Time Complexity and Zeros of the Hypervolume Indicator Gradient Field

Michael Emmerich and André Deutz

Leiden University, Leiden Institute for Advanced Computer Science
2333 CA Leiden, The Netherlands
{emmerich,deutz}@liacs.nl
http://natcomp.liacs.nl

Abstract. In multi-objective optimization the hypervolume indicator is a measure for the size of the space within a reference set that is dominated by a set of μ points. It is a common performance indicator for judging the quality of Pareto front approximations. As it does not require a-priori knowledge of the Pareto front it can also be used in a straightforward manner for guiding the search for finite approximations to the Pareto front in multi-objective optimization algorithm design.

In this paper we discuss properties of the gradient of the hypervolume indicator at vectors that represent approximation sets to the Pareto front. An expression for relating this gradient to the objective function values at the solutions in the approximation set and their partial derivatives is described for arbitrary dimensions $m \geq 2$ as well as an algorithm to compute the gradient field efficiently based on this information. We show that in the bi-objective and tri-objective case these algorithms are asymptotically optimal with time complexity in $\Theta(\mu d + \mu \log \mu)$ for d being the dimension of the search space and μ being the number of points in the approximation set. For the case of four objective functions the time complexity is shown to be in $\mathcal{O}(\mu d + \mu^2)$. The tight computation schemes reveal fundamental structural properties of this gradient field that can be used to identify zeros of the gradient field. This paves the way for the formulation of stopping conditions and candidates for optimal approximation sets in multi-objective optimization.

Keywords: Set Oriented Optimization, Multiobjective Gradient, Hypervolume Indicator, Computational Complexity, Optimality Conditions.

1 Introduction

The gradient field assigns to each vector in the search space (or decision space) a vector of all partial derivatives at this vector that is called the gradient at this point. Gradients play an important role in the formulation of optimization algorithms, as they are vectors that point in the direction where function values will increase the most and thus can guide the search towards better solutions. Moreover, for differentiable functions the gradient at local optima is zero, which can be used to identify candidates for local optima.

O. Schütze et al. (eds.), *EVOLVE - A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation III,* Studies in Computational Intelligence 500,
DOI: 10.1007/978-3-319-01460-9_8,

The problem of solving multi-objective optimization problems, is often restated as finding a finite approximation set to the Pareto front of the problem. In this case the hypervolume indicator provides a figure of merit for an approximation set. Loosely speaking, it measures the volume of the subspace that is Pareto dominated by the approximation set. The hypervolume indicator gradient at a set of decision vectors points in the direction that locally yields maximal improvement of this indicator by simultaneously updating all points. It was first described in [1], but analysis and computation schemes were mainly restricted to the bi-objective case. This chapter presents a substantially extended analysis and efficient algorithms for computing the hypervolume indicator gradient field. In the bi- and tri-objective cases these algorithms are even asymptotically optimal. In particular the following research questions will be addressed:

Given information on the objective function vectors and partial derivatives of the objective functions for all points in the approximation set:

- Can we concisely define the hypervolume indicator gradient field and the points where it is defined for an arbitrary number of objective functions?
- Can structural properties of the gradient expression be exploited to find efficient algorithms for computing the hypervolume gradient?
- Can these structural properties be used to identify compact equations for the zeros of the hypervolume gradient field?

As will be shown, the answer to all three questions is affirmative.

In the following discussion we will first establish a formal framework for defining the hypervolume indicator gradient at an approximation set. Actually, we will be talking about two gradient fields:

1. The gradient field for the mapping from a set of decision vectors to the hypervolume indicator
2. The gradient field for the mapping from a set of objective vectors to the hypervolume indicator

We will proceed with the definition of these gradient fields and identify at which domains consisting of approximation sets the gradient fields are well-defined. Efficient algorithms for the computation of the gradient field at an approximation set will be provided, for both mappings. Their asymptotic optimality for the bi- and tri-objective case will be proven. Finally, a locality property of the hypervolume indicator will be discussed. It yields concise formulations of conditions of points where the gradient field of the first mapping obtains values of zero. This can be used in optimality conditions. The same property gives rise to a new interpretation of the hypervolume indicator gradient field and a technique for its visualization. The final section is also devoted to the discussion of implications of the new theoretical results for set-oriented multi-objective optimization in the future.

2 Related Work

The idea to use gradient information in multi-objective optimization is not new.

Fliege [2] suggests a steepest descent method that searches within the cone of dominating solutions in the direction where the net decrease of objective function

values is expected to be maximal among all vectors with a given length, added to the current variable vector. This direction, obtained by quadratic maximization based on the Jacobian (the matrix of the objective functions gradients), is termed *multi-criterion gradient*. Variations and generalizations of this approach have been proposed by Brown and Smith [3] and Bosman and de Jong [4]. More recently a gradient based method that approximates the gradient from points that are generated in an evolutionary search in [5] was suggested. A similar line of research is given by methods that generate non-dominated points by linear combinations of the negative gradients with positive weights [6,7]. For small step-sizes this yields non-dominated or dominating solutions. Thereby, the Euler method is used to integrate along a path of such solutions. Recently, these methods have been hybridized for evolutionary multi-criterion optimization by Shukla et al. [8] by computing favorable directions for generating offspring individuals. Unlike the aforementioned methods, homotopy and continuation methods as described by Hillermeier [9] and Schütze et al. [10] use gradient-based search not in the first place to move search points closer to the Pareto front, but for finding a well-distributed set of points covering the Pareto fronts. The basic idea is to gradually extend the manifold around a given Karush-Kuhn-Tucker point. This way, given a smooth and connected Pareto front, accurate approximations can be achieved. A technique called *directed search* uses gradient information to steer the search in a desired direction given by a vector in the objective space [11].

In this chapter we will further explore an alternative use of gradients in multi-objective optimization that was proposed in [1]. Here the gradient field is formulated on the (μd)-dimensional space of concatenated sets of μ decision vectors in $\mathbb{R}^d$ or, respectively, at multi-sets of decision vectors $\mathbb{R}^{\mu d}$. Following this paradigm, *the improvement of a single decision vector is measured explicitly and solely in how much this vector improves with respect to its contribution to a scalar performance measure stated on an entire set of decision vectors.*

3 Formal Definition of the Hypervolume Indicator Gradient Field

A central concept in this work is that of a gradient at a vector and that of a gradient field. To avoid ambiguity of language we will provide elementary definitions, first.

3.1 Gradient at a Vector and Gradient Field

We introduce partial derivatives via one-sided partial derivatives for a function $\varphi : \mathbb{R}^n \to \mathbb{R}$.

$$\frac{\partial_+ \varphi}{\partial x_i}(\mathbf{x}) = \lim_{t \downarrow 0} \frac{\varphi(\mathbf{x} + t\mathbf{e}_i) - \varphi(\mathbf{x})}{t}$$

denotes the right one-sided partial derivative at $\mathbf{x}$ for x_i, and

$$\frac{\partial_- \varphi}{\partial x_i}(\mathbf{x}) = \lim_{t \uparrow 0} \frac{\varphi(\mathbf{x} + t\mathbf{e}_i) - \varphi(\mathbf{x})}{t}$$

is the left one-sided partial derivative. If both values exists at a point $\mathbf{x}$ and are equal, we denote by

$$\frac{\partial \varphi}{\partial x_i}(\mathbf{x}) := \frac{\partial_+ \varphi}{\partial x_i}(\mathbf{x}) = \frac{\partial_- \varphi}{\partial x_i}(\mathbf{x})$$

the partial derivative at $\mathbf{x}$ with respect to x_i.

The gradient of a function $\mathbb{R}^n \to \mathbb{R}$ at a vector is a vector pointing in the direction of the steepest ascent at that point. The steepness of the slope is given by the length of this vector. It is defined via partial derivatives as:

$$\nabla \varphi(\mathbf{x}) := \left(\frac{\partial \varphi}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial \varphi}{\partial x_n}(\mathbf{x}) \right)^\top . \tag{1}$$

The function $\nabla \varphi : \mathbb{R}^n \to \mathbb{R}^n$ is commonly referred to as the *gradient field* associated to φ.

3.2 Multi-objective Optimization, Efficient Set, and Pareto Front

In multi-objective (or: multicriteria) optimization, we consider an m-tuple of functions

$$(f_1 : \mathbb{R}^d \to \mathbb{R}, \ldots\ldots, f_k : \mathbb{R}^d \to \mathbb{R}, \ldots\ldots, f_m : \mathbb{R}^d \to \mathbb{R}),$$

each function of which is to be minimized or maximized. Without loss of generality, in the following we assume the goal is maximization. We denote by $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ the corresponding vector valued function $(f_1, \ldots, f_m)^\top$. The practically very important special cases $m = 2$ and $m = 3$ are called bi-objective (or: bicriteria) and tri-objective (or: tricriteria) problems.

In the following discussion it will be important to clearly distinguish between decision vectors $\mathbf{x} \in \mathbb{R}^d$, that is the domain of $\mathbf{f}$ or *decision space*, and objective vectors in $\mathbf{y} \in \mathbb{R}^m$, that is the co-domain of $\mathbf{f}$ or *objective space*. As $\mathbf{f}$ is not necessarily surjective the following definition is made: An objective vector $\mathbf{y}$ is *attainable* if $\mathbf{y} = \mathbf{f}(\mathbf{x})$ for some $\mathbf{x} \in \mathbb{R}^d$. The set of attainable objective vectors is termed *attainable objective space*.

The above problem of multi-objective optimization is not well stated, as it is not clear how to deal with *conflicting objective functions*, that is pairs f_k and $f_{k'}$ with $\arg\min_{\mathbf{x} \in \mathbb{R}^d}(f_k) \cap \arg\min_{\mathbf{x} \in \mathbb{R}^d}(f_{k'}) = \emptyset$. However, Pareto dominance establishes a partial order on the objective space. The maximal elements of this partial order for the attainable objective space we will term *Pareto optimal objective vectors* and their pre-images with respect to $\mathbf{f}$ we will term *efficient decision vectors*. Accordingly, the set of all Pareto optimal objective vectors we term *Pareto front*, whereas the *efficient set* will be the set of all efficient decision vectors. See also Ehrgott [12] for these definitions.

