Gradient-Based/Evolutionary Relay Hybrid for Computing Pareto Front Approximations Maximizing the S-Metric

Michael Emmerich[1], André Deutz[1], and Nicola Beume[2]

[1]University of Leiden, Leiden Institute for Advanced Computer Science,
2333 CA Leiden, The Netherlands
http://natcomp.liacs.nl/
[2]University of Dortmund, Chair of Algorithm Engineering,
44221 Dortmund, Germany
{emmerich,deutz}@liacs.nl, nicola.beume@udo.edu,
http://Ls11-www.cs.uni-dortmund.de

Abstract. The problem of computing a good approximation set of the Pareto front of a multiobjective optimization problem can be recasted as the maximization of its S-metric value, which measures the dominated hypervolume. In this way, the S-metric has recently been applied in a variety of metaheuristics. In this work, a novel high-precision method for computing approximation sets of a Pareto front with maximal S-Metric is proposed as a high-level relay hybrid of an evolutionary algorithm and a gradient method, both guided by the S-metric. First, an evolutionary multiobjective optimizer moves the initial population close to the Pareto front. The gradient-based method takes this population as its starting point for computing a local maximal approximation set with respect to the S-metric. Thereby, the population is moved according to the gradient of the S-metric.

This paper introduces expressions for computing the gradient of a set of points with respect to its S-metric on basis of the gradients of the objective functions. It discusses singularities where the gradient is vanishing or differentiability is one sided. To circumvent the problem of vanishing gradient components of the S-metric for dominated points in the population a penalty approach is introduced.

In order to test the new hybrid algorithm, we compute the precise maximizer of the S-metric for a generalized Schaffer problem and show, empirically, that the relay hybrid strategy linearly converges to the precise optimum. In addition we provide first case studies of the hybrid method on complicated benchmark problems.

1 Introduction and Mathematical Preliminaries

In multiobjective optimization, a solution has to fulfill several objectives in the best possible way. Maximization problems can be reformulated as minimization problems, thus, without loss of generality, we can restrict our attention to those. Formally, the problem reads as follows:

$$\mathbf{f} = (f_1, \ldots, f_m)^T, \quad f_1(\mathbf{x}) \to \min, \ldots, f_m(\mathbf{x}) \to \min, \quad \mathbf{x} \in \mathcal{X}. \tag{1}$$

T. Bartz-Beielstein et al. (Eds.): HM 2007, LNCS 4771, pp. 140–156, 2007.

The domain $\mathcal{X}$ is called the *decision space* or *search space* and contains all feasible solutions, and the co-domain $\mathcal{Y}$ of all m objectives is called *objective space*. Here we assume continuous functions, so $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}^m$.

Since the objectives are typically conflicting, there is no single best solution and the aim is to generate sets of good compromise solutions. These solutions are suggestions to the decision maker who finally chooses one for realization.

A partial order holds among the points, which is defined in the objective space and is transferred to the preimages in the search space. A point $\mathbf{x}$ is said to (weakly) dominate a point $\mathbf{x}'$ ($\mathbf{x} \prec \mathbf{x}'$), iff $\mathbf{f}(\mathbf{x}) \neq \mathbf{f}(\mathbf{x}')$ and $\forall i \in \{1, \ldots, m\} : f_i(\mathbf{x}) \leq f_i(\mathbf{x}')$. A point $\mathbf{x}$ strictly dominates a point $\mathbf{x}'$ ($\mathbf{x} < \mathbf{x}'$), iff $\forall i = 1, \ldots, m : f_i(\mathbf{x}) < f_i(\mathbf{x}')$. The points that are minimal with respect to the partial order $\prec$ within a set are called non-dominated. The non-dominated points within the whole search space are called *efficient set or Pareto set* $\mathcal{X}_E$ and the set of their corresponding images is called *Pareto front* $\mathcal{Y}_N$.

Since continuous problems cannot be expected to be solved optimally, a good approximation of the Pareto front is aspired. Two sets are already incomparable, if one set contains a point that is incomparable to each point of the other set. Thus, a qualitative ranking is mostly impossible. Instead, auxiliary demands which suggest high quality are formulated for sets, such as: (1) many non-dominated points, (2) closeness to Pareto front, and (3) well-distributed along the Pareto front. From our point of view the term well-distributed means to have a regular spacing between points in regions with similar trade-off and a higher concentration of points in regions with a more balanced trade-off among the objectives.

Among the developed quality measures, the *S-metric* or *dominated hypervolume* by Zitzler and Thiele [1] is of utmost importance. It is defined as

$$\mathcal{S}(X) = \text{Lebesgue}\{\mathbf{y} \mid \exists \mathbf{y}^{(i)} : \ \mathbf{y}^{(i)} \prec \mathbf{y} \ \wedge \ \mathbf{y} \prec \mathbf{y}^{ref}\}, \tag{2}$$

where $\mathbf{y}^{(i)} = \mathbf{f}(\mathbf{x}^{(i)})$ are the image points of the set $X \subseteq \mathcal{X}$ under $\mathbf{f}$, and X is an approximation of the Pareto set. The reference point $\mathbf{y}^{ref}$ confines the dominated hypervolume. Note that the same definition can be used to define the S-metric for subsets of $\mathbb{R}^m$ directly. It is an alleged drawback that the reference point influences the absolute value of the metric. However, in practical settings it is often possible to state bounds for the objective function values and thus the reference point can be chosen as that upper bound vector. In addition, recent results on generalizations of the S-metric show, that the distribution of points on the Pareto front can be influenced by weighting parameters according to the user's preferences [2].

The maximal S-metric value is achieved by the Pareto front. For compact image sets of $\mathbf{f}$ and appropriately chosen reference points the maximization of the S-metric for a given number of points always results in a non-dominated set of solutions. Further properties of the S-metric were studied by Fleischer [3] and Zitzler et al. [4].

The maximization of the S-metric receives increasingly more attention as a solution principle for approximating Pareto fronts by means of a well-distributed

non-dominated set. Recently, the S-metric has been used as a single-objective substitute function to guide the process of multiobjective optimizers. Accordingly, the problem of finding a good approximation of the Pareto front of the original multiobjective optimization problem can be re-stated as:

$$\mathcal{S}(X) \rightarrow \max, X \subseteq_{\mu} \mathcal{X} \tag{3}$$

where $X \subseteq_{\mu} \mathcal{X}$ means that X is a set of at most μ elements from $\mathcal{X}$.

