \tag{A8}$$

to the current iterate value. If the surrogates are linear or quadratic polynomials and the trust region use a p-norm with $p \in \{1, 2, \infty\}$ these sub-problems are linear or quadratic programs.

A convergence proof for the unconstrained case is given in [33]. It relies on a sufficient decrease bound similar to (20). However, it is not shown that $\kappa^{\mathrm{sd}} \in (0,1)$ exists independent of the iteration index t but stated as an assumption.

Furthermore, constraints (in particular box constraints) are integrated into the definition of $\omega(\bullet)$ and $\omega_{\mathrm{m}}^{(t)}(\bullet)$ using an active set strategy (see [38]). Consequently, both values are no longer Cauchy continuous. We can remedy both drawbacks by relating the (possibly constrained) Pascoletti–Serafini trial point to the strict modified Pareto–Cauchy point in our projection framework. To this end, we allow in (A7) and (A8) any feasible set fulfilling Assumption 1. Moreover, we recite the following assumption:

Assumption A1 (Assumption 4.10 in [33])**.** *There is a constant* $\mathrm{r} \in (0,1]$ *so that if* $\mathbf{x}^{(t)}$ *is not Pareto critical, the components* $r_1^{(t)}, \dots, r_k^{(t)}$*, of* $\mathbf{r}^{(t)}$ *satisfy* $\dfrac{\min_\ell r_\ell^{(t)}}{\max_\ell r_\ell^{(t)}} \geq \mathrm{r}$.

The assumption can be justified because $r_\ell^{(t)} > 0$ if $\mathbf{x}^{(t)}$ is not critical and $r_\ell^{(t)}$ can be bounded above and below by expressions involving $\omega_{\mathrm{m}}^{(t)}(\bullet)$, see Remark 4 and [33] (Lemma 4.9). We can then derive the following lemma:

Lemma A1. *Suppose Assumptions 1 and 2 and Appendix B hold. Let* $(\tau^+, \mathbf{x}_+^{(t)})$ *be the solution to* (A7)*. Then there exists a constant* $\tilde{\kappa}_{\mathrm{m}}^{\mathrm{sd}} \in (0,1)$ *such that it holds*

$$\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_+^{(t)}) \geq \tilde{\kappa}_{\mathrm{m}}^{\mathrm{sd}}\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right) \min\left\{ \frac{\omega_{\mathrm{m}}^{(t)}\left(\mathbf{x}^{(t)}\right)}{\mathrm{c}H_{\mathrm{m}}^{(t)}}, \Delta^{(t)}, 1 \right\}.$$

Proof. If $\mathbf{x}^{(t)}$ is critical for (MOPm), then $\tau^{+} = 0$ and $\mathbf{x}_{+}^{(t)} = \mathbf{x}^{(t)}$ and the bound is trivial [5]. Otherwise, we can use the same argumentation as in [33] ([Lemma 4.13]) to show that for the strict modified Pareto–Cauchy point $\hat{\mathbf{x}}_{\mathrm{PC}}^{(t)}$ it holds that

$$\Phi_{\mathrm{m}}^{(t)}(\mathbf{x}^{(t)}) - \Phi_{\mathrm{m}}^{(t)}(\mathbf{x}_{+}^{(t)}) \geq \mathrm{r} \min_{\ell} \left\{ m_{\ell}^{(t)}(\mathbf{x}^{(t)}) - m_{\ell}^{(t)}(\hat{\mathbf{x}}_{\mathrm{PC}}^{(t)}) \right\}$$

and the final bound follows from Corollary 2 with the new constant $\tilde{\kappa}_{\mathrm{m}}^{\mathrm{sd}} = \mathrm{r}\kappa_{\mathrm{m}}^{\mathrm{sd}}$. □