In Pareto optimization we are interested in finding the efficient set and Pareto front for $\mathbf{f}$. The Pareto front is interesting, because it reveals the nature of the trade-off between different objectives and contains all objective vectors that cannot be strictly improved anymore without additional statements about preferences.

Remark 1. Note that we restrict ourselves here to the continuous and unconstrained case, but definitions can be generalized in a straightforward way to decision spaces with (integrity) constraints. This does however not hold for the gradient computations that will be discussed in this paper.

In continuous multi-objective optimization we face the problem that the efficient set and the Pareto front of a function can be innumerably large sets. One approach is to approximate the Pareto front with a finite multi-set of, say μ, attainable objective vectors[1]. We will term a multi-set Y of μ solutions in the attainable objective space an approximation set to the Pareto front , and a multi-set X of μ solutions in the decision space an approximation set to the efficient set.

3.3 Hypervolume Indicator

One approach to state optimality of an approximation set in the decision space is to require for an approximation set of maximal hypervolume indicator H. Roughly speaking, this indicator assigns a better (higher) value to approximation sets to the Pareto front that dominate many objective function vectors than to approximation sets that dominate fewer objective vectors. We define

$$\mathrm{DomSet}(\mathrm{Y}) = \{\mathbf{y}' \in \mathbb{R}^m \mid \exists \mathbf{y} \in \mathrm{Y} : \mathbf{y} \text{ Pareto dominates } \mathbf{y}'\}$$

As this set has infinite measure, its size cannot serve as an indicator. Instead the hypervolume indicator measures the size of the dominated volume within the reference set $[\mathbf{r}, \infty)$ for a reference vector $\mathbf{r} \in \mathbb{R}^m$. Hence, the definition of the *hypervolume indicator* reads:

$$H(\mathrm{Y}, \mathbf{r}) = \lambda(\mathrm{DomSet}(\mathrm{Y}) \cap [\mathbf{r}, \infty)),$$

and λ denotes the Lebesgue measure on $\mathbb{R}^m$, that is the area of the dominated set in the reference space is measured in case $m = 2$ and its volume in the case $m = 3$. The choice of a proper reference point is a task that is typically delegated to the user. Ideally it should be dominated by all attainable objective vectors. We write $H(\mathrm{Y})$ instead of $H(\mathrm{Y}, \mathbf{r})$ if the definition of $\mathbf{r}$ is clear from the context.

For geometrical considerations the following equivalent definition (for $m > 1$) is sometimes more accessible, but requires $\mathbf{r} \leq \mathbf{y}$, componentwise, for all $\mathbf{y} \in \mathbf{Y}$:

$$H(\mathrm{Y}, \mathbf{r}) = \lambda(\cup_{\mathbf{y} \in \mathrm{Y}} [\mathbf{r}, \mathbf{y}]).$$

The hypervolume indicator (or: S-metric) was first introduced as a *unary performance indicator* [13,14] and is nowadays also widely used in bounded-size archiving and to guide the search towards the Pareto front. It is commonly used and analyzed in the context of evolutionary multi-objective optimization [15][16], but has hardly been considered so far in deterministic algorithms for finding the

[1] Note that the symbol μ is used, as it is a common symbol for denoting the size of a population in evolutionary multi-objective optimization.

Pareto front (cf. [17]). Recently, the hypervolume indicator received attention in computational geometry as it is a special case of Klee's measure problem and it may serve to establish lower complexity bounds for this problem [18]. In general it is likely that the time complexity of the hypervolume indicator is exponential in dimension m, while fast algorithms with subquadratic time complexity in the number of points in the approximation set μ exists for the practically relevant cases with $m = 2$, $m = 3$ (cf. [19]), and $m = 4$ (cf. [20]).

3.4 Gradients at Approximation Sets

As gradients are defined at vectors and not at multi-sets, a mapping from multi-sets to vectors that represent these will be established next. A multi-set X of μ decision vectors in $\mathbb{R}^d$, that may serve as an approximation to the efficient set, is represented as a concatenation of its elements and called a μd-vector. We say a μd-vector

$$\mathbf{X} := (x_1^{(1)}, \ldots, x_d^{(1)}, \ldots\ldots, x_1^{(i)}, \ldots, x_d^{(i)}, \ldots\ldots, x_1^{(\mu)}, \ldots, x_d^{(\mu)})^\top \in \mathbb{R}^{\mu \cdot d}$$

represents the multi-set $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\mu)}\}$. We name the subsequence with upper index $i \in \{1, \ldots, \mu\}$, the i-th *subvector* of the μd vector. Accordingly, we can represent multi-sets in $\mathbb{R}^m$, that may serve as approximations to the Pareto front, as μm-vectors $\mathbf{Y}$. We say

$$\mathbf{Y} := (y_1^{(1)}, \ldots, y_m^{(1)}, \ldots\ldots, y_1^{(i)}, \ldots, y_m^{(i)}, \ldots\ldots, y_1^{(\mu)}, \ldots, y_m^{(\mu)})^\top \in \mathbb{R}^{\mu \cdot m}$$

represents the multi-set $\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(\mu)}\}$. We name the subsequence with upper index $i \in \{1, \ldots, \mu\}$, the i-th *subvector* of the μm vector. As μ (the number of points in the approximation set), d (the number of dimensions of the decision space) and m (the number of objective functions) are constant in the search algorithms that we consider, it turns out to be a convenient notational convention.

The mapping we have just defined is, in general, not injective.

Proposition 1. *Every multi-set of size μ with elements in $\mathbb{R}^d$ (or, respectively, $\mathbb{R}^m$) has at least one and at most $\mu!$ representing μd-vectors (or, respectively μm-vectors). Each μd-vector (or, respectively μm-vector) represents exactly one multi-set in $\mathbb{R}^d$ (or, respectively, $\mathbb{R}^m$).*

Proof. The concatenation of vectors from the multi-set can be done in $\mu!$ different orders. Due to duplicates the number of distinguishable μd vectors might be less than $\mu!$. □

The proposition takes into account the possibility of duplicates in the multi-set, in which case the number of representations will be less than $\mu!$.

To establish a connection between μd-vectors and μm-vectors, define the mapping $\mathbf{F} : \mathbb{R}^{\mu d} \to \mathbb{R}^{\mu m}$ with

$$\mathbf{X} \mapsto (f_1(\mathbf{x}^{(1)}), \ldots, f_m(\mathbf{x}^{(1)}), \ldots\ldots, f_1(\mathbf{x}^{(\mu)}), \ldots, f_m(\mathbf{x}^{(\mu)}))$$

Remark 2. The reformulation of multi-sets to concatenated vectors will not be needed in the long run, as we will show that the gradient can be decomposed into subgradients associated with single points. To say it with Wittgenstein's metaphor, our construction serves as a 'ladder' that after we climbed it can be discarded again.

For technical reasons, first the definition of the hypervolume indicator needs to be slightly adapted to be compatible with the vector representation:

$$\mathcal{H}(\mathbf{Y}) = \lambda \left(\bigcup_{i=1,\ldots,\mu} (-\infty, (y_1^{(i)}, \ldots, y_m^{(i)})^\top] \cap [\mathbf{r}, \infty) \right). \tag{2}$$

Proposition 2. *Let* $\mathbf{Y}$ *denote a* μm*-vector that represents some multi-set* Y *in* $\mathbb{R}^m$*. Then* $\mathcal{H}(\mathbf{Y}) = H(\mathrm{Y})$*.*
Proof. □

For a given μd-vector $\mathbf{X}$ of μ points we define:

$$\mathcal{H}_\mathbf{F}(\mathbf{X}) := \mathcal{H}(\mathbf{F}(\mathbf{X})). \tag{3}$$

The introduced formal framework is sound, as by optimizing $\mathcal{H}_\mathbf{F}$ over the set of μd-vectors we will obtain multi-sets of maximal hypervolume. For precision, the following lemma is stated:

Lemma 1. *Each multi-set of size* μ *that maximizes* $H_\mathbf{F}$ *is represented by at least one and at most* $\mu!$ *maxima of* $\mathcal{H}_\mathbf{F}$*. A* μd*-vector that is not maximal with respect to* $\mathcal{H}_\mathbf{F}$ *does not represent a maximal multi-set of size* μ *for* $H_\mathbf{F}$*.*

Proof. This follows from Propositions 1 and 2. □

4 The Hypervolume Gradient Field

The gradient field $\nabla\mathcal{H}_\mathbf{F}$ is defined by Equation 1 for the mapping $\mathcal{H}_\mathbf{F}$ at any μd-vector where $\mathcal{H}_\mathbf{F}$ is differentiable, that is for any μd-vector for which all partial derivatives with respect to $\mathcal{H}_\mathbf{F}$ are well defined. Analogously, the gradient field $\nabla\mathcal{H}$ is defined by Equation 1 at any μm-vector where $\mathcal{H}$ is differentiable.

We will first look at how the partial derivatives of the gradient field $\nabla\mathcal{H}$ and $\nabla\mathcal{H}_\mathbf{F}$ can be computed given the information (in the points where the functions are partially differentiable):

$$f_k(\mathbf{x}^{(i)}) \text{ for } i = 1, \ldots, \mu; k = 1, \ldots, m$$

and

$$\frac{\partial f_k}{\partial x_j^{(i)}}(\mathbf{x}^{(i)}) \text{ for } i = 1, \ldots, \mu; j = 1, \ldots, m; k = 1, \ldots, m.$$

Subsequently we will classify regions of differentiability.