Recent work proposed methods for S-metric maximization that are based on simulated annealing, particle swarms, and evolutionary algorithms. *Evolutionary multiobjective optimization algorithms (EMOA, MOEA)* [5,6] established as efficient and robust optimizers and modern EMOA like IBEA [7], ESP [8], and SMS-EMOA successfully apply an S-metric based function to evaluate and select promising solutions, or use it for archiving [9]. The SMS-EMOA by Emmerich et al. [10] uses the S-metric in the selection method of a steady-state EMOA. It has been tested extensively on benchmarks and real-world applications, receiving results competitive or better than state-of-the-art methods in the field. In this paper we continue in the same spirit, and derive a gradient based method for solving multiobjective problems.

In this work the gradient of the S-metric at a point, representing an approximation set, is introduced to solve the optimization problem of positioning the given μ points of the set such that the S-metric value of the set is maximized. Using this gradient, we apply a simple steepest ascent method. We propose a hybridization of the gradient method with SMS-EMOA as a high-level relay (cf. Talbi [11]), meaning that autonomous algorithms are executed sequentially. The gradient method is applied after SMS-EMOA to locally optimize its final population. Thus, we combine efficient local optimization based on a new gradient-based method, with more exhaustive global optimization techniques.

As opposed to previous work on gradient based multiobjective optimization (e.g. [12,13,14,15,16]) this approach does not use gradients to improve points of the population independent of each other but, by aiming at improving the S-metric (that considers the population as one aggregate), it looks at the distribution of the entire population.

The paper is structured as follows. In Section 2 expressions of the gradient of the S-metric are derived and discussed. In Section 3 the maximal S-metric is determined analytically to verify the gradient formulation. Section 4 introduces a steepest descent gradient method for S-metric maximization. Afterward, the hybridization of this method with the evolutionary algorithm SMS-EMOA is proposed and studied on multimodal test problems (Section 5). The new methods form starting points for further studies. A summary of the results and discussion of open questions is provided in Section 6.

2 Gradient of the S-Metric

In this section we discuss expressions for gradient computation with respect to the S-metric and discuss its differentiability properties.

2.1 Mathematical Notation

In order to compute gradients of the S-metric, we represent a population P of size μ, $P \subseteq_\mu \mathcal{X}$, as a vector of length $\mu \cdot d$:

$$\mathbf{p} = (x_1^{(1)}, \ldots, x_d^{(1)}, \ldots, x_1^{(\mu)}, \ldots, x_d^{(\mu)})^\top = (p_1, \ldots, p_{\mu \cdot d})^\top.$$

For notational convenience we introduce *blocks* of a μd-vector as

$$\Pi(i, \mathbf{p}) = (x_1^{(i)}, \ldots, x_d^{(i)}) = (p_{(i-1)\cdot d+1} \ldots p_{i \cdot d}).$$

The mapping from μd-vectors to populations is defined as:

$$\Psi(\mathbf{p}) = \{\ (x_1^{(i)}, \ldots, x_d^{(i)})^\top \mid i \in \{1, \ldots, \mu\}\ \}. \tag{4}$$

Different μd-vectors may represent the same population (but not vice-versa). Every non-empty population $P \subseteq_\mu \mathcal{X}$ is represented by at least one tuple of the form above.

For optimization purposes it is sufficient to work with μd-vectors. This holds, because the set of global optima of the problem

$$\mathcal{S}(\Psi(\mathbf{p})) \to \max, \text{ subject to } \Psi(\mathbf{p}) \subseteq_\mu \mathcal{X},\ \mathbf{p} \in \mathbb{R}^{\mu d} \tag{5}$$

can be mapped to the set of global optima of the original problem (Eq. 3) via Ψ. Note that for $\mathcal{X} = \mathbb{R}^d$ the constraint $\Psi(\mathbf{p}) \subseteq_\mu \mathcal{X}$ is trivially fulfilled. Moreover, the number of local optima of the new problem is usually increased, as different μd-vectors may give rise to the same population. Given one μd-vector, all equivalent representations can be obtained by permuting its blocks.

2.2 Definition and Analytical Calculation of S-Metric's Gradient

A general definition of the gradient for the space of μd-vectors is

$$\nabla_{\mathbf{p}} \mathcal{S} = \left(\frac{\partial \mathcal{S}}{\partial p_1}, \ldots, \frac{\partial \mathcal{S}}{\partial p_{\mu \cdot d}}\right)^\top \tag{6}$$

In order to express the gradient of the S-metric in terms of the gradients of the objective functions the following structure of the composition of mappings is applied:

$$\mathbb{R}^{\mu \cdot d} \underbrace{\xrightarrow{\mathbf{F}}}_{\text{decision to objective space}} \mathbb{R}^{\mu \cdot m} \underbrace{\xrightarrow{\mathcal{S}}}_{\text{objective space to S-metric}} \mathbb{R}^+. \tag{7}$$

where $\mathbf{F}$ is defined by using the objective functions $\mathbf{f} = (f_1, f_2, \cdots, f_m)^\top$ so that $\mathbf{F}(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(\mu)}) = (\mathbf{f}(\mathbf{x}^{(1)}), \mathbf{f}(\mathbf{x}^{(2)}), \cdots, \mathbf{f}(\mathbf{x}^{(\mu)}))^\top$ with the functions f_i defined as above and $\mathcal{S}$ as the S-metric function.

The S-metric is defined on sets of points (Eq. 2), but for notational convenience, we also apply it directly to vectors which can be interpreted as sets

according to the mapping Ψ (Eq. 5). Using the chain rule the gradient can be rewritten as follows. Let $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(\mu)}$ be μ points in the decision space, then $\nabla\mathcal{S}(\mathbf{p})$ can be written as:

$$\mathcal{S}' \text{ at } \begin{pmatrix} \mathbf{f}(\mathbf{x^{(1)}}) \\ \mathbf{f}(\mathbf{x^{(2)}}) \\ \cdots \\ \mathbf{f}(\mathbf{x}^{(\mu)}) \end{pmatrix} \circ \begin{pmatrix} \mathbf{f}' \text{ at } \mathbf{x}^{(1)} & 0 & 0 \cdots & 0 \\ 0 & \mathbf{f}' \text{ at } \mathbf{x}^{(2)} & 0 \cdots & 0 \\ \vdots & \vdots & \vdots \cdots & \vdots \\ 0 & 0 & 0\ 0 & \mathbf{f}' \text{ at } \mathbf{x}^{(\mu)} \end{pmatrix} \tag{8}$$

The top level structure of the matrix associated to the linear mapping $\mathbf{F}'$ is a diagonal matrix of size μ whose diagonal elements are matrices of size $m \times d$ associated to the linear maps $\mathbf{f}'$ at $\mathbf{x}^{(j)}$, where $j = 1, 2, \cdots, \mu$ and each of the off-diagonal elements is the zero matrix of size $m \times d$ as well.

