Bounding the Effectiveness of Hypervolume-based $(\mu + \lambda)$-Archiving Algorithms

Tamara Ulrich and Lothar Thiele

Computer Engineering and Networks Laboratory, ETH Zurich,
8092 Zurich, Switzerland
firstname.lastname@tik.ee.ethz.ch

Abstract. In this paper, we study bounds for the α-approximate effectiveness of non-decreasing $(\mu+\lambda)$-archiving algorithms that optimize the hypervolume. A $(\mu + \lambda)$-archiving algorithm defines how μ individuals are to be selected from a population of μ parents and λ offspring. It is non-decreasing if the μ new individuals never have a lower hypervolume than the μ original parents. An algorithm is α-approximate if for any optimization problem and for any initial population, there exists a sequence of offspring populations for which the algorithm achieves a hypervolume of at least $1/\alpha$ times the maximum hypervolume.

Bringmann and Friedrich (GECCO 2011, pp. 745–752) have proven that all non-decreasing, locally optimal $(\mu+1)$-archiving algorithms are $(2+\epsilon)$-approximate for any $\epsilon > 0$. We extend this work and substantially improve the approximation factor by generalizing and tightening it for any choice of λ to $\alpha = 2-(\lambda-p)/\mu$ with $\mu = q\cdot\lambda-p$ and $0 \leq p \leq \lambda-1$. In addition, we show that $1 + \frac{1}{2\lambda} - \delta$, for $\lambda < \mu$ and for any $\delta > 0$, is a lower bound on α, i.e. there are optimization problems where one can not get closer than a factor of $1/\alpha$ to the optimal hypervolume.

Keywords: Multiobjective Evolutionary Algorithms, Hypervolume, Submodular Functions

1 Introduction

When optimizing multiple conflicting objectives, there usually is no single best solution. Instead, there are incomparable tradeoff solutions, where no solution is strictly better than any other solution. *Better* in this case refers to Pareto-dominance, i.e. one solution is said to be better than another, or dominate it, if it is equal or better in all objectives, and strictly better in at least one objective. The set of non-dominated solutions is called the Pareto-optimal set. Usually, this Pareto-optimal set can contain a large number of solutions, and it is infeasible to calculate all of them. Instead, one is interested in finding a relatively small, but still *good* subset of this Pareto-optimal set.

It is not a priori clear how a good subset should look like, i.e. how the goodness of a subset can be measured. One of the most popular measures for subset quality is the hypervolume indicator, which measures the volume of the dominated space. Therefore, one possibility to pose a multiobjective optimization problem is to look for a solution set $\mathcal{P}^*$ of fixed size, which maximizes the hypervolume.

Algorithms that optimize the hypervolume face several problems. First, the number of possible solutions can become very large, so it is not possible to select from all solutions. Second, even if all solutions are known and the non-dominated solutions can be identified, the number of subsets explodes and not all of them can be enumerated for comparison.

In this paper, we consider $(\mu + \lambda)$-evolutionary algorithms, or $(\mu + \lambda)$-EAs. They iteratively improve a set of solutions, where the set is named 'population' and the iteration is denoted as 'generation'. In particular, they maintain a population of size μ, generate λ offspring from the μ parents and then select μ solutions from the μ parents and the λ offspring that are to survive into the next generation. Note that we here only consider non-decreasing algorithms, i.e. algorithms whose hypervolume cannot decrease from one generation to the next.

Several questions arise in this setting. First, what are upper and lower bounds on the hypervolume that a population of a fixed size will achieve? Is it possible to prove that a set of size μ with the maximal hypervolume can be found, without explicitly testing all possible sets? To answer these questions, the term *effectiveness* has been defined. An algorithm is effective if for any optimization problem[1] and for any initial population[2], there is a sequence of offspring[3] which leads to the population with maximum hypervolume. Obviously, $(\mu + \mu)$-EAs are always effective: We just choose the first set of offspring to be exactly the population with the maximal hypervolume and then we select this set as the new population. It has also been shown by Zitzler et al.[4] that $(\mu + 1)$-EAs, on the other hand, are ineffective. Recently, it has been shown by Bringmann and Friedrich [1] that all $(\mu + \lambda)$-EAs with $\lambda < \mu$ are ineffective.

Bringmann and Friedrich then raised the follow-up question: If it is not possible to reach the optimal hypervolume for all optimization problems and all initial populations, is it at least possible to give a lower bound on the achieved hypervolume? To this end, they introduced the term *α-approximate effectiveness*. An algorithm is α-approximate if for any optimization problem and for any initial population there is a sequence of offspring with which the algorithm achieves at

[1] We only consider finite search spaces here, such that mutation operators exist which produce offspring with a probability larger than zero. Note that any search space coded on a computer is finite.