4.1 C $\mathcal{H}_{\mathbf{F}}$ at a μd-Vector

Using a different notation the mapping $\mathcal{H}_{\mathbf{F}}$ in Equation 3 can be defined by the following composition of mappings:

$$\mathbb{R}^{\mu \cdot d} \underbrace{\xrightarrow{\mathbf{F}}}_{\text{decision space to objective space}} \mathbb{R}^{\mu \cdot m} \underbrace{\xrightarrow{\mathcal{H}}}_{\text{objective space to single value}} \mathbb{R}. \quad (4)$$

According to Equation 1 the hypervolume indicator gradient $\nabla \mathcal{H}_{\mathbf{F}}(\mathbf{X})$ of the composition $\mathcal{H}_{\mathbf{F}} = \mathcal{H} \circ \mathbf{F}$ is defined as:

$$\nabla \mathcal{H}_{\mathbf{F}}(\mathbf{X}) = \left(\frac{\partial \mathcal{H}_{\mathbf{F}}(\mathbf{X})}{\partial x_1^{(1)}}, \ldots, \frac{\partial \mathcal{H}_{\mathbf{F}}(\mathbf{X})}{\partial x_d^{(1)}}, \ldots, \frac{\partial \mathcal{H}_{\mathbf{F}}(\mathbf{X})}{\partial x_1^{(\mu)}}, \ldots, \frac{\partial \mathcal{H}_{\mathbf{F}}(\mathbf{X})}{\partial x_d^{(\mu)}} \right)^{\top} \quad (5)$$

The chain rule provides us with the gradient of $\mathcal{H}_{\mathbf{F}}$ at a point $\mathbf{X}$:

$$\nabla \mathcal{H}_{\mathbf{F}}(\mathbf{X}) = \left(\left(\nabla \mathcal{H} \begin{pmatrix} \mathbf{f}(\mathbf{x^{(1)}}) \\ \mathbf{f}(\mathbf{x^{(2)}}) \\ \cdots \\ \mathbf{f}(\mathbf{x}^{(\mu)}) \end{pmatrix} \right)^{\top} \cdot \begin{pmatrix} \mathbf{f}' \text{ at } \mathbf{x}^{(1)} & 0 & 0 \cdots & 0 \\ 0 & \mathbf{f}' \text{ at } \mathbf{x}^{(2)} & 0 \cdots & 0 \\ \vdots & \vdots & \vdots \cdots & \vdots \\ 0 & 0 & 0 \ 0 \ \mathbf{f}' \text{ at } \mathbf{x}^{(\mu)} & \end{pmatrix} \right)^{\top} \quad (6)$$

To visualize the structure of the composition we give a detailed description:

$$\underbrace{\left(\left(\begin{matrix} \frac{\partial \mathcal{H}}{\partial y_1^{(1)}} \\ \vdots \\ \frac{\partial \mathcal{H}}{\partial y_m^{(1)}} \\ \vdots \\ \frac{\partial \mathcal{H}}{\partial y_1^{(\mu)}} \\ \vdots \\ \frac{\partial \mathcal{H}}{\partial y_m^{(\mu)}} \end{matrix} \right) (\mathbf{F}(\mathbf{X})) \right)^{\top}}_{\nabla \mathcal{H}(\mathbf{F}(\mathbf{X}))} \cdot \underbrace{\begin{pmatrix} \frac{\partial f_1(\mathbf{x}^{(1)})}{\partial x_1^{(1)}} & \cdots & \frac{\partial f_1(\mathbf{x}^{(1)})}{\partial x_d^{(1)}} & 0 \cdots 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \ \vdots \ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_m(\mathbf{x}^{(1)})}{\partial x_1^{(1)}} & \cdots & \frac{\partial f_m(\mathbf{x}^{(1)})}{\partial x_d^{(1)}} & 0 \cdots 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \vdots \ \cdots \ \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \ \vdots \ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \vdots \ \cdots \ \vdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 \cdots 0 & \frac{\partial f_1(\mathbf{x}^{(\mu)})}{\partial x_1^{(\mu)}} & \cdots & \frac{\partial f_1(\mathbf{x}^{(\mu)})}{\partial x_d^{(\mu)}} \\ \vdots & \vdots & \vdots & \vdots \ \vdots \ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \cdots 0 & \frac{\partial f_m(\mathbf{x}^{(\mu)})}{\partial x_1^{(\mu)}} & \cdots & \frac{\partial f_m(\mathbf{x}^{(\mu)})}{\partial x_d^{(\mu)}} \end{pmatrix}}_{\mathbf{F}'(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\mu)})} \quad (7)$$

It is clear that $\mathbf{F}'(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\mu)})$ depends solely on the gradient functions ∇f_i at the subvectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(\mu)}$ that correspond with the decision vectors of the original problem. Hence, if these $m \cdot \mu$ local gradients are known, the Jacobian matrix $\mathbf{F}'(\mathbf{X})$ can be computed.

4.2 Gradient of the Mapping $\mathcal{H}$ at a μm-Vector

The computation of the components $\nabla \mathcal{H}((y_1^{(1)}, \ldots, y_m^{(1)}, \ldots, y_1^{(\mu)}, \ldots, y_m^{(\mu)}))$ can be traced back to a geometrical problem as depicted in Figure 1. In two dimensions these components are simply the lengths of the line segments of the 'staircase' (or attainment curve). For details, see [1].

Let us next focus on the general case $m \geq 2$: Let $\mathbf{Y}$ denote a μm-vector that is given by the mapping $\mathbf{F}$ at some μd-vector $\mathbf{X}$ in which case $\mathbf{Y} = \mathbf{F}(\mathbf{X})$.

We first look at the case of non-duplicate coordinates in $(y_k^{(1)}, \ldots, y_k^{(\mu)})$ for each $k = 1, \ldots, m$ and points that do not occur at the boundary of the reference space $[\mathbf{r}, \infty)$.

Definition 1. *Let $\pi_{1,\ldots,\check{k},\ldots,m}(\mathbf{y}) \in \mathbb{R}^{m-1}$ denote the projection of a subvector $\mathbf{y}$ in the μm-vector onto the coordinates $1, \ldots, \check{k}, \ldots, m$, where $\check{k}$ means that k is omitted.*

Theorem 1. *Let $i \in \{1, \ldots, \mu\}$. Let H_{m-1} denote the hypervolume indicator for the $(m-1)$-dimensional objective space with reference space $[\pi_{1,\ldots,\check{k},\ldots,m}(\mathbf{r}), \infty)$. Let $Y_{(i)}^{>k}$ denote the multi-set of projections $\pi_{1,\ldots,\check{k},\ldots,m}(\mathbf{y}^{(i)})$ of subvectors $\mathbf{y}^{(i)}$ of a μm-vector $\mathbf{Y}$ with a higher k-th coordinate than the k-th coordinate of the subvector $\mathbf{y}^{(i)}$.*

$$\frac{\partial \mathcal{H}_m}{\partial y_k^{(i)}}(\mathbf{Y}) = H_{m-1}(Y_{(i)}^{>k} \cup \{\pi_{1,\ldots,\check{k},\ldots,m}(\mathbf{y}^{(i)})\}) - H_{m-1}(Y_{(i)}^{>k}).$$

Proof. The theorem follows from the geometrical insight that for a sufficiently small Δ a small variation of the m-th coordinate in positive direction by the amount of Δ will cause a linear increment of the hypervolume indicator by the size of a slice, given by the face of the attainment surface [21] adjacent to this point in the $(m-1)$-dimensional projection times Δ. See also Figure 1 and Figure 2 for a visualization of the geometrical construction in 2-D and, respectively, 3-D.

Example 1. The construction in Figure 2 shows, here for $i = 2$ and $k = 3$, that one-sided partial derivatives are equal to areas of the visible face A which is adjacent to the i-th subvector. In this example $\partial_- \mathcal{H}/\partial y_3^{(2)}$ is smaller than $\partial_+ \mathcal{H}/\partial y_3^{(2)}$ and hence $\partial \mathcal{H}/\partial y_3^{(2)}$ is not defined. Also for $y_3^{(3)}$ one-sided partial derivatives are unequal, while for all other coordinates the one-sided partial derivates are equal in the positive and negative coordinate direction and thus the partial derivatives are defined.

4.3 Characterization of the Set of Differentiable Points

Partial derivatives of a μm vector are either positive, zero, or undefined. In case they are undefined, still all one-sided derivatives exist, but are unequal for at least one coordinate of the μm vector. Next, we provide criteria based on properties of subvectors that allow in most cases to decided whether or not a μm vector is differentiable.

Let us partition the multi-set of subvectors of a given μm-vector $\mathbf{Y}$ with respect to Pareto dominance, relative to the other subvectors, and relative to the reference space:

1. Partitioning into subsets based on Pareto dominance relative to the other subvectors in $\mathbf{Y}$

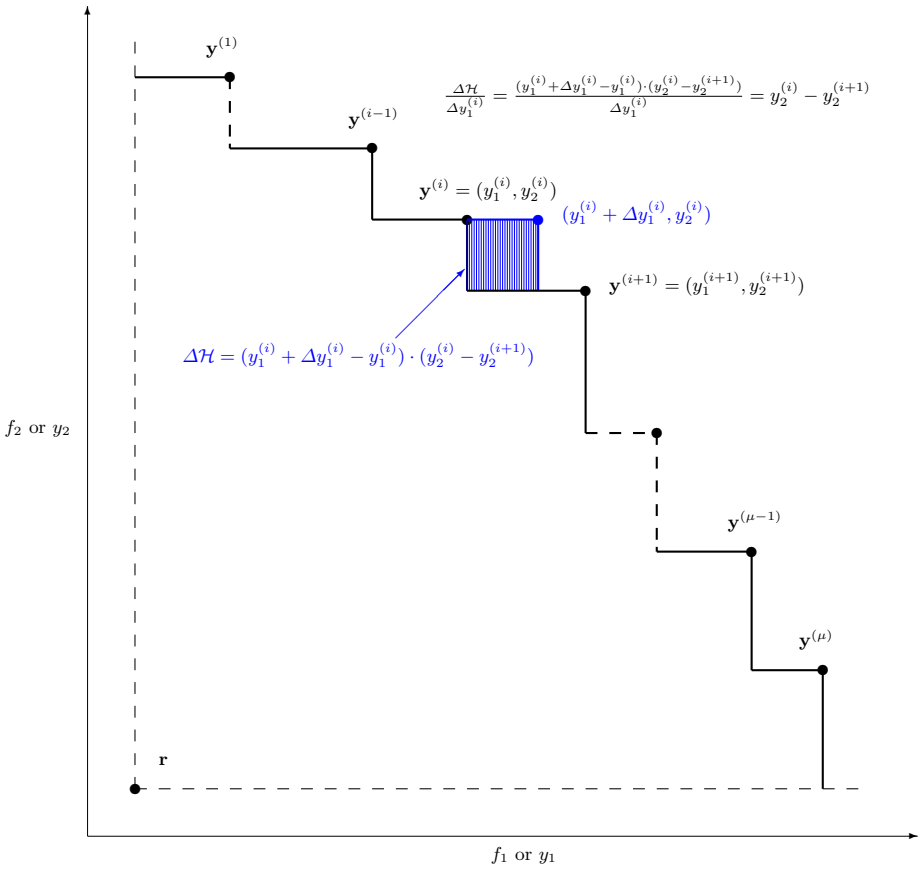

Fig. 1. Geometrical construction used for identifying the partial derivatives $\partial\mathcal{H}/\partial y_j^{(i)}$ of the gradient for $m = 2$ at some non-dominated μm-vector