A more detailed description of this matrix is given as:

$$\underbrace{\begin{pmatrix} \frac{\partial\mathcal{S}}{\partial y_1^{(1)}} \\ \vdots \\ \frac{\partial\mathcal{S}}{\partial y_m^{(1)}} \\ \vdots \\ \frac{\partial\mathcal{S}}{\partial y_1^{(\mu)}} \\ \vdots \\ \frac{\partial\mathcal{S}}{\partial y_m^{(\mu)}} \end{pmatrix}^{\top}}_{\nabla\mathcal{S}(\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(\mu)})} \cdot \underbrace{\begin{pmatrix} \frac{\partial f_1(\mathbf{x}^{(1)})}{\partial x_1^{(1)}} & \cdots & \frac{\partial f_1(\mathbf{x}^{(1)})}{\partial x_d^{(1)}} & 0 \cdots 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots\ \vdots\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_m(\mathbf{x}^{(1)})}{\partial x_1^{(1)}} & \cdots & \frac{\partial f_m(\mathbf{x}^{(1)})}{\partial x_d^{(1)}} & 0 \cdots 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \vdots \cdots \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots\ \vdots\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \vdots \cdots \vdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 \cdots 0 & \frac{\partial f_1(\mathbf{x}^{(\mu)})}{\partial x_1^{(\mu)}} & \cdots & \frac{\partial f_1(\mathbf{x}^{(\mu)})}{\partial x_d^{(\mu)}} \\ \vdots & \vdots & \vdots & \vdots\ \vdots\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \cdots 0 & \frac{\partial f_m(\mathbf{x}^{(\mu)})}{\partial x_1^{(\mu)}} & \cdots & \frac{\partial f_m(\mathbf{x}^{(\mu)})}{\partial x_d^{(\mu)}} \end{pmatrix}}_{\mathbf{F}'(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(\mu)})} \tag{9}$$

Note that $\mathbf{F}'(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\mu)})$ depends solely on the gradient functions ∇f_i at the sites $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\mu)}$. Hence, if these $m \cdot \mu$ local gradients are known, the desired gradient $\nabla\mathcal{S}(\mathbf{p})$ can be computed.

The computation of $\nabla\mathcal{S}(\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(\mu)})$ is discussed next. Three cases of the set $\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(\mu)}\}$ need to be considered: (1) mutually non-dominated sets, (2) sets with strictly dominated points, and (3) sets with weakly dominated points.

(1) Mutually non-dominated sets. For $m = 1$ holds $\frac{\partial\mathcal{S}}{\partial y_1^{(i)}} = 1$, and for $m = 2$ holds (assuming vectors are sorted $\mathbf{y}^{(i)}$ in descending order of f):

$$\frac{\partial\mathcal{S}}{\partial y_1^{(i)}} = y_2^{(i-1)} - y_2^{(i)} \quad \text{and} \quad \frac{\partial\mathcal{S}}{\partial y_2^{(i)}} = y_1^{(i-1)} - y_1^{(i)}, \quad i = 1, \ldots, \mu \tag{10}$$

as illustrated in Fig. 1. Note that extremal points need special treatment, as their contribution to the gradient is influenced by the reference point. In three dimensions ($m = 3$), the computation of the partial derivative gets more tedious. The general principle is sketched in Fig. 2.

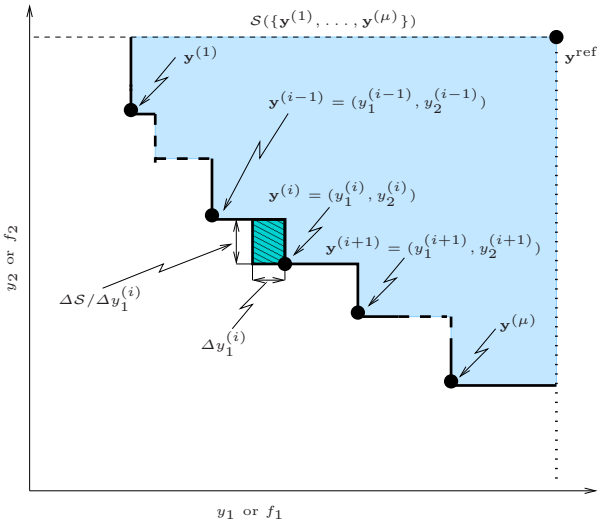

Fig. 1. Partial Derivative of the S-metric for $m = 2$ and non-dominated sets. The lengths of the line-segments of the attainment curve correspond to the values of the partial derivatives of $\mathcal{S}$. Only for extremal points do the values of the partial derivatives depend on the reference point.

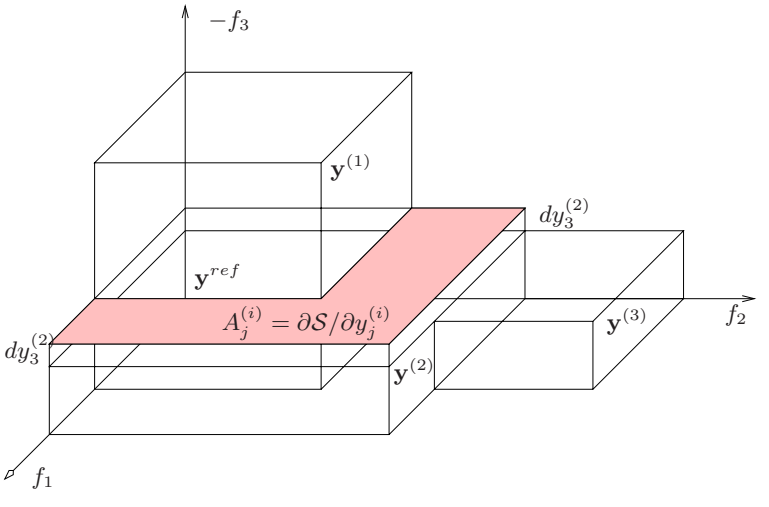

Fig. 2. Partial derivative for $m = 3$. By changing a point $\mathbf{y}^{(\mathbf{i})}$ differentially in the j-th coordinate direction, the hypervolume grows with the area $A_j^{(i)}$ of the 'visible' face of the exclusively contributed hypervolume of that point in the direction of the movement. Hence $A_j^{(i)}$ is the partial derivative $\partial\mathcal{S}/\partial y_j^{(i)}$.