References

1. Ehrgott, M. *Multicriteria Optimization*, 2nd ed.; Springer: Berlin, Germany, 2005.
2. Jahn, J. *Vector Optimization: Theory, Applications, and Extensions*, 2nd ed.; Springer: Berlin, Germany, 2011; OCLC: 725378304.
3. Miettinen, K. *Nonlinear Multiobjective Optimization*; Springer: Berlin, Germany, 2013; OCLC: 1089790877.
4. Eichfelder, G. Twenty Years of Continuous Multiobjective Optimization. Available online: http://www.optimization-online.org/DB_FILE/2020/12/8161.pdf (accessed on 8 April 2021).
5. Eichfelder, G. *Adaptive Scalarization Methods in Multiobjective Optimization*; Springer: Berlin, Germany, 2008. [CrossRef]
6. Fukuda, E.H.; Drummond, L.M.G. A Survay on Multiobjective Descent Methods. *Pesqui. Oper.* **2014**, *34*, 585–620. [CrossRef]
7. Fliege, J.; Svaiter, B.F. Steepest descent methods for multicriteria optimization. *Math. Method. Operat. Res. (ZOR)* **2000**, *51*, 479–494. [CrossRef]
8. Graña Drummond, L.; Svaiter, B. A steepest descent method for vector optimization. *J. Comput. Appl. Math.* **2005**, *175*, 395–414. [CrossRef]
9. Lucambio Pérez, L.R.; Prudente, L.F. Nonlinear Conjugate Gradient Methods for Vector Optimization. *SIAM J. Optim.* **2018**, *28*, 2690–2720. [CrossRef]
10. Lucambio Pérez, L.R.; Prudente, L.F. A Wolfe Line Search Algorithm for Vector Optimization. *ACM Transact. Math. Softw.* **2019**, *45*, 1–23. [CrossRef]
11. Gebken, B.; Peitz, S.; Dellnitz, M. A Descent Method for Equality and Inequality Constrained Multiobjective Optimization Problems. In *Numerical and Evolutionary Optimization—NEO 2017*; Trujillo, L., Schütze, O., Maldonado, Y., Valle, P., Eds.; Springer: Cham, Switzerland, 2019; pp. 29–61.
12. Hillermeier, C. *Nonlinear Multiobjective Optimization: A Generalized Homotopy Approach*; Springer Basel AG: Basel, Switzerland, 2001; OCLC: 828735498.
13. Gebken, B.; Peitz, S.; Dellnitz, M. On the hierarchical structure of Pareto critical sets. *J. Glob. Optim.* **2019**, *73*, 891–913. [CrossRef]
14. Wilppu, O.; Karmitsa, N.; Mäkelä, M. *New Multiple Subgradient Descent Bundle Method for Nonsmooth Multiobjective Optimization*; Report no. 1126; Turku Centre for Computer Science: Turku, Sweden, 2014.
15. Gebken, B.; Peitz, S. An Efficient Descent Method for Locally Lipschitz Multiobjective Optimization Problems. *J. Optim. Theor. Appl.* **2021**. [CrossRef]
16. Custódio, A.L.; Madeira, J.F.A.; Vaz, A.I.F.; Vicente, L.N. Direct Multisearch for Multiobjective Optimization. *SIAM J. Optim.* **2011**, *21*, 1109–1140. [CrossRef]
17. Audet, C.; Savard, G.; Zghal, W. Multiobjective Optimization Through a Series of Single-Objective Formulations. *SIAM J. Optim.* **2008**, *19*, 188–210. [CrossRef]
18. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [CrossRef]
19. Deb, K. *Multi-Objective Optimization Using Evolutionary Algorithms*; Wiley: Hoboken, NJ, USA, 2001.
20. Coello, C.A.C.; Lamont, G.B.; Veldhuizen, D.A.V. *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2nd ed.; Springer: New York, NY, USA, 2007.
21. Abraham, A.; Jain, L.C.; Goldberg, R. (Eds.) Evolutionary multiobjective optimization: Theoretical advances and applications. In *Advanced Information and Knowledge Processing*; Springer: New York, NY, USA, 2005.
22. Zitzler, E. Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications. Ph.D. Thesis, ETH, Zurich, Switzerland, 1999.
23. Peitz, S.; Dellnitz, M. A Survey of Recent Trends in Multiobjective Optimal Control—Surrogate Models, Feedback Control and Objective Reduction. *Math. Comput. Appl.* **2018**, *23*, 30. [CrossRef]
24. Chugh, T.; Sindhya, K.; Hakanen, J.; Miettinen, K. A survey on handling computationally expensive multiobjective optimization problems with evolutionary algorithms. *Soft Comput.* **2019**, *23*, 3137–3166. [CrossRef]
25. Deb, K.; Roy, P.C.; Hussein, R. Surrogate Modeling Approaches for Multiobjective Optimization: Methods, Taxonomy, and Results. *Math. Comput. Appl.* **2020**, *26*, 5. [CrossRef]
26. Roy, P.C.; Hussein, R.; Blank, J.; Deb, K. Trust-Region Based Multi-objective Optimization for Low Budget Scenarios. In *Evolutionary Multi-Criterion Optimization*; Series Title: Lecture Notes in Computer Science; Deb, K., Goodman, E., Coello Coello, C.A., Klamroth, K., Miettinen, K., Mostaghim, S., Reed, P., Eds.; Springer International Publishing: Cham, Switzerland, 2019; Volume 11411; pp. 373–385. [CrossRef]