(a) S: Is the set of strictly dominated subvectors, that is subvectors for which there exists a subvector in $\mathbf{Y}$ that is strictly better in all coordinates.
(b) W: Is the set of weakly dominated subvectors, that is subvectors for which there exists no subvector in $\mathbf{Y}$ that is strictly better in all coordinates and that are Pareto dominated by at least one subvector in $\mathbf{Y}$.
(c) N: Is the set of non-dominated[2] subvectors that in no objective space coordinate have a duplicate value with another non-dominated subvector at this coordinate, e.g. a subvector $(1, 2, 3)^\top$ and another subvector $(3, 2, 1)^\top$ have no duplicate but $(2, 1, 3)^\top$ and $(3, 1, 2)^\top$ have.
(d) D: Is the set of non-dominated subvectors with duplicate coordinates for some objective space coordinate.

[2] where non-domination means here Pareto domination with respect to another subvector in $\mathbf{Y}$

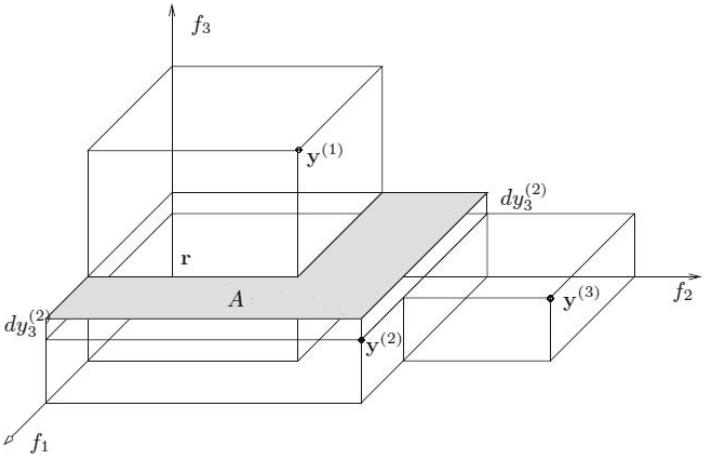

Fig. 2. Geometrical construction for identifying a one-sided partial derivative $\partial_+ \mathcal{H}/\partial y_3^{(2)} := A$ at some non-dominated μd-vector $\mathbf{Y} = (y_1^{(1)}, y_2^{(1)}, y_3^{(1)}, \ldots, y_1^{(3)}, y_2^{(3)}, y_3^{(3)})^T$ for $m = 3$.

2. Partitioning into subsets relative to the reference space $[\mathbf{r}, \infty)$
   (a) I: Is the set of subvectors in the interior of the reference space.
   (b) B: Is the set of subvectors on the boundary of the reference space.
   (c) E: Is the set of subvectors in the exterior of the reference space.

Furthermore, we can partition subvectors with respect to differentiability:

1. Z: Is the set of subvectors for which all partial derivatives are zero.
2. U: Is the set of subvectors for which some partial derivatives are undefined, but as always is the case for the hypervolume indicator $\mathcal{H}$ one-sided partial derivatives exist.
3. P: Is the set of subvectors for which the partial derivatives are all positive.

The relation between these subsets is summarized in the following proposition

Proposition 3

$$Z = E \cup S \tag{8}$$
$$U = D \cup (W \setminus E) \cup (B \setminus S) \tag{9}$$
$$P = N \cap I \tag{10}$$

Proof. In the exterior E and the strictly dominated subspace any differential move of a subvector will leave the hypervolume unchanged, therefore all partial derivatives are zero. In case of $N \cap I$ the size of the face that determines the one-sided partial derivative is the same for the positive and negative direction of a differential move of a single coordinate. It is positive, because the hypervolume will increase (decrease) at the same linear rate when moving the point up or down

in the k-th coordinate. It needs to be shown that the rate is positive. The rate is given by the increment of the $m-1$ dimensional hypervolume to the hypervolume of $Y_{(i)}^{>k}$. This increment must be strictly positive, because the projected point is non-dominated with respect to $Y_{(i)}^{>k}$, and when adding a non-dominated point to a set the hypervolume increases (strict monotonicity property [22]). The projected subvector must be non-dominated in the $m-1$ dimensional projection with respect to the subvectors in $Y_{(i)}^{>k}$ because these vectors are already 'better' in the k-th coordinate and points in N must be non-dominated in m dimensions. For U we cannot decide based on the proposition whether all partial derivatives exist, but the one sided derivatives exist as by changing a single coordinate the hypervolume changes at a linear rate (proportional to the size of a $m-1$ dimensional cuboid) or it remains constant. Clearly Z, U and P do not overlap and cover the set of possible subvectors and thus $\{Z, U, P\}$ forms a partition of the set of subvectors. □

Theorem 2. *A μm-vector with partitionings $\{S, W, N, D\}$ and $\{I, B, E\}$ is differentiable, if $U = D \cup (W \setminus E) \cup (B \setminus S) = \emptyset$.*

Proof. Because Z, U, P is a partition, if the condition is satisfied all subvectors are either in Z or in P and therefore their partial derivatives are defined (either zero or positive). □

Remark 3. In three and more dimensions it is possible that all partial derivatives are defined at subvectors that are non-adjacent but have one coordinate in common. An example would be $\mathbf{Y} = ((1,5,2) \circ (5,1,2) \circ (3,3,3))^\top$. Here we use $\circ$ as a symbol for *concatenation* of tuples, e.g. $((a,b) \circ (c,d)) = (a,b,c,d)$. An example where partial derivatives are undefined due to duplicate coordinates is given with Figure 2, for the 3-rd subvector and the 2-nd subvector.

Example 2. In Figure 3 a μm-vector with $\mu = 10$ and $m = 2$ is depicted. We obtain these subsets:

Partition based on dominance: $S = \{6, 9, 10\}, W = \{5, 7\}, N = \{1, 2, 3, 4, 8\}$, $D = \emptyset$

Partition based on reference space: $I = \{1, 3, 4, 5, 6, 7\}$, $B = \{8, 10\}$, $E = \{2, 9\}$

Partition based on differentiability: $Z = \{2, 6, 9, 10\}$, $U = \{5, 7, 8\}$, $P = \{1, 3, 4\}$.

Clearly, only subvectors in $U \neq \emptyset$ might have unequal one-sided partial derivatives. This is indeed the case for the 5-th and 8-th subvector, while for the 7-th subvector the one-sided partial derivatives are equal and zero.

Duplicates among subvectors in the same coordinate can be checked for easily, and whenever they are obtained and the subvectors are neither in S or in E a deeper investigation might be required for checking differentiability. Next, a neccessary condition for differentiability of such μm vectors will be derived for cases where subvectors are in the interior of the reference space $I = (\mathbf{r}, \infty)$.

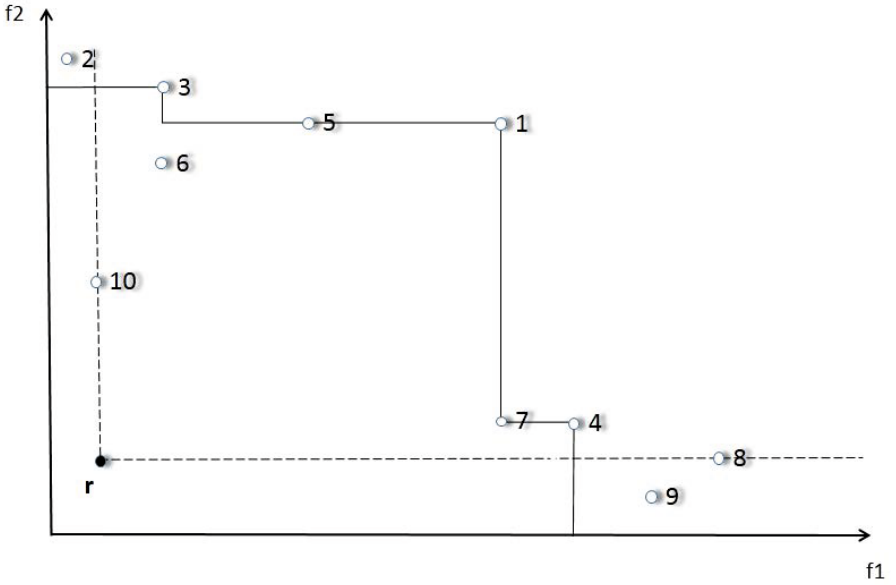

Fig. 3. Differentiability regions

Definition 2. *A subvector* $\mathbf{y}^{(i1)}$ *of* $\mathbf{Y}$ *is said to be* interlaced *with a subvector* $\mathbf{y}^{(i2)}$ *in* $\mathbf{Y}$*, iff* $\exists k \in \{1, ..., m\}$*:* $y_k^{(i1)} = y_k^{(i2)}$ *and*

$$\lambda(([\mathbf{r}_{\check{k}}, \pi_{1,\ldots,\check{k},\ldots m}(\mathbf{y}^{(i1)})] \cap [\mathbf{r}_{\check{k}}, \pi_{1,\ldots,\check{k},\ldots m}(\mathbf{y}^{(i2)})]) \setminus \mathrm{DomSet}(Y_{(i1)}^{>k})) > \delta \quad (11)$$

for some $\delta > 0$*, where* λ *denotes the* $m-1$ *dimensional Lebesgue measure and* $\mathbf{r}_{\check{k}} = \pi_{1,\ldots,\check{k},\ldots,m}(\mathbf{r})$*.*

Proposition 4. *If any two subvectors in* I *are interlaced for some* μm*-vector, the function* $\mathcal{H}$ *is not differentiable.*

Proof. If two vectors are interlaced for the k-th coordinate for which the condition is satisfied it clearly holds that $\partial_- \mathcal{H}/y_k^{(i1)} + \delta \leq \partial_+ \mathcal{H}/y_k^{(i1)}$, and hence the two one-sided partial derivatives are not the same. □

It is conjectured that differentiability is given exactly when all subvectors in the interior of the reference space I are mutually non-interlaced. However, a further investigation of this question and the question of how to check the condition in Equation 11 efficiently is left to the future work.