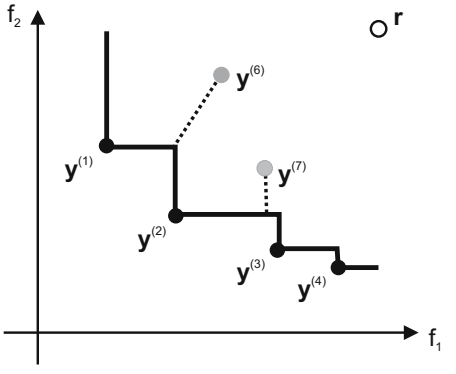

Fig. 3. The penalty function is defined as the sum of the euclidean distance (dashed lines) of the dominated points (gray) to the attainment curve (solid line) shaped by the non-dominated points (black) and bounded by the reference point **r**. The penalty is subtracted from the S-metric value to give an influence to the dominated points.

(2) Sets with strictly dominated points. The gradient equals zero in case of dominated points—provided that a slight perturbation does not make them non-dominated—since no improvement of the S-metric can be observed for any movement. Therefore, dominated points do not move during a search with gradient methods but just remain in their position. To enable an improvement of dominated points, a *penalty value* can be subtracted from the S-metric value, that is negative if and only if points are dominated and otherwise zero. For each dominated point, the minimal Euclidean distance to the attainment surface shaped by the non-dominated points is calculated (Fig. 3). The sum of these values is subtracted from the S-metric value of the whole set of points. This way, the movement of dominated points influences the improvement of the penalized S-metric and a local gradient of the dominated points is computed that points in the direction of the nearest point on the attainment curve. In a gradient descent method the movement of the non-dominated points is delayed by the dominated ones. Anyway, this drawback is a smaller deficit than completely losing the dominated points. Since any non-dominated point contributes to the S-metric value, the primary aim is to make all points non-dominated.

(3) Sets with weakly dominated points. Points that are dominated but not strictly dominated (we call them *weakly dominated*) lie on the attainment surface of the non-dominated points. Slight movements can make the points either remain weakly dominated, become strictly dominated or non-dominated. Thus, the gradient at these points is not continuous. The left-sided derivative $\frac{\partial^- \mathcal{S}}{y_j^{(i)}}$ may be positive, while the right-sided derivative $\frac{\partial^+ \mathcal{S}}{y_j^{(i)}}$ is always zero. For $m = 2$ positive one-sided derivatives can be determined as the length of the segment of the attainment curve. Let $\mathbf{y}^{(i_L)}$ determine the neighbor of the weakly dominated point

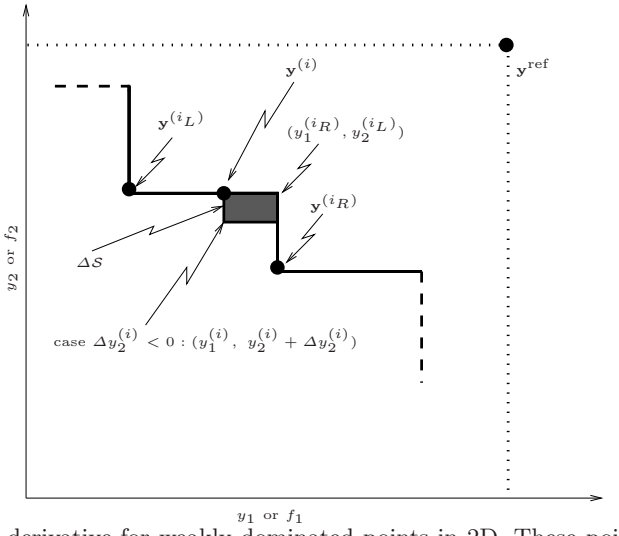

Fig. 4. Partial derivative for weakly dominated points in 2D. These points are dominated but not strictly dominated.

$\mathbf{y}^{(i)}$ on the upper left corner of the attainment curve, and $\mathbf{y}^{(i_R)}$ the neighbor on the lower right corner (see Fig. 4). If the point $\mathbf{y}^{(i)}$ lies on the segment $\mathbf{y}^{(i_L)}$ to $(y_1^{(i_R)}, y_2^{(i_L)})^\top$, then $\frac{\partial^+\mathcal{S}}{\partial y_1^{(i)}} = 0$ and $\frac{\partial^-\mathcal{S}}{\partial y_2^{(i)}} = y_1^{(i_R)} - y_1^{(i)}$ (see also Fig. 4); else if the point lies on the segment $\mathbf{y}^{(i_R)}$ to $(y_1^{(i_R)}, y_2^{(i_L)})^\top$, then $\frac{\partial^-\mathcal{S}}{\partial y_1^{(i)}} = y_2^{(i_L)} - y_2^{(i)}$ and $\frac{\partial^+\mathcal{S}}{\partial y_2^{(i)}} = 0$. The fact that $\mathcal{S}(\mathbf{p})$ is in general not continuously differentiable at weakly dominated points makes it problematic to work with gradient-based methods that make use of second order derivatives.

Weakly dominated points can also cause non-dominated points to have discontinuous local derivatives, which is comprehensible by arguments similar to the ones above. Besides degenerate points in the search space can cause discontinuous derivatives. These are, loosely defined, search points (or blocks) with the same image.

2.3 Empirical Determination of the Gradient

In practice the computation of the gradient can be approximated for example by using numerical differentiation. Since weakly non-dominated points of the population are not continuously differentiable, we need to take one-sided derivatives in both directions into account. For a small positive ϵ we compute them via:

$$\frac{\partial\mathcal{S}}{\partial p_i} \approx \frac{\mathcal{S}((p_1, \ldots, p_i \pm s\epsilon, \ldots, p_{\mu d})^\top) \pm \mathcal{S}((p_1, \ldots, p_i, \ldots, p_{\mu d})^\top)}{\epsilon} \tag{11}$$

The algebraic signs we need to use depend on the gradients of the objective function. In case of continuously differentiable objective functions, it is numerically

safer to compute the derivatives of the objective functions first, and then use the chain rule to compute the derivatives of the S-metric taking special care of weakly non-dominated points whenever they occur. Both the computation of Equation 11 and the computation of the gradients of all objective functions at all points (that can be used to compute the gradient via the chain rule) requires μd evaluations of the objective function vectors.

3 Analytical Solution of S-Metric Maximization

We exemplarily verify the maximization of the S-metric with the gradient by an analytical calculation for a problem with a linear Pareto front $\{(y_1, y_2) \mid y_2 = 1 - y_1 \text{ and } y_1 \in [0,1]\}$ and a fixed number of points. Using analytical arguments and partial derivatives, the optimal positions of the points are calculated. Later we will use this problem and its solution for testing the local convergence behavior of the gradient-based method.