[2] Note that the term *for any initial population* implies that at any point during the algorithm, there exists a sequence of offspring with which an effective algorithm can achieve the optimal hypervolume.

[3] Note that the term *there is a sequence of offspring* assumes that we are given variation operators that produce any sequence of offspring with probability greater than zero.

least $1/\alpha \cdot H^{\max}$, where $H^{\max}$ is the maximum achievable hypervolume of a population of size μ. They proved in their paper that a $(\mu+1)$-EA is 2-approximate and conjectured that for larger λ, a $(\mu+\lambda)$-EA is $O(1/\lambda)$-approximate.

On the other hand, we might also be interested in upper bounds on the achievable hypervolume. Bringmann and Friedrich [1] have found an optimization problem where no algorithm can achieve more than $1/(1+0.1338(1/\lambda-1/\mu)-\epsilon)$ of the optimal hypervolume, i.e. there is no $(1+0.1338(1/\lambda-1/\mu)-\epsilon)$-approximate archiving algorithm for any $\epsilon > 0$.

Why is knowledge of the bounds of the α-approximate effectiveness useful? Assume that we are using an exhaustive mutation operator, which produces any offspring with a probability larger than zero. Therefore, the probability of generating an arbitrary sequence of offspring is also larger than zero. The $\frac{1}{2}$-approximate effectiveness of $(\mu+1)$-EAs now tells us that if we execute the evolutionary algorithm for a sufficiently large number of generations, we will end up with a population that has at least half of the maximal hypervolume. In case of a $(\mu+\mu)$-EA, on the other hand, we know that we will finally achieve a population with maximum hypervolume, i.e. $\alpha = 1$. We are therefore interested in deriving bounds on the effectiveness of evolutionary algorithms.

This paper extends the work of Bringmann and Friedrich by (a) computing the α-approximate effectiveness of $(\mu+\lambda)$-EAs for general choices of λ, (b) tightening the previously known upper bound on α, and (c) tightening the previously known lower bound on α. The results for (a) and (b) are based on the theory of submodular functions, see [2]. For (c) we show that for $\lambda < \mu$, there exist optimization problems where any $(\mu+\lambda)$-EA does not get closer than a factor of $1/\alpha$ to the optimal hypervolume with $\alpha = 1 + \frac{1}{2\lambda} - \delta$, for any $\delta > 0$.

The paper is organized as follows: The next section presents the formal setting, including the definition of the hypervolume, the algorithmic setting, definitions for the effectiveness and approximate effectiveness and an introduction into submodular functions. In Section 3 we determine an upper bound on α for general choices of μ and λ, thereby giving a quality guarantee in terms of a lower bound of the achievable hypervolume. Finally in Section 4, we will determine a lower bound on α for general choices of μ and λ.

2 Preliminaries

Consider a multiobjective minimization problem with a decision space $\mathcal{X}$ and an objective space $\mathcal{Y} \subseteq \mathbb{R}^m = \{f(x)|x \in \mathcal{X}\}$, where $f : \mathcal{X} \to \mathcal{Y}$ denotes a mapping from the decision space to the objective space with m objective functions $f = \{f_1, ..., f_m\}$ which are to be minimized.

The underlying preference relation is weak Pareto-dominance, where a solution $a \in \mathcal{X}$ weakly dominates another solution $b \in \mathcal{X}$, denoted as $a \preceq b$, if and only if solution a is better or equal than b in all objectives, i.e. iff $f(a) \leqslant f(b)$, or equivalently, iff $f_i(a) \leq f_i(b), \forall i \in \{1, ..., m\}$. In other words, a point $p \in \mathcal{X}$ weakly dominates the region $\{y \in \mathbb{R}^m : f(p) \leqslant y\} \subset \mathbb{R}^m$.

2.1 Hypervolume Indicator

The hypervolume indicator of a given set $\mathcal{P} \subseteq \mathcal{X}$ is the volume of all points in $\mathbb{R}^m$ which are dominated by at least one point in $\mathcal{P}$ and which dominate at least one point of a reference set $\mathcal{R} \subset \mathbb{R}^m$.[4] Roughly speaking, the hypervolume measures the size of the dominated space of a given set. Sets with a larger hypervolume are considered better. More formally, the hypervolume indicator can be written as

$$H(\mathcal{P}) := \int_{y \in \mathbb{R}^m} A_{\mathcal{P}}(y)\, dy$$

where $A_{\mathcal{P}}(y)$ is called the *attainment function* of set $\mathcal{P}$ with respect to a given reference set $\mathcal{R}$, and is defined as follows:

$$A_{\mathcal{P}}(y) = \begin{cases} 1 & \text{if } \exists p \in \mathcal{P},\, r \in \mathcal{R} \,:\, f(p) \leqslant y \leqslant r \\ 0 & \text{else} \end{cases}$$