Finally, the following proposition states a sufficient condition for differentiability of $\mathcal{H}_{\mathbf{F}}$.

Proposition 5. *The set of differentiable points of* $\mathcal{H}_{\mathbf{F}}$ *comprises all* μd*-vectors* $\mathbf{X}$ *for which* $\mathcal{H}$ *is differentiable at* $\mathbf{F}(\mathbf{X})$ *and* $\mathbf{f}$ *is differentiable at all subvectors of* $\mathbf{X}$*.*

Proof. This follows from the well-known fact that the composition of differentiable functions is differentiable and the fact that F is differentiable at $\mathbf{X}$ iff $\mathbf{f}$ is differentiable at each subvector of $\mathbf{X}$. □

Remark 4. Note, that there are points for which $\mathcal{H}_{\mathbf{F}}$ is differentiable that are not captured in the above proposition. In these cases $\mathbf{f}'$ has at some subvectors of $\mathbf{X}$ zero components. Because of these zero components the one-sidedness of components in $\nabla\mathcal{H}(\mathbf{F}(\mathbf{X}))$ might not influence the differentiability at $\mathbf{X}$, if the position of the zeros matches the position of the one-sided derivatives.

5 Efficient Computation

Next, the computational time complexity of computing the gradient field $\nabla\mathcal{H}_{\mathbf{F}}$ at μd-vectors, given the Jacobian matrices $\mathbf{f}'(\mathbf{x}^{(i)})$, $i = 1, \ldots, \mu$ and $\mathbf{F}(\mathbf{X})$ is discussed. Note that the input data requires memory space in $\Theta(\mu dm)$, and the output data requires memory space in $\Theta(\mu d)$. Only worst case complexities are considered here.

A naïve implementation of the scheme proposed above has super-quadratic complexity in the number of points of the approximation set, because a straightforward computation of Equation 6, that is $\nabla\mathcal{H}(\mathbf{F}(\mathbf{X}))^{\top}\mathbf{F}'(\mathbf{x})$ requires no less than $\mu^2 m^2 d$ multiplications and memory resources proportional to $\mu^2 md$.

The computation of the hypervolume indicator can be done efficiently by utilizing

1. the sparsity of the Jacobian matrix $\mathbf{F}'(\mathbf{X})$, and
2. fast dimension sweep algorithms for incremental hypervolume updates when computing $\nabla\mathcal{H}(\mathbf{Y})$ at a given μm-vector $\mathbf{Y}$.

5.1 Exploiting Sparsity in Matrix Multiplication

An observation from studying the structure in Equation 7 is that many components have a zero value and for each column of the matrix only m components need to be considered in the scalar multiplication with the vector on the right hand side.

Theorem 3. *Given a vector valued objective function* $\mathbf{f} : \mathbb{R}^{\mathbf{d}} \to \mathbb{R}^{\mathbf{m}}$*, a* μd*-vector* $\mathbf{X}$*, the partial derivatives* $\frac{\partial\mathcal{H}}{\partial y_k^i}(\mathbf{F}(\mathbf{X}))$ *and* $\frac{\partial f_k(\mathbf{x}^{(i)})}{\partial x_i^{(j)}}$ *for* $i = 1, ..., \mu$*;* $j = 1, \ldots, d$*; and* $k = 1, \ldots, m$ *the* μd *components of* $\frac{\partial\mathcal{H}_{\mathbf{F}}}{\partial x_j^{(i)}}(\mathbf{X})$ *can be computed with a computational complexity in* $\mathcal{O}(\mu dm)$ *by means of*

$$\frac{\partial\mathcal{H}_{\mathbf{F}}}{\partial x_j^{(i)}}(\mathbf{X}) = \sum_{k=1}^{m} \frac{\partial\mathcal{H}}{\partial y_k^{(i)}}(\mathbf{F}(\mathbf{X})) \cdot \frac{\partial f_k(\mathbf{x}^{(i)})}{\partial x_j^{(i)}}, i = 1, \ldots, \mu, j = 1, \ldots, d. \tag{12}$$

Proof. This follows immediately when omitting all zero terms in Equation 6. □

5.2 Dimension Sweep Algorithms for Computing $\nabla\mathcal{H}$

The next goal is to efficiently compute the components of the gradient of the mapping from the objective space to the hypervolume indicator, that is $\frac{\partial\mathcal{H}}{\partial y_k^{(i)}}(\mathbf{Y})$, $i = 1, \ldots, \mu$ and $k = 1, \ldots, m$, for some μm-vector $\mathbf{Y} \in \mathbb{R}^{\mu m}$ after checking for differentiability of $\mathcal{H}$ in $\mathbf{Y} \in \mathbb{R}^{\mu m}$.

Recall, that Theorem 1 states that the partial derivative is given by the incremental change in the dominated hypervolume of the $(m-1)$-dimensional projection, after adding a single point.

Our algorithm is inspired by dimension sweep algorithms for computing the hypervolume indicator as described in [19] and, for 4-D, in [23].

Figure 4 outlines the details of the algorithm to compute hypervolume components. The first part of the algorithms determines all subvectors that evaluate to zero and subvectors with undefined partial derivatives (lines 1-3). This requires a classification of subvectors using Proposition 3. If the set of undefined subvectors is non-empty the μm-vector, $\mathcal{H}$ will be classified as undefined in (cf. Theorem 2). For the Z, U, P partition different sets need to be identified. The non-dominated set can be identified with time complexity in $O(\mu(\log\mu)^{\max(1,m-2)})$ using the algorithm of Kung et al. [24]; all other sets and set-operations can be computed with time complexity in $\mathcal{O}(m\mu\log\mu)$ either using elementary algorithms or based on sorting [25].

In the remainder the algorithm computes gradient components for subvectors in P based on the definition in Theorem 1. This is done by m dimension sweeps, each one computing the partial derivatives of subvectors in P of the k-th objective function.

Following Theorem 1 starting from the subvector with highest k-th coordinate the algorithm adds one by one in descending order of the k-th coordinate the projected subvectors `q` to a balanced tree data structure `T` and computes the incremental change in the $(m-1)$-dimensional hypervolume indicator of the set of points processed so far (all points higher in the k-th coordinate as it follows from Theorem 1) caused by this insertion. For the computation it is only required to maintain the set of the non-dominated points in the $(m-1)$-dimensional projection among the points that have been processed so far in the k-th sweep. The tree data structure `T` is used to maintain this set and quickly identify dominated points to be removed. This way a fast amortized logarithmic-time update schemes (2-D) or amortized linear-time update schemes (3-D) for the hypervolume indicator can be achieved. These update algorithms can be derived from the algorithms described by Beume et al. [19] and, in more than three dimensions, by Guerreiro et al. [23]. For a discussion of the reformulation of these dimension sweep algorithms for computing the m-dimensional indicator, as incremental update schemes for $(m-1)$-dimensional hypervolume indicators, see Hupkens and Emmerich [26]. In the last step of the algorithm's iteration dominated points are removed from the tree. The time for this step amortizes to the cost of identifying a single dominated point, as elements can be removed only once.

Algorithm: GradMultiSweep
Input: μm vector $\mathbf{Y}$ with subvectors $\mathbf{y}^{(1)} \in \mathbb{R}^m, \ldots, \mathbf{y}^{(\mu)} \in \mathbb{R}^m$, reference point $\mathbf{r}$
Output: Partial derivatives $\frac{\partial \mathcal{H}}{\partial y_k^{(i)}}$, $i = 1, \ldots, \mu$; $k = 1, \ldots, m$.

1. Determine the partition Z, U, and P of the subvectors of $\mathbf{Y}$ using Proposition 3.
2. **if** $U \neq \emptyset$ **output** ("Partial derivatives might be only one-sided in " + U)
3. Assign 0 to all partial derivatives of subvectors in Z.
4. **Remark**: In the remainder compute partial derivatives for all subvectors in P.
5. **For** $k \in \{1, \ldots, m\}$
   (a) Compute P_k as the set of all $(m-1)$-dimensional projections of subvectors of P by omitting their k-th coordinate.
   (b) Add subvectors in P_k in descending order of the k-th coordinate to a queue Q.
   (c) Initialize tree data structure for collecting non-dominated point T as empty.
   (d) **While** Q is not empty:
      i. q ← Lop off first (greatest) element from the queue Q.
      ii. Compute increment $\Delta H(q, \mathtt{T})$ of $(m-1)$-dimensional hypervolume indicator when adding $\mathbf{q}$ to T using efficient update schemes (for $m = 2$ sorting can be used, for $m = 3$ see Beume et al. [19], and for $m \geq 4$ see Guerreiro et al.[23]).
      iii. Set $\frac{\partial \mathcal{H}}{\partial y^{(i(\mathbf{q}))}} = \Delta H(q, T)$, where $i(\mathbf{q})$ is the index that corresponds to the index of the original subvector in $\mathbf{Y}$ of which $\mathbf{q}$ is the projection.
      iv. Add $\mathbf{q}$ to T and remove all elements that are Pareto dominated in the $(m-1)$-dimensional projection by q from T.