Due to the monotonicity of the S-metric the μ points of the approximation set that maximizes $\mathcal{S}$ lie on the Pareto front. In order to consider the hypervolume of the approximation set we fix $(1,1)$ as the reference point and we consider $\mu + 2$ points on this Pareto curve whose y_1-coordinates we denote by u_i, with $i = 0, \ldots, n+1$. For any such collection of $n+2$ points we always require $u_0 = 0$ and $u_{n+1} = 1$. We want to maximize the hypervolume with respect to $(1,1)$. This is equivalent to minimizing the sum of the area of the triangles which are bounded by the Pareto curve and the sides of the rectangles shaping the attainment curve. Let v_i denote the length of the interval between u_i and u_{i+1}, then $\sum_{i=1}^{v+1} v_i^2$ is twice the area we want to minimize under the constraints $\sum_{i=1}^{n+1} v_i = 1$ and $\forall i : 0 \leq v_i$. This area is minimal in case the $n+2$ points are uniformly distributed (with the understanding that two of the points are the end points). It is easy and worthwhile to prove this fact geometrically, yet we revert to an analytical verification as follows. Let $g := \sum_{i=1}^{n+1} v_i^2$. Incorporating the constraint $v_{n+1} = 1 - \sum_{i=1}^{n} v_i$ yields $g = \sum_{i=1}^{n} v_i^2 + (1 - \sum_{i=1}^{n} v_i)^2$. Computing the partial derivatives of g results in $\frac{\partial g}{\partial v_j} = 2v_j - 2(1 - \sum_{i=1}^{n} v_i)$ where $j = 1, \ldots n$. Each of these partial derivatives has a value of zero at $v_1 = \frac{1}{n+1}, \ldots, v_n = \frac{1}{n+1}$ and at this point the minimum occurs. Translations back to the original problem result in $v_1 = \frac{1}{n+1}, \ldots, v_n = \frac{1}{n+1}$ and $v_{n+1} = \frac{1}{n+1}$. Hence, the points maximizing the S-metric are equidistant (with two occupying the end points).

Note that by approximating the Pareto front $\{(y_1, y_2) \mid y_i \in \mathbb{R} \text{ with } 0 \leq y_1 \leq 1 \text{ and } y_2 = 1 - y_2\}$ with a set consisting of μ points plus two extremal points $(0,1), (1,0)$ the maximal S-metric is $\frac{1}{2} \cdot \frac{\mu}{\mu+1}$. Moreover this maximum value can only be attained if the μ non-extremal points are equally spaced between the two extremal points.

With the generalized Schaffer problem Emmerich and Deutz [17] proposed a scalable-dimension problem that gives rise to the discussed linear Pareto front $\{(y, 1-y) \mid y \in [0,1]\}$ for $\alpha = 0.5$: $f_1(x) = \frac{1}{d^\alpha}(\sum_{i=1}^{d} x_i^2)^\alpha \rightarrow \min$ and

$f_2(x) = \frac{1}{d^\alpha}(\sum_{i=1}^{d}(1-x_i)^2)^\alpha \rightarrow \min$ for $x_i \in \mathbb{R}_+$, where $i = 1, ..., d$. In the following section, this problem and its solution set are consulted for a proof of concept result for the numerical optimization routines.

4 Gradient-Based Pareto Optimization

Due to the known problems with second-order gradient methods, which require twice continuous differentiability, a first-order gradient method, namely the steepest descent/ascent method with backtracking line search has been implemented [18]. The pseudo-code of our implementation is provided in Algorithm 1. The line-search algorithm has been kept simple to maintain transparency of the search process. It will however converge to a local maximizer relative to the line search direction. Note, that the line search may move to the same point in two subsequent iterations. In this case the evaluation of the objective function vectors of the population can be omitted. The convergence speed and accuracy of the line search can be controlled with the parameters τ and α_{min}, respectively. We recommend a setting of $\tau = 0.1$, while the setting of α_{min} depends on the problem. Since the length of the gradient decreases when the algorithm converges to the optimum of a differentiable function, α_{min} does not have to be very low, because the length of the gradient influences the step-size as well.

Algorithm 1. Gradient-ascent S-metric maximization

input variables: initial population as μd vector $\mathbf{p}$
control variables: accuracy of line search α_{min}, step reduction rate $\tau \in (0, 1)$
$\alpha \leftarrow 1$ {Initialize step size α}
$i \leftarrow 0; \mathbf{p}^{best} \leftarrow \mathbf{p}^0$
$\mathbf{d}^{(0)} \leftarrow \nabla\mathcal{S}(\mathbf{p}^{best})$ {Initialize search direction}
while $|\mathbf{d}^{(i)}| > \epsilon$ {Gradient larger than ϵ} **do**
  $\alpha \leftarrow 1$
  while $\alpha > \alpha_{min}$ {Line search in gradient direction} **do**
    $\mathbf{p}^{new} \leftarrow \mathbf{p}^{best} + \alpha\mathbf{d}^{(i)}$ {Try positive direction}
    if $\mathcal{S}(\Psi(\mathbf{p}^{best})) \geq \mathcal{S}(\Psi(\mathbf{p}^{new}))$ **then**
      $\mathbf{p}^{new} \leftarrow \mathbf{p}^{best} - \alpha\mathbf{d}^{(i)}$ {Try negative direction}
      if $\mathcal{S}(\Psi(\mathbf{p}^{best})) \geq \mathcal{S}(\Psi(\mathbf{p}^{new}))$ {No success with both moves} **then**
        $\alpha \leftarrow \alpha \cdot \tau$ {Reduce step size α}
        $\mathbf{p}^{new} \leftarrow \mathbf{p}^{best}$ {New current best point is old current best point}
      end if
    end if
    $\mathbf{p}^{best} \leftarrow \mathbf{p}^{new}$
  end while
  $\mathbf{d}^{(i+1)} \leftarrow \nabla\mathcal{S}(\mathbf{p}^{new})$, $i \leftarrow i + 1$ {Compute new gradient direction}
end while
return $\mathbf{p}^{best}$

5 SMS-EMOA-Gradient Hybrid

The gradient-descent method requires a good starting point in order to converge to the Pareto front. For this purpose an EMOA is applied which generates a good approximation of the Pareto front. We propose the SMS-EMOA because it has shown excellent results concerning the optimization of test functions and real-world problems (cf. [10,19,20]). The SMS-EMOA uses a steady-state selection scheme, i.e. in each generation one new solution is generated and one solution is discarded. A population of μ individuals is optimized without additional archives (which are often used in other EMOA). The S-metric is used within the selection operator to determine the subset of μ individuals with the highest S-metric value. Thereby, the individual with the least exclusive contribution of dominated hypervolume is discarded. As mentioned in Section 1, the maximization of the S-metric results in a well-distributed solution set with an emphasis of solutions in regions with fair trade-offs. The SMS-EMOA's final population functions as the starting point of the gradient strategy which does only a fine-tuning of the solutions. This sequential application of autonomous algorithms is called high-level relay hybridization according to the taxonomy introduced by Talbi [11]. The total number of function evaluations is partitioned among the algorithms.