27. Conn, A.R.; Scheinberg, K.; Vicente, L.N. *Introduction to Derivative-Free Optimization*; Number 8 in MPS-SIAM Series on Optimization; Society for Industrial and Applied Mathematics/Mathematical Programming Society: Philadelphia, PA, USA, 2009; OCLC: Ocn244660709.
28. Larson, J.; Menickelly, M.; Wild, S.M. Derivative-free optimization methods. *arXiv* **2019**, arXiv:1904.11585.
29. Qu, S.; Goh, M.; Liang, B. Trust region methods for solving multiobjective optimisation. *Optim. Method. Softw.* **2013**, *28*, 796–811. [CrossRef]
30. Villacorta, K.D.V.; Oliveira, P.R.; Soubeyran, A. A Trust-Region Method for Unconstrained Multiobjective Problems with Applications in Satisficing Processes. *J. Optim. Theor. Appl.* **2014**, *160*, 865–889. [CrossRef]
31. Ryu, J.H.; Kim, S. A Derivative-Free Trust-Region Method for Biobjective Optimization. *SIAM J. Optim.* **2014**, *24*, 334–362. [CrossRef]
32. Audet, C.; Savard, G.; Zghal, W. A mesh adaptive direct search algorithm for multiobjective optimization. *Eur. J. Oper. Res.* **2010**, *204*, 545–556. [CrossRef]
33. Thomann, J.; Eichfelder, G. A Trust-Region Algorithm for Heterogeneous Multiobjective Optimization. *SIAM J. Optim.* **2019**, *29*, 1017–1047. [CrossRef]
34. Wild, S.M.; Regis, R.G.; Shoemaker, C.A. ORBIT: Optimization by Radial Basis Function Interpolation in Trust-Regions. *SIAM J. Sci. Comput.* **2008**, *30*, 3197–3219. [CrossRef]
35. Conn, A.R.; Scheinberg, K.; Vicente, L.N. Global Convergence of General Derivative-Free Trust-Region Algorithms to First- and Second-Order Critical Points. *SIAM J. Optim.* **2009**, *20*, 387–415. [CrossRef]
36. Conn, A.R.; Gould, N.I.M.; Toint, P.L. *Trust-Region Methods*; MPS-SIAM series on optimization; Society for Industrial and Applied Mathematics: Harrisburg, PA, USA, 2000.
37. Luc, D.T. *Theory of Vector Optimization*; Lecture Notes in Economics and Mathematical Systems; Springer: Berlin, Heidelberg, 1989; Volume 319. [CrossRef]
38. Thomann, J. A Trust Region Approach for Multi-Objective Heterogeneous Optimization. Ph.D. Thesis, TU Ilmenau, Illmenau, Germany, 2018.
39. Nocedal, J.; Wright, S.J. *Numerical Optimization*, 2nd ed.; Springer Series in Operations Research; Springer: Berlin, Germany, 2006; OCLC: Ocm68629100.
40. Wendland, H. *Scattered Data Approximation*, 1st ed.; Cambridge University Press: Cambridge, UK, 2004. [CrossRef]
41. Wild, S.M. *Derivative-Free Optimization Algorithms for Computationally Expensive Functions*; Cornell University: Ithaca, NY, USA, 2009.
42. Wild, S.M.; Shoemaker, C. Global Convergence of Radial Basis Function Trust Region Derivative-Free Algorithms. *SIAM J. Optim.* **2011**, *21*, 761–781. [CrossRef]