Fig. 4. Computing gradient components

The following theorem summarizes the complexity results of computing gradients of the hypervolume in the objective function space:

Theorem 4. *Given a μm-set $\mathbf{Y}$ of μ concatenated vectors of size m with no duplicate coordinates among subvectors. Then the computation of all components $\partial \mathcal{H} / \partial y_k^{(i)}(\mathbf{Y})$ for $k = 1, \ldots, m$ has a time complexity in $\Theta(\mu \log \mu)$ for $m = 2, 3$ and a time complexity in $\mathcal{O}(\mu^2)$ for $m = 4$.*

Proof. The lower bound of $\Omega(\mu \log \mu)$ for $m = 2$ can be proven by reduction of uniform gap as in [19]. For a given set $\{u_1, \ldots, u_m\}$ we need to represent this set as an instance of to the hypervolume gradient in linear time by duplication of coordinates, yielding $(u_1, -u_1, u_2, -u_2, \ldots, u_\mu, -u_\mu)$. After computing the hypervolume partial derivatives for a reference point $\mathbf{r} = (\min_{i=1,\ldots,\mu}\{u_i\}, -\max_{i=1,\ldots,\mu}\{u_i\})$, the uniform gap is decided positive if and only if all non-zero partial derivatives are the same, which can be checked in a linear number of comparisons.

For proving a lower bound for $m = 3$, we show that there exists a linear time reduction of the hypervolume indicator in two dimensions to the problem of computing the gradient components in three dimensions. As the complexity of computing the hypervolume indicator in two dimensions was proven by Beume et al. [19] to be in $\Omega(\mu \log \mu)$, a time complexity faster than $\Omega(\mu \log \mu)$ would yield a contradiction. The reduction reads as follows: Given μ mutually non-dominated vectors in 2-D, say $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(\mu)}$, and assume they are all dominating the 2-D reference point $(r_1, r_2)^\top$. Now we can construct a problem with reference point $(r_1, r_2, 0)^\top$ and a μm-vector $\mathbf{Y} = (u_1^{(1)}, u_2^{(1)}, 1)^\top, \ldots, (u_1^{(1)}, u_2^{(1)}, \mu)^\top$, then $H(\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(\mu)}) = \sum_{i=1}^{\mu} \frac{\partial \mathcal{H}}{\partial y_3^{(k)}}$. □

Example 3. This example illustrates the computation of $\nabla\mathcal{H}$ at a μm-vector $\mathbf{Y} = ((12, 11, 7) \circ (1, 3, 7) \circ (3, 10, 8) \circ (14, 4, 5) \circ (6, 12, 4) \circ (-1, 2, 9))^\top$ using Algorithm 4. Reference point is $(0, 0, 0)^\top$. The algorithm first partitions the multi-set of subvectors into $U = \emptyset$, $Z = \{(1, 3, 7)^\top, (-1, 2, 9)^\top\}$ and $P = \{(12, 11, 7)^\top$ $(3, 10, 8)^\top$, $(14, 4, 5)^\top$ $(6, 12, 4)^\top\}$. All partial derivatives of the 2nd and 6th subvector are set to zero. Figure 5 visualizes a sweep of P for the final outer loop with index $k = 3$: We initialize the queue as $\mathtt{Q} = [(6, 12)^\top \rightsquigarrow (14, 4)^\top \rightsquigarrow (12, 11)^\top \rightsquigarrow (3, 10)^\top]$. The pictures from the left to the right Figure 5 show the situations right after each iteration of the inner loop. First the algorithm lops off $\mathbf{q}$ at the front of the queue and inserting it to $\mathtt{T}$. The hypervolume update in the $(m-1)$-dimensional projection to f_1 and f_2 is now 30 and the partial derivative $\partial\mathcal{H}/\partial y_3^{(3)}$ is set to this value, because 3 is the upper index of the subvector from which the current $\mathbf{q}$ originated. Now, $\mathtt{Q} = [(6, 12)^\top \rightsquigarrow (14, 4)^\top \rightsquigarrow (12, 11)^\top]$ and the tree contains element $(3, 10)^\top$. In the next iteration $\mathbf{q} = (12, 11)^\top$ is drawn from the queue. The hypervolume update is now 102 and assigned to $\partial\mathcal{H}/\partial y_3^{(1)}$, as 1 is the index of the subvector in $\mathbf{Y}$ from which $\mathbf{q}$ originated. The vector $(3, 10)^\top$ is removed from the tree, because $\mathbf{q} = (12, 11)^\top$ dominates it in the first two dimensions. The next two iterations will not remove points from the tree and the partial derivatives $\partial\mathcal{H}/\partial y_3^{(4)} = 8$ and $\partial\mathcal{H}/\partial y_3^{(5)} = 6$ will be computed in this order. Thereafter the queue is empty and the algorithm terminates.

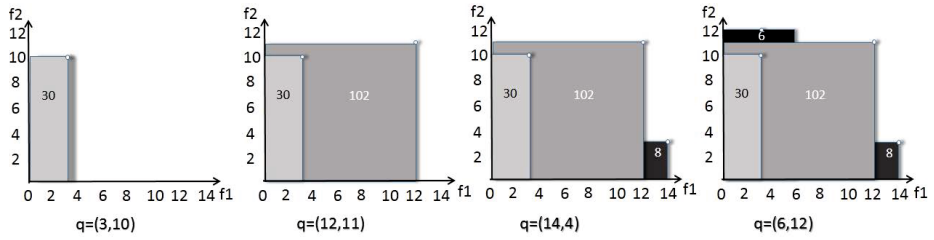

Fig. 5. Gradient computation for a 3-D gradient $\nabla\mathcal{H}(\mathbf{Y})$ for $\mathbf{Y} = ((3, 10, 8) \circ (12, 11, 7) \circ (14, 4, 5) \circ (6, 12, 4))^\top$ and reference point $\mathbf{r} = (0, 0, 0)^\top$

5.3 Time Complexity of Computing the Gradient at a μd-Vector

When putting the results of Theorems 4 and 3 together we obtain that the time complexity in the number of points in the approximation set, μ, is governed by the bounds given in Theorem 4. However, when dealing with a large number of dimensions the influence of m and d might be considerable. The cost for the matrix multiplication is influenced by the search space dimension and scales with $\mathcal{O}(\mu dm)$. Here μd is the same complexity as computing the gradients of all points in the approximation set and thus is at its lower bound.

Theorem 5. *Given an objective function* $\mathbf{f}$*, a* μd *vector* $\mathbf{X}$*, the partial derivatives* $\frac{\partial \mathcal{H}}{\partial y_k^i}(\mathbf{F}(\mathbf{X}))$ *and* $\frac{\partial f_k(\mathbf{x}^{(i)})}{\partial x_i^{(j)}}$ *for* $i = 1, ..., \mu$*;* $j = 1, \ldots, d$*; and* $k = 1, \ldots, m$ *the time complexity of computing all* μd *components of* $\frac{\partial \mathcal{H}_\mathbf{F}}{\partial x_j^{(i)}}(\mathbf{X})$ *of the hypervolume gradient* $\mathcal{H}_\mathbf{F}(\mathbf{X})$ *is given by* $\Theta(d\mu + \mu \log \mu)$ *in* $m = 2$ *and* $m = 3$ *dimensions, and by* $\mathcal{O}(\mu d + \mu^2)$ *in* $m = 4$ *dimensions.*

Proof. The output size is μd, therefore this is a lower bound for the complexity. Then the result follows from Theorem 3 and Theorem 4 and the fact that m is assumed to be constant. □

6 Gradient Components and Hypervolume Contributions

Revealing the relation between hypervolume contributions of points and the gradient components provides an important insight into the structure of the gradient field, that can yield (1) an alternative algorithm for computing hypervolume contributions, and (2) a concise formulation of an optimality criterion.

The hypervolume contribution $\Delta H(\mathbf{y}, Y)$ of a multi-set Y and a point $y \in Y$ is defined as:

$$\Delta H(\mathbf{y}, Y) = H(Y) - H(Y \setminus \{y\}) \tag{13}$$

Accordingly, define the hypervolume contribution $\Delta\mathcal{H}(i, Y), i = 1, \ldots, \mu$ of the i-th subvector in the μm-vector $\mathbf{Y}$ as the size of the truncated dominated subspace that is dominated by the i-th subvector but not by any other subvector. Putting this into more concrete terms, let $\pi_{1,\ldots,\overset{\times}{i},\ldots,\mu}(\mathbf{Y})$ denote the projection of $\mathbf{y}$ with the i-th subvector removed. Then

$$\Delta\mathcal{H}(i, \mathbf{Y}) = \mathcal{H}(\mathbf{Y}) - \mathcal{H}(\pi_{1,\ldots,\overset{\times}{i},\ldots\mu}(\mathbf{Y}))$$

From the geometrical situation described in Theorem 1 we obtain:

$$\nabla\Delta\mathcal{H}(i, \mathbf{Y}) = \left(\frac{\partial \Delta\mathcal{H}(i, \mathbf{Y})}{\partial y_1^{(i)}}, \ldots, \frac{\partial \Delta\mathcal{H}(i, \mathbf{Y})}{\partial y_m^{(i)}} \right)^\top = \tag{14}$$

$$= \left(\frac{\partial \mathcal{H}(\mathbf{Y})}{\partial y_1^{(i)}}, \ldots, \frac{\partial \mathcal{H}(\mathbf{Y})}{\partial y_m^{(i)}} \right)^\top \tag{15}$$

Furthermore, let us define the following subgradient at $\mathbf{X}$:

$$\nabla\mathcal{H}_{\mathbf{F}}(i,\mathbf{X}) = \left(\frac{\partial\mathcal{H}_{\mathbf{F}}}{\partial x_1^{(i)}},\ldots,\frac{\partial\mathcal{H}_{\mathbf{F}}}{\partial x_d^{(i)}}\right)^{\top},$$

that is $\mathcal{H}_{\mathbf{F}}(i,\mathbf{X})$ is equal to the i-th subvector of $\mathcal{H}_{\mathbf{F}}(\mathbf{X})$.