Experiment on the generalized Schaffer problem: We conducted two experiments to analyze the limit behavior of the hybrid algorithm on the generalized Schaffer problem (Section 3) which reads $f_1(\mathbf{x}) = 1/d^\alpha(\sum_{i=1}^d x_i^2)^\alpha$, $f_2(\mathbf{x}) = 1/d^\alpha(\sum_{i=1}^d(1-x_i)^2)^\alpha, x \in \mathcal{X} = [0,1]^d, \alpha \in \mathbb{R}^+$, and both objectives to be minimized. The first 1000 evaluations are always performed by SMS-EMOA. Figures 5 and 6 show a clipping of the subsequent behavior of typical runs, at which SMS-EMOA is always started using the same random seed.

In Fig. 5 the results pertaining to the generalized Schaffer problem with $d = 10$, $\alpha = \frac{1}{2}$ (hence the Pareto front is linear, cf. Section 3) of the following experiment are shown. The population size μ is 5, 10, or 15, and dimension d is 10, 15, or 20. The purpose of this experiment was to study the convergence behavior of the gradient part of the algorithm. We see that the convergence (after a reasonable starting population has been found by the SMS part) is linear or almost linear. The former is especially true for small sizes of the approximation sets. The dimension of the search space has less effect on the speed of the methods. This can be explained by the relatively long time needed to perform line searches, as the dimension of the search space only influences the time needed for the gradient computation.

Fig. 6 shows the results for the generalized Schaffer problem with $\alpha = 1$, the dimension of the search space $d = 10$, and a population size (i.e., the size of the approximation set) of 10. The Pareto front is equal to $\{(y_1, y_2) \mid y_2 = 1 - 2\sqrt{y_1} + y_1 \text{ and } 0 \leq y_1 \leq 1\}$ and the maximally attainable S-metric is $1 - \frac{1}{6} \approx 0.833333$. The discontinuities in the progress correspond to the end of a line search, and a gap indicates that function evaluations are spend on the gradient calculation. The picture shows that once the gradient part of the hybrid method is supplied

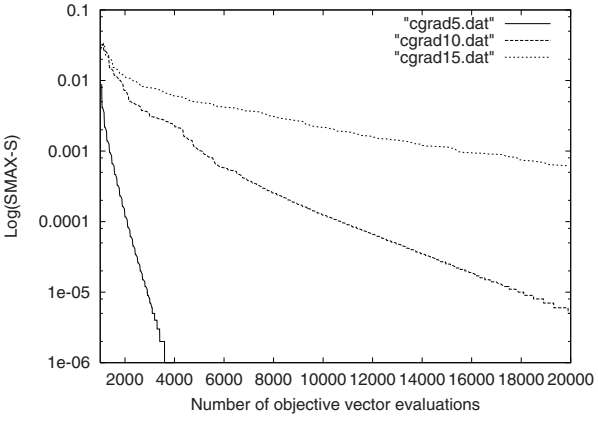

Fig. 5. The limit behavior of the gradient method starting from a population evolved over 1000 iterations with the SMS-EMOA for different problem dimensions d and population sizes μ. The logarithmic distance to the known optimum of the S-metric is plotted for different strategies.

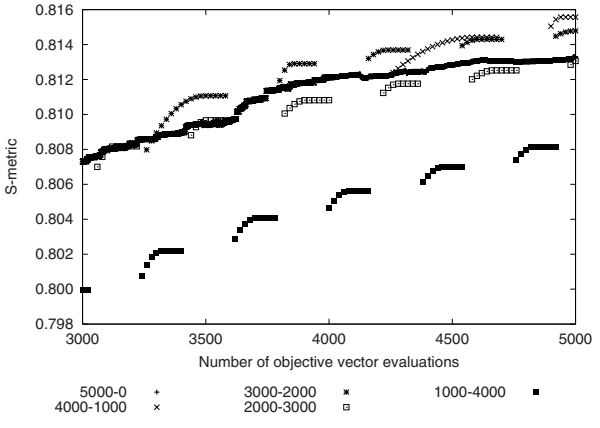

Fig. 6. The limit behavior of the gradient method starting from a population evolved over 1000 iterations by SMS-EMOA. The S-metric is plotted for different strategies, where the first number denotes the number of evaluations of the SMS part and the second of the gradient part.

with a reasonably good approximation set to the Pareto front the gradient part of the method outperforms the pure SMS-EMOA.

Studies on the ZDT Test Suite: Fig. 7 refers to the experiments run on the problem ZDT6 of the ZDT benchmark [5]. The size of the approximation set was chosen to be 20. Runs without penalty (Fig. 7, top) and with penalty (Fig. 7, bottom) on dominated points have been conducted. The total number of function evaluations

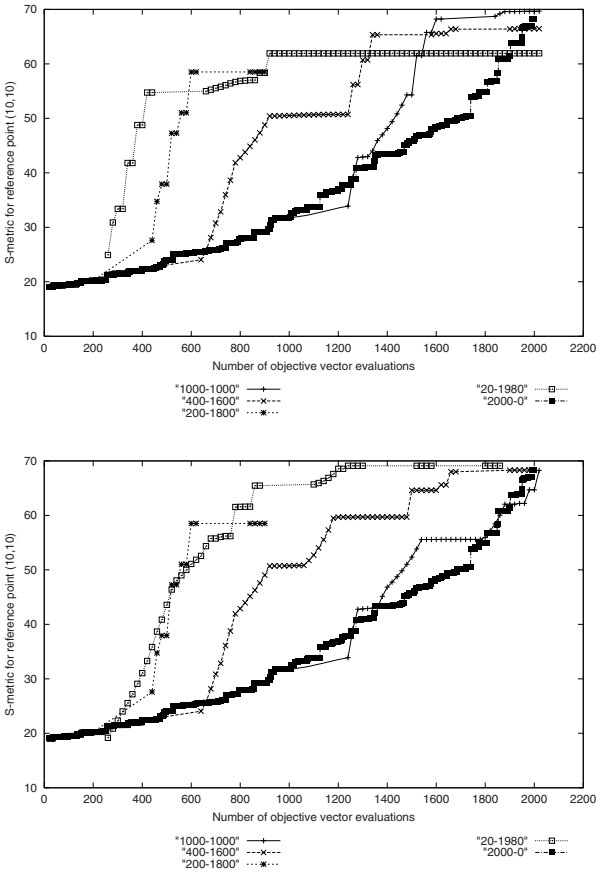

Fig. 7. Convergence of the hybrid algorithm for different switching points at which the gradient based solver takes over on the 10-D ZDT6 problem. In the upper (lower) figure the strategy is without (with) penalty. The numbers in the legend determine the number of function evaluation before and after switching. All strategies computed a total of 2000 evaluations and used a population size of 20.

in each run was 2000. Five different strategies were performed, listed with increasing number of function evaluations dedicated to the SMS part: 20, 200, 400, 1000, and 2000, respectively. The remainder of the 2000 function evaluations was used for the gradient part.