43. Regis, R.G.; Wild, S.M. CONORBIT: Constrained optimization by radial basis function interpolation in trust regions. *Optim. Methods Softw.* **2017**, *32*, 552–580. [CrossRef]
44. Fleming, W. *Functions of Several Variables*; Undergraduate Texts in Mathematics; Springer: New York, NY, USA, 1977. [CrossRef]
45. Stellato, B.; Banjac, G.; Goulart, P.; Bemporad, A.; Boyd, S. OSQP: An operator splitting solver for quadratic programs. *Math. Program. Comput.* **2020**, *12*, 637–672. [CrossRef]
46. Johnson, S.G. The NLopt Nonlinear-Optimization Package. Available online: https://nlopt.readthedocs.io/en/latest/ (accessed on 8 April 2021).
47. Svanberg, K. A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM J. Optim.* **2002**, *12*, 555–573. [CrossRef]
48. Legat, B.; Timme, S.; Weisser, T.; Kapelevich, L.; Rackauckas, C.; TagBot, J. JuliaAlgebra/DynamicPolynomials.jl: V0.3.15. 2020. Available online: https://zenodo.org/record/4153432#.YG5wjj8RVPY (accessed on 8 April 2021).
49. Runarsson, T.P.; Yao, X. Search biases in constrained evolutionary optimization. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* **2005**, *35*, 233–243. [CrossRef]
50. Revels, J.; Lubin, M.; Papamarkou, T. Forward-Mode Automatic Differentiation in Julia. *arXiv* **2016**, arXiv:1607.07892.
51. Zitzler, E.; Deb, K.; Thiele, L. Comparison of Multiobjective Evolutionary Algorithms: Empirical Results. *Evol. Comput.* **2000**, *8*, 173–195. [CrossRef]
52. Deb, K.; Thiele, L.; Laumanns, M.; Zitzler, E. Scalable Test Problems for Evolutionary Multiobjective Optimization. In *Evolutionary Multiobjective Optimization*; Series Title: Advanced Information and Knowledge Processing; Abraham, A., Jain, L., Goldberg, R., Eds.; Springer: London, UK, 2005; pp. 105–145. [CrossRef]
53. Powell, M.J. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in Optimization and Numerical Analysis*; Gomez, S., Hennart, J.P., Eds.; Springer: Dordrecht, The Netherlands, 1994; pp. 51–67.
54. Prinz, S.; Thomann, J.; Eichfelder, G.; Boeck, T.; Schumacher, J. Expensive multi-objective optimization of electromagnetic mixing in a liquid metal. *Optim. Eng.* **2020**. [CrossRef]
55. Thomann, J.; Eichfelder, G. Representation of the Pareto front for heterogeneous multi-objective optimization. *J. Appl. Numer. Optim.* **2019**, *1*, 293–323.
56. Deshpande, S.; Watson, L.T.; Canfield, R.A. Multiobjective optimization using an adaptive weighting scheme. *Optim. Methods Softw.* **2016**, *31*, 110–133. [CrossRef]

57. Regis, R.G. Multi-objective constrained black-box optimization using radial basis function surrogates. *J. Comput. Sci.* **2016**, *16*, 140–155. [CrossRef]
58. Schütze, O.; Cuate, O.; Martín, A.; Peitz, S.; Dellnitz, M. Pareto Explorer: A global/local exploration tool for many-objective optimization problems. *Eng. Optim.* **2020**, *52*, 832–855. [CrossRef]