Let us recall the equation from Theorem 3:

$$\frac{\partial\mathcal{H}_{\mathbf{F}}}{\partial x_j^{(i)}}(\mathbf{X}) = \sum_{k=1}^{m}\frac{\partial\mathcal{H}}{\partial y_k^{(i)}}(\mathbf{F}(\mathbf{X}))\cdot\frac{\partial f_k(\mathbf{x}^{(i)})}{\partial x_j^{(i)}}, i=1,\ldots,\mu, j=1,\ldots,d. \tag{16}$$

It can be written in a compact form:

Theorem 6. *Let $\mathbf{f}'(\mathbf{x}^{(i)})$ denote the Jacobian matrix of $\mathbf{f}:\mathbb{R}^d\to\mathbb{R}^m$ at $\mathbf{x}^{(i)}$ and $\nabla\Delta H(i,\mathbf{F}(\mathbf{X}))$ the m partial derivatives of the hypervolume contribution. Then*

$$\nabla\mathcal{H}_{\mathbf{F}}(i,\mathbf{X}) = \nabla\Delta H(i,\mathbf{F}(\mathbf{X}))\cdot\mathbf{f}'(\mathbf{x}^{(i)}),\ i=1,\ldots,\mu. \tag{17}$$

Proof. This follows by rewriting Equation 16. □

According to this new interpretation of Theorem 3, it can be said that *the i-th subvector of $\nabla\mathbf{F}(\mathbf{X})$ is the gradient of the hypervolume contributions at the i-th subvector of* $\mathbf{X}$ for all other values in $\mathbf{X}$ being constant.

Remark 5 (Visualization of 2-D and 3-D gradient). The fact that the components of the gradient are related to the gradients of the hypervolume contributions can be used for a graphical representation of the gradient of $\mathcal{H}$ at a μm-vector. For each subvector (that is for each point in the Pareto front approximation) the gradient vector is drawn as an arrow starting in that point. Normalization by dividing by the length of the subgradient, that is $||\nabla\Delta\mathcal{H}(i,\mathbf{Y})||$, makes the visualization more readable. Examples follows.

Example 4. The visualization in Figure 6 is based on the data of Example 3 and subvectors that contribute only zero gradient components are omitted.

Example 5. In Figure 7 a visualization for $m=2$ and $\mu=5$ is depicted. See Figure 8 for an example with $m=3$ and with 100 points distributed randomly on the positive part of a sphere with radius 10 and a reference point of 0. Here normalization is used to make the picture more transparent.

Remark 6 (Implementation of 3-D Gradient). To implement the 3-D example in Figure 8 a fast computation of the 3-D Gradient field computation has been implemented in C++. It is based on the algorithm of Fonseca and Emmerich [27] that computes all contributions to the hypervolume indicator within a single sweep and with a time complexity in $\mathcal{O}(\mu\log\mu)$. This algorithm can be easily modified to compute the visible facets of the volumes that are dominated by precisely one single subvector and therewith the components of $\nabla\mathcal{H}$ at some μm vector within a *single* sweep. The details of this implementation are omitted in this paper, but the code is made available under `http://natcomp.liacs.nl`.

6.1 Optimality Conditions

From the theoretical observations in Theorem 6 necessary conditions for optimality of μd vectors w.r.t. the hypervolume indicator can be stated in a concise way.

Let us restrict our attention first to differentiable μd vectors $\mathbf{X}$ with all subvectors of $\mathbf{F}(\mathbf{X})$ being non-dominated and in the interior of $[\mathbf{r}, \infty)$. These μd vectors will be termed *proper μd-vectors*. Note, that for proper μd vectors all partial derivatives of $\mathbf{H}(\mathbf{F}(\mathbf{X}))$ are non-zero.

As $\mathbf{H}_F$ is differentiable in $\mathbf{X}$ the following optimality condition holds:

Theorem 7. *A necessary condition for $\mathcal{H}_F$ being optimal is that*

$$\nabla \Delta H(i, \mathbf{F}(\mathbf{X})) \cdot \mathbf{f}'(\mathbf{x}^{(i)}) = 0 \tag{18}$$

for all $i = 1, \ldots, \mu$, or in different notation

$$\sum_{k=1}^{m} \frac{\partial \Delta \mathcal{H}(i, \mathbf{F}(\mathbf{X}))}{\partial y_k^{(i)}} \cdot \frac{\partial f_k(\mathbf{x}^{(i)})}{\partial x_j^{(i)}} = 0 \tag{19}$$

for all $i = 1, \ldots, \mu; j = 1, \ldots, d$.

Proof. This is the usual condition for stationarity of points and decomposition of the gradient described in Theorem 6. □

Loosely speaking, Theorem 7 means that by finding solutions for which all hypervolume contribution gradients turn zero, candidates for optimal approximation sets can be obtained. This observation yields μd equations to be satisfied for μd variables to be determined.

We note that this condition holds also for non-proper μd vectors, although we can already a-priori conclude that optima of non-proper μd-vectors are of minor interest. If our aim is to approximate an non-degenerate Pareto front, that is a $(m-1)$-dimensional manifold, every one of the μ points in the approximation set should contribute.

A close look at Equation 18 shows that there can be two reasons that a proper μd-vector satisfies the equation for a particular index i:

1. Some components of the Jacobian matrix are zero.
2. The partial derivatives of the contribution (which remain constant for a fixed value of d) are canceled out by the components of the column vectors of the Jacobian matrix.

In the unconstrained bi-objective case the Fritz John necessary conditions for a differentiable point $\mathbf{x}$ (with respect to f_1 and f_2) to belong to the efficient set read:

$$\exists \lambda_1, \lambda_2 \geq 0 : \lambda_1 \neq \lambda_2 \text{ and } \lambda_1 \nabla f_1(\mathbf{x}) + \lambda_2 \nabla f_2(\mathbf{x}) = 0 \tag{20}$$

This means either at least one of the gradient vectors is zero, or the gradient vectors point in the opposite direction.

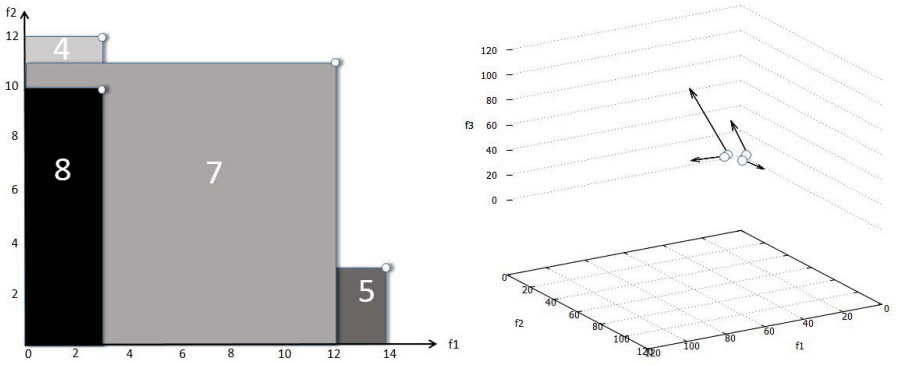

Fig. 6. Gradient computation for a 3-D gradient $\nabla\mathcal{H}(\mathbf{Y})$ for $\mathbf{Y} = ((3, 10, 8) \circ (12, 11, 7) \circ (14, 4, 5) \circ (6, 12, 4))^\top$ and reference point $\mathbf{r} = (0, 0, 0)^\top$

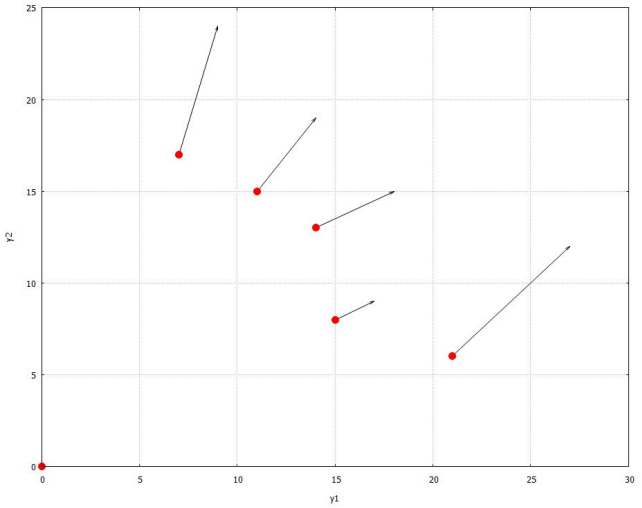

Fig. 7. Gradient computation for a 2-D gradient $\nabla\mathcal{H}(\mathbf{Y})$ for $\mathbf{Y} = ((7, 17) \circ (11, 15) \circ (14, 13) \circ (15, 8) \circ (21, 6))^\top$ and reference point $\mathbf{r} = (0, 0, 0)^\top$

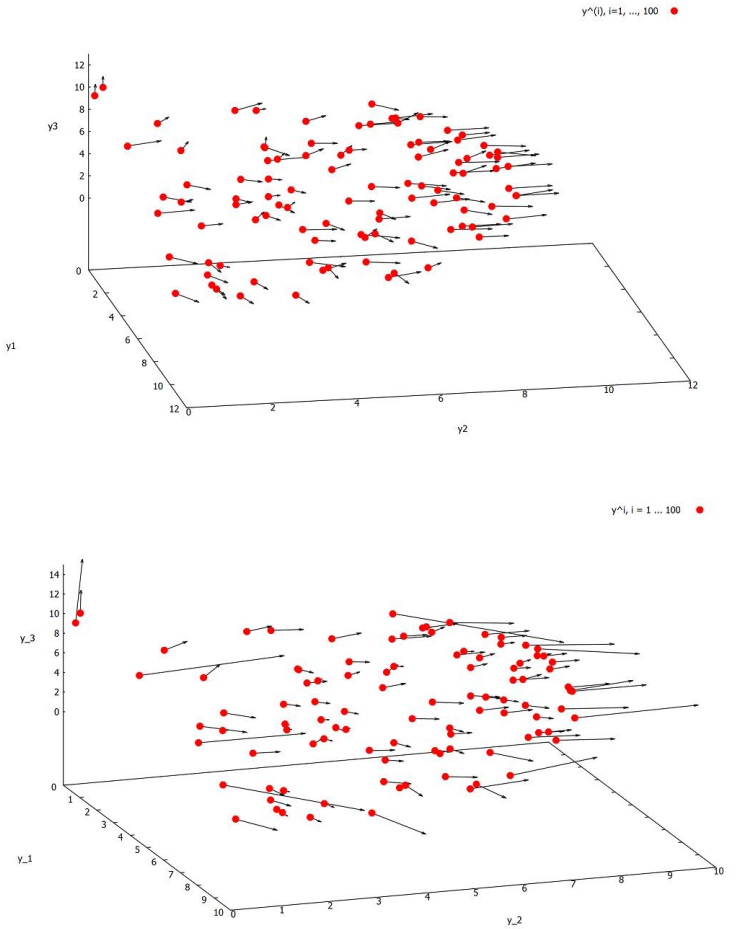

Fig. 8. Normalized gradient of $\mathcal{H}$ at $\mathbf{Y}$ (upper picture) and non-normalized gradient at $\mathbf{Y}$, for $\mathbf{Y}$ given as a random set of objective vectors distributed randomly on the positive section of a sphere with radius 10