The two pictures reveal that it pays off to apply the gradient part of the algorithm as soon as a rough approximation set has been found. The speed-up occurs especially at the beginning and thus the hybrid approach is useful in case you would like to get very good results with few function evaluations. Secondly the picture also shows that giving a penalty to points in the population which are dominated gives far better approximation sets w.r.t. the S-metric.

Table 1. Runs with the relay hybrid obtained on the ZDT test suite. For each variant five runs have been performed. For ZDT4 the pure gradient approach failed to find a point dominating the reference point, thus the S-metric value remained zero.

Problem	ST	MIN 1000	AVG 1000	MAX 1000	MIN 1500	AVG 1500	MAX 1500	MIN 2000	AVG 2000	MAX 2000
ZDT1	1	21.876223	23.275247	24.199412	21.876223	23.850845	24.487564	21.876223	23.903422	24.505799
ZDT1	2	21.118595	23.384878	24.342832	23.403862	23.974269	24.449935	23.604535	24.121280	24.457263
ZDT1	3	17.202933	20.021265	21.845488	23.774588	24.101479	24.361921	24.175715	24.375030	24.482113
ZDT1	4	17.202933	20.021265	21.845488	21.583049	22.643311	23.524577	24.113695	24.256499	24.403682
ZDT1	5	17.202933	20.021265	21.845488	21.583049	22.643311	23.524577	23.060069	23.726008	24.365373
ZDT2	1	19.162475	21.245294	24.064990	19.412864	22.237161	24.106690	19.412880	22.808071	24.127194
ZDT2	2	18.126020	20.581421	23.101347	19.060637	21.727331	23.428600	19.695295	22.106164	23.962237
ZDT2	3	14.661332	18.103744	19.873194	14.661332	20.404380	23.680670	14.661332	20.654252	23.686246
ZDT2	4	14.661332	18.103744	19.873194	18.657467	19.972369	21.240527	19.999733	22.152473	23.640263
ZDT2	5	14.661332	18.103744	19.873194	18.657467	19.972369	21.240527	19.995868	20.817935	22.205402
ZDT3	1	22.399148	25.882488	27.149451	22.399148	26.109937	27.331745	22.399148	26.242184	27.373348
ZDT3	2	24.461667	26.156686	27.223846	24.517026	26.290136	27.464974	24.535586	26.645181	27.488230
ZDT3	3	19.142756	21.624740	23.348263	23.360338	25.773577	27.284952	23.797245	25.967965	27.398976
ZDT3	4	19.142756	21.624740	23.348263	22.030572	24.145441	25.792775	23.555302	26.097489	27.326041
ZDT3	5	19.142756	21.624740	23.348263	22.030572	24.145441	25.792775	24.719521	25.779436	27.260427
ZDT4	1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ZDT4	2	1.991208	8.738555	12.296512	1.992431	9.560956	15.533157	1.993505	10.070364	17.979625
ZDT4	3	6.070645	10.526529	14.042237	9.518876	12.649215	16.239738	10.526157	13.008094	16.239738
ZDT4	4	6.070645	10.526529	14.042237	8.052115	11.965294	15.086376	8.052115	12.311928	16.310129
ZDT4	5	6.070645	10.526529	14.042237	8.052115	11.965294	15.086376	8.587789	12.613113	16.121092
ZDT6	1	57.826184	70.856967	78.100751	60.275019	72.771141	79.744496	60.364094	73.882299	83.297907
ZDT6	2	36.830943	61.742804	72.663663	38.297947	68.691263	78.227012	51.547110	72.630236	79.137021
ZDT6	3	38.012356	51.377244	63.308468	62.527358	68.822536	77.404330	71.326729	78.212762	85.345639
ZDT6	4	38.012356	51.377244	63.308468	53.872935	75.230301	83.236934	75.959491	80.438820	85.029665
ZDT6	5	38.012356	51.377244	63.308468	53.872935	75.230301	83.236934	81.736268	88.169766	91.648977

The finding of a reasonable approximation set to be used as a starting point for the gradient method is always done by the SMS. In the nearly pure gradient method also a very tiny fraction of the total number of functions evaluations is used by SMS-EMOA (20 evaluations). Clearly, the hybrid algorithm converges in each case to a population with maximum S-metric. Also the pure SMS method eventually catches up with the hybrid algorithm and converges to the maximum.

Table 1 shows the results of running the hybrid algorithm on the ZDT test suite (ZDT1 - ZDT4, and ZDT6). On each of the five problems the five different distributions of 2000 function evaluations among the hybrid parts are applied: (1) SMS: 20, gradient: 1980, (2) SMS: 500, gradient: 1500, (3) SMS: 1000 gradient: 1000, (4) SMS: 1500, gradient: 500, (5) SMS: 2000, gradient: 0. Each version of the hybrid algorithm is repeated five times with different random seeds. The reference point for each of the first four ZDTs was chosen as $(5, 5)$ and for ZDT6 it was $(10, 10)$. There are three checkpoints (at 1000, 1500, and 2000 evaluations) at which the minimal, average, and maximal S-metric are recorded (calculated concerning the five repetitions of a strategy). All strategies used the penalty function for dominated points. For ZDT1 and ZDT2 it is clear that the hybrid method is outperforming the pure SMS algorithm. In case of ZDT3 the pure gradient method is somewhat worse than the pure SMS on the other hand in case the first half of the function evaluations is spent on SMS (line 3 of ZDT3) the hybrid method outperforms the pure SMS again. A similar remark can be made about ZDT4 except that the pure gradient method in this case does not give good results due to reference point sensitivity. The reference point has been chosen too close to the Pareto front so that no point dominates it after a small number of function evaluations and the gradient strategy cannot work. The reference point sensitivity is not present in the SMS part of the algorithm as it only looks for relative increments of the hypervolume and (if $d = 2$) always selects extremal points directly. We see that when 500 or more evaluations are first spent on the SMS the hybrid is again competitive with the pure SMS. In case of ZDT6 which is multimodal the hybrid strategies do worse than the pure SMS. In all cases we see that the gradient method gives a speed-up especially in the beginning of the optimization.