We can combine this with the previous result and obtain:

Corollary 1. *A necessary condition for a proper μd-vector consisting of efficient subvectors with respect to the bi-objective optimization problem to represent a locally optimal approximation set of the hypervolume indicator is given by:*

$$\frac{\frac{\partial \mathcal{H}}{\partial y_1^{(i)}}(\mathbf{F}(\mathbf{X}))}{\frac{\partial \mathcal{H}}{\partial y_2^{(i)}}(\mathbf{F}(\mathbf{X}))} = \frac{||\nabla f_2(\mathbf{x}^{(i)})||}{||\nabla f_1(\mathbf{x}^{(i)})||} \tag{21}$$

for all $i = 1, \ldots, \mu$.

In other words, the differential change in the i-th subvector of $\mathbf{X}$ in the decision space causes a growth of the hypervolume contribution of the i-th subvector of $\mathbf{F}(X)$ in the y_1 direction, that is compensated by a decrease of the hypervolume contribution of that subvector in direction y_2.

It is expected that a more general analysis of the findings presented in this section will reveal refined optimality conditions and a better understanding of the properties of (locally) optimal approximation sets. It may also shed new light on the yet unanswered question of how points in bounded size sets that maximize the hypervolume indicator distribute on a given Pareto front (see also [28]).

7 Conclusions and Outlook

This chapter refined the definition of the hypervolume indicator gradient field for the higher dimensional case. The size of the faces of the boundary of the dominated subspace are the gradient components at a set of objective vectors in the decision space. Partial derivatives of the gradient can be readily computed by using algorithms for computing the incremental hypervolume contributions. This yields algorithms with asymptotically optimal computational time complexity $\Theta(\mu d + \mu \log \mu)$ for computing the gradient at an approximation set from the Jacobian matrices of $\mathbf{f}$ at the points, and the values of the objective vectors in the bi- and tri-objective case. In the four objective case the time complexity can be guaranteed to be in $\mathcal{O}(\mu d + \mu^2)$. Further progress in incremental update schemes for the hypervolume indicator will also yield sharper bounds for gradient computations. Finally by deriving tight computation schemes, structural properties of the hypervolume indicator gradient field were revealed that entail a set of μd simple equations to be satisfied for an proper approximation set to be optimal. The analysis of these conditions may shed a light on the fundamental laws that govern the distribution of points in hypervolume indicator optimal approximations sets to the Pareto front (see also [28]). Moreover, the formulation of stopping criteria that guarantee local optimality for hypervolume-indicator based Pareto optimization is now in reach.

References

1. Emmerich, M.T.M., Deutz, A.H., Beume, N.: Gradient-Based/Evolutionary Relay Hybrid for Computing Pareto Front Approximations Maximizing the S-Metric. In: Bartz-Beielstein, T., Blesa Aguilera, M.J., Blum, C., Naujoks, B., Roli, A., Rudolph, G., Sampels, M. (eds.) HCI/ICCV 2007. LNCS, vol. 4771, pp. 140–156. Springer, Heidelberg (2007)
2. Fliege, J., Svaiter, B.F.: Steepest Descent Methods for Multicriteria Optimization. Mathematical Methods of Operations Research 51(3), 479–494 (2000)
3. Brown, M., Smith, R.E.: Effective Use of Directional Information in Multi-objective Evolutionary Computation. In: Cantú-Paz, E., et al. (eds.) GECCO 2003. LNCS, vol. 2723, pp. 778–789. Springer, Heidelberg (2003)
4. Bosman, P.A., de Jong, E.D.: Exploiting Gradient Information in Numerical Multi-Objective Evolutionary Optimization. In: Beyer, H.G., et al. (eds.) GECCO 2005, vol. 1, pp. 755–762. ACM Press, New York (2005)
5. Lara, A., Schütze, O., Coello, C.A.C.: On Gradient-Based Local Search to Hybridize Multi-objective Evolutionary Algorithms. In: Tantar, E., Tantar, A.-A., Bouvry, P., Del Moral, P., Legrand, P., Coello Coello, C.A., Schütze, O. (eds.) EVOLVE- A bridge between Probability, Set Oriented Numerics and Evolutionary Computation. SCI, vol. 447, pp. 303–330. Springer, Heidelberg (2013)
6. Timmel, G.: Ein stochastisches Suchverfahren zur Bestimmung der Optimalen Kompromißlösungen bei statistischen polykriteriellen Optimierungsaufgaben. Journal TH Ilmenau 6, 139–148 (1980)
7. Schäffler, S., Schultz, R., Wienzierl, K.: Stochastic Method for the Solution of Unconstrained Vector Optimization Problems. Journal of Optimization Theory and Applications 114(1), 209–222 (2002)
8. Shukla, P.K., 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)
9. Hillermeier, C.: Generalized Homotopy Approach to Multiobjective Optimization. Journal of Optimization Theory and Applications 110(3), 557–583 (2001)
10. Schütze, O., Dell'Aere, A., Dellnitz, M.: 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, vol. 04461. IBFI, Schloss Dagstuhl, Germany (2005)
11. Schütze, O., Lara, A., Coello Coello, C.A.: The Directed Search Method for Unconstrained Multi-Objective Optimization Problems. In: Proceedings of the EVOLVE–A Bridge Between Probability, Set Oriented Numerics, and Evolutionary Computation (2011)
12. Ehrgott, M.: Multicriteria Optimization. Springer (2005)
13. 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 1998. LNCS, vol. 1498, pp. 292–301. Springer, Heidelberg (1998)
14. Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C.M., da Fonseca, V.G.: Performance Assessment of Multiobjective Optimizers: an Analysis and Review. IEEE Trans. Evolutionary Computation 7(2), 117–132 (2003)

15. Auger, A., Bader, J., Brockhoff, D., Zitzler, E.: Hypervolume-based Multiobjective Optimization: Theoretical Foundations and Practical Implications. Theor. Comput. Sci. 425, 75–103 (2012)
16. Beume, N.: Hypervolume-Based Metaheuristics for Multiobjective Optimization. PhD Thesis. Eldorado (2011)
17. Custódio, A.L., Emmerich, M., Madeira, J.F.A.: Recent Developments in Derivative-free Multiobjective Optimization. In: Topping, B. (ed.) Computational Technology Reviews, vol. 5, pp. 1–30. Saxe-Coburg Publications (2012)
18. Bringmann, K.: Bringing Order to Special Cases of Klee's Measure Problem. CoRR abs/1301.7154 (2013)
19. Beume, N., Fonseca, C.M., López-Ibáñez, M., Paquete, L., Vahrenhold, J.: On the Complexity of Computing the Hypervolume Indicator. IEEE Trans. Evolutionary Computation 13(5), 1075–1082 (2009)
20. Yıldız, H., Suri, S.: On Klee's Measure Problem for Grounded Boxes. In: Dey, T.K., Whitesides, S. (eds.) Symposium on Computational Geometry, pp. 111–120. ACM (2012)
21. Fonseca, C.M., Guerreiro, A.P., López-Ibáñez, M., Paquete, L.: On the Computation of the Empirical Attainment Function. In: [29], pp. 106–120
22. 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)
23. Guerreiro, A.P., Fonseca, C.M., Emmerich, M.T.M.: A Fast Dimension-Sweep Algorithm for the Hypervolume Indicator in Four Dimensions. In: CCCG, pp. 77–82 (2012)
24. Kung, H.T., Luccio, F., Preparata, F.P.: On Finding the Maxima of a Set of Vectors. Journal of the ACM 22(4), 469–476 (1975)
25. Baeza-Yates, R.: A Fast Set Intersection Algorithm for Sorted Sequences. In: Sahinalp, S.C., Muthukrishnan, S.M., Dogrusoz, U. (eds.) CPM 2004. LNCS, vol. 3109, pp. 400–408. Springer, Heidelberg (2004)
26. Hupkens, I., Emmerich, M.: Logarithmic-time Updates in SMS-EMOA and Hypervolume-based Archiving. In: Emmerich, M., et al. (eds.) EVOLVE - A Bridge between Probability, Set Oriented Numerics,and Evolutionary Computation IV. AISC, vol. 227, pp. 155–169. Springer, Heidelberg (2013)
27. Emmerich, M.T.M., Fonseca, C.M.: Computing Hypervolume Contributions in Low Dimensions: Asymptotically Optimal Algorithm and Complexity Results. In: [27], pp. 121–135
28. Auger, A., Bader, J., Brockhoff, D., Zitzler, E.: Theory of the Hypervolume Indicator: Optimal μ-Distributions and the Choice of the Reference Point. In: Foundations of Genetic Algorithms (FOGA 2009), pp. 87–102. ACM, New York (2009)
29. Takahashi, R.H.C., Deb, K., Wanner, E.F., Greco, S. (eds.): EMO 2011. LNCS, vol. 6576. Springer, Heidelberg (2011)