6 Conclusions and Outlook

This paper introduces the gradient computation of the S-metric with respect to a population of points. Using the chain rule, the gradient of the S-metric can be computed from the gradients of the objective functions. It is important to distinguish between strictly dominated, weakly dominated, and non-dominated points. While for non-dominated sets differentiability is inherited from the objective functions, in the presence of weakly dominated points one-sided derivatives occur. For strictly dominated points sub-gradients with value zero occur. They make it impossible to improve these points by means of gradient methods. This problem can be partly circumvented by introducing a penalty approach.

However, the experiments in this paper show that it is advantageous to start the search with non-dominated sets close to the Pareto front, computed by

an evolutionary algorithm, preferably one which maximizes the S-metric, too. Therefore, the proposed relay hybrid between the SMS-EMOA and a gradient method seems promising, though refined rules for phase switching still needs to be worked out. The study on the generalized Schaffer problem shows the potential of the new approach to find high precision approximations of finite populations maximizing the S-metric.

Future research should extend the empirical work on benchmarks and study problems of higher objective space dimension. Though some basic ideas of the gradient computation for more than two objectives using the chain rule have been sketched, details of the implementation need to be worked out.

Acknowledgments

M. Emmerich acknowledges financial support by the Netherlands Organisation for Scientific Resarch (NWO). Nicola Beume is partly supported by the *Deutsche Forschungsgemeinschaft (DFG)* as part of the *Collaborative Research Center 'Computational Intelligence' (SFB 531).*

References

1. Zitzler, E., Thiele, L.: Multiobjective Optimization Using Evolutionary Algorithms—A Comparative Case Study. In: Eiben, A.E., Bäck, T., Schoenauer, M., Schwefel, H.-P. (eds.) PPSN V. LNCS, vol. 1498, pp. 292–301. Springer, Heidelberg (1998)
2. Zitzler, E., Brockhoff, D., Thiele, L.: The hypervolume indicator revisited: On the design of pareto-compliant indicators via weighted integration. In: Obayashi, S., Deb, K., Poloni, C., Hiroyasu, T., Murata, T. (eds.) EMO 2007. LNCS, vol. 4403, pp. 862–876. Springer, Heidelberg (2007)
3. Fleischer, M.: The measure of pareto optima. Applications to multi-objective metaheuristics. In: Fonseca, C.M., Fleming, P.J., Zitzler, E., Deb, K., Thiele, L. (eds.) EMO 2003. LNCS, vol. 2632, pp. 519–533. Springer, Heidelberg (2003)
4. Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C.M., Grunert da Fonseca, V.: Performance assessment of multiobjective optimizers: An analysis and review. IEEE TEC 7(2), 117–132 (2003)
5. Deb, K.: Multi-Objective Optimization using Evolutionary Algorithms. Wiley, Chichester, UK (2001)
6. Coello Coello, C.A., Van Veldhuizen, D.A., Lamont, G.B.: Evolutionary Algorithms for Solving Multi-Objective Problems. Kluwer Academic Publishers, New York (2002)
7. Zitzler, E., Künzli, S.: Indicator-based selection in multiobjective search. In: Yao, X., Burke, E.K., Lozano, J.A., Smith, J., Merelo-Guervós, J.J., Bullinaria, J.A., Rowe, J.E., Tiňo, P., Kabán, A., Schwefel, H.-P. (eds.) PPSN VIII. LNCS, vol. 3242, pp. 832–842. Springer, Heidelberg (2004)
8. Huband, S., Hingston, P., While, L., Barone, L.: An evolution strategy with probabilistic mutation for multi-objective optimisation. In: CEC03, vol. 4, pp. 2284–2291. IEEE Computer Society Press, Los Alamitos (2003)

9. Knowles, J.: Local-Search and Hybrid Evolutionary Algorithms for Pareto Optimization. Phd thesis, Department of Computer Science, University of Reading, UK (2002)
10. Emmerich, M., Beume, N., Naujoks, B.: An EMO Algorithm Using the Hypervolume Measure as Selection Criterion. In: Coello Coello, C.A., Hernández Aguirre, A., Zitzler, E. (eds.) EMO 2005. LNCS, vol. 3410, pp. 62–76. Springer, Heidelberg (2005)
11. Talbi, E.G.: A Taxonomy of Hybrid Metaheuristics. Journal of Heuristics 8(5), 541–564 (2002)
12. Timmel, G.: Ein stochastisches Suchverfahren zur Bestimmung der Optimalen Kompromißlösungen bei statistischen polykriteriellen Optimierungsaufgaben. Journal TH Ilmenau 6, 139–148 (1980)
13. Fliege, J., Svaiter, B.F.: Steepest descent methods for multicriteria optimization. Mathematical Methods of Operations Research 51(3), 479–494 (2000)
14. Shukla, P., Deb, K., Tiwari, S.: Comparing Classical Generating Methods with an Evolutionary Multi-objective Optimization Method. In: Coello Coello, C.A., Hernández Aguirre, A., Zitzler, E. (eds.) EMO 2005. LNCS, vol. 3410, pp. 311–325. Springer, Heidelberg (2005)
15. Schütze, O., Dell'Aere, A., Dellnitz, M.: On continuation methods for the numerical treatment of multi-objective optimization problems. In: Branke, J., Deb, K., Miettinen, K., Steuer, R. (eds.) Practical Approaches to Multi-Objective Optimization. Dagstuhl Seminar Proceedings, IBFI, Schloss Dagstuhl, Germany, vol. 04461 (2005)
16. Bosman, P.A., de Jong, E.D.: Combining gradient techniques for numerical multi-objective evolutionary optimization. In: Keijzer, M., et al. (eds.) GECCO06, vol. 1, pp. 627–634. ACM Press, Seattle, USA (2006)
17. Emmerich, M., Deutz, A.: Test Problems based on Lamé Superspheres. In: Obayashi, S., Deb, K., Poloni, C., Hiroyasu, T., Murata, T. (eds.) EMO 2007. LNCS, vol. 4403, pp. 922–936. Springer, Heidelberg (2007)
18. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge, UK (2006)
19. Wagner, T., Beume, N., Naujoks, B.: Pareto-, Aggregation-, and Indicator-based Methods in Many-objective Optimization. In: Obayashi, S., Deb, K., Poloni, C., Hiroyasu, T., Murata, T. (eds.) EMO 2007. LNCS, vol. 4403, pp. 742–756. Springer, Heidelberg (2007)
20. Beume, N., Naujoks, B., Emmerich, M.: SMS-EMOA: Multiobjective Selection Based on Dominated Hypervolume. European Journal of Operational Research 181(3), 1653–1669 (2007)