The goal of a $(\mu + \lambda)$-EA is to find a population $\mathcal{P}^* \subseteq \mathcal{X}$ of size μ with the maximum hypervolume:

$$H(\mathcal{P}^*) = \max_{\mathcal{P} \subseteq \mathcal{X}, |\mathcal{P}| = \mu} H(\mathcal{P}) = H_\mu^{\max}(\mathcal{X})$$

2.2 Algorithmic Setting

Algorithm 1 General $(\mu + \lambda)$-EA framework: μ denotes the population size; λ the offspring size; the algorithm runs for g generations.

```
1: function EA($\mu$, $\lambda$, $g$)
2:     $\mathcal{P}^0 \leftarrow$ initialize with $\mu$ random solutions
3:     for $t = 1$ to $g$ do
4:         $\mathcal{O}^t \leftarrow$ generate $\lambda$ offspring
5:         $\mathcal{P}^t \leftarrow$ select $\mu$ solutions from $\mathcal{P}^{t-1} \cup \mathcal{O}^t$
6:     end for
7:     return $\mathcal{P}^g$
8: end function
```

The general framework we are considering here is based on a $(\mu + \lambda)$ evolutionary algorithm (EA) as shown in Algorithm 1. The selection step of Line 5 is done by a $(\mu + \lambda)$-archiving algorithm[5]. We here assume that the archiving algorithm is non-decreasing, i.e. $H(\mathcal{P}^t) \geq H(\mathcal{P}^{t-1})$, $1 \leq t \leq g$. We use the following formal definition (as given in [1]) to describe an archiving algorithm:

[4] No assumptions on the reference set have to be made, as our results have to hold for any objective space, including the one only containing solutions that dominate at least one reference point. If that set is empty, all algorithms are effective, as the hypervolume is always zero.

[5] We use the term *archiving algorihm* here to be compliant with [1]. It does not mean that we keep a separate archive in addition to the population $\mathcal{P}^t$.

Definition 1. *A $(\mu+\lambda)$-*archiving algorithm *A is a partial mapping $A : 2^{\mathcal{X}} \times 2^{\mathcal{X}} \rightarrow 2^{\mathcal{X}}$ such that for a μ-population $\mathcal{P}$ and a λ-population $\mathcal{O}$, $A(\mathcal{P},\mathcal{O})$ is a μ-population and $A(\mathcal{P},\mathcal{O}) \subseteq \mathcal{P} \cup \mathcal{O}$.*

Using this definition, the for-loop in Algorithm 1 can be described as follows, see also [1]:

Definition 2. *Let $\mathcal{P}^0$ be a μ-population and $\mathcal{O}^1, ..., \mathcal{O}^N$ a sequence of λ-populations. Then*

$$\mathcal{P}^t := A(\mathcal{P}^{t-1}, \mathcal{O}^t) \quad \textit{for all } t = 1, ..., N$$

We also define

$$\begin{aligned} A(\mathcal{P}^0, \mathcal{O}^1, ..., \mathcal{O}^t) &:= A(A(\mathcal{P}^0, \mathcal{O}^1, ..., \mathcal{O}^{t-1}), \mathcal{O}^t) \\ &= A(...A(A(\mathcal{P}^0, \mathcal{O}^1), \mathcal{O}^2), ..., \mathcal{O}^t) \\ &= \mathcal{P}^t \quad \textit{for all } t = 1, ..., N \end{aligned}$$

As mentioned above, we only consider non-decreasing archiving algorithms which are defined as follows, see also [1]:

Definition 3. *An archiving algorithm A is non-decreasing, if for all inputs $\mathcal{P}$ and $\mathcal{O}$, we have*

$$H(A(\mathcal{P},\mathcal{O})) \geq H(\mathcal{P})$$

2.3 Effectiveness and Approximate Effectiveness

Following Bringmann and Friedrich [1], we here assume a worst-case view on the initial population and a best-case view on the choice of offspring. This means that we would like to know for any optimization problem, starting from any initial population, whether there exists a sequence of offspring populations such that the EA is able to find a population with the maximum possible hypervolume. If so, the archiving algorithm is called *effective*:

Definition 4. *A $(\mu+\lambda)$-archiving algorithm A is* effective*, if for all finite sets $\mathcal{X}$, all objective functions f and all μ-populations $\mathcal{P}^0 \subseteq \mathcal{X}$, there exists an $N \in \mathbb{N}$ and a sequence of λ-populations $\mathcal{O}^1, ..., \mathcal{O}^N \subseteq \mathcal{X}$ such that*

$$H(A(\mathcal{P}^0, \mathcal{O}^1, ..., \mathcal{O}^N)) = H_\mu^{max}(\mathcal{X})$$

Similarly, we use the following definition for the approximate effectiveness, which quantifies the distance to the optimal hypervolume that can be achieved:

Definition 5. *Let $\alpha \geq 1$. A $(\mu+\lambda)$-archiving algorithm A is* α-approximate *if for all finite sets $\mathcal{X}$, all objective functions f and all μ-populations $\mathcal{P}^0 \subseteq \mathcal{X}$, there exists an $N \in \mathbb{N}$ and a sequence of λ-populations $\mathcal{O}^1, ..., \mathcal{O}^N$ such that*

$$H(A(\mathcal{P}^0, \mathcal{O}^1, ..., \mathcal{O}^N)) \geq \frac{1}{\alpha} H_\mu^{max}(\mathcal{X})$$

Of course, an effective archiving algorithm is 1-approximate. Here, we are interested in deriving bounds on α for any choice of μ and λ.

2.4 Submodular Functions

The theory of submodular functions has been widely used to investigate problems where one is interested in selecting optimal subsets of a given size. But what exactly is a submodular function? At first, they map subsets of a given base set to real numbers, just like the hypervolume indicator defined above. In addition, submodular functions show a diminishing increase when adding points to sets that become larger. In other words, let us define the set function $z : 2^{\mathcal{X}} \to \mathbb{R}$, where $2^{\mathcal{X}}$ is the power set of the decision space. Then the contribution of a point $s \in \mathcal{X}$ with respect to set $\mathcal{A} \subset \mathcal{X}$ is $c(s, \mathcal{A}) = z(\mathcal{A} \cup \{s\}) - z(\mathcal{A})$. When z is a submodular function, the contribution $c(s, \mathcal{A})$ gets smaller when $\mathcal{A}$ becomes larger. More formally, a submodular function z is defined as follows:

$$\forall \mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{X}, \forall s \in \mathcal{X} \backslash \mathcal{B} : z(\mathcal{A} \cup \{s\}) - z(\mathcal{A}) \geq z(\mathcal{B} \cup \{s\}) - z(\mathcal{B}) \tag{1}$$

i.e. if set $\mathcal{A}$ is contained in set $\mathcal{B}$, the contribution of adding a point s to $\mathcal{A}$ is larger or equal than the contribution of adding s to $\mathcal{B}$. A submodular function is non-decreasing if it is monotone in adding points:

$$\forall \mathcal{B} \subseteq \mathcal{X}, \forall s \in \mathcal{X} \backslash \mathcal{B} : z(\mathcal{B} \cup \{s\}) \geq z(\mathcal{B})$$

Now, we show that the hypervolume indicator as defined above is non-decreasing and submodular.

Theorem 1. *The hypervolume indicator $H(\mathcal{P})$ is non-decreasing submodular.*

Proof. At first, we define the contribution of a solution s to a set $\mathcal{B}$ as

$$H(\mathcal{B} \cup \{s\}) - H(\mathcal{B}) = \int_{y \in \mathbb{R}^m} C(\mathcal{B}, s, y)\, dy$$

with

$$C(\mathcal{B}, s, y) = A_{\mathcal{B} \cup \{s\}}(y) - A_{\mathcal{B}}(y)$$

Using the definition of the attainment function A we find

$$C(\mathcal{B}, s, y) = \begin{cases} 1 & \text{if } (\exists r \in \mathcal{R} \,:\, f(s) \leqslant y \leqslant r) \wedge (\nexists p \in \mathcal{B} : f(p) \leqslant y) \\ 0 & \text{else} \end{cases}$$

As $C(\mathcal{B}, s, y)$ is non-negative, the hypervolume indicator is non-decreasing.

Consider two arbitrary sets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{X}$ with $\mathcal{A} \subseteq \mathcal{B}$, and an arbitrary solution $s \in \mathcal{X}$, $s \notin \mathcal{B}$. To prove that the hypervolume indicator is submodular, we have to show that

$$H(\mathcal{A} \cup s) - H(\mathcal{A}) \geq H(\mathcal{B} \cup s) - H(\mathcal{B}) \tag{2}$$

or equivalently

$$\int_{y \in \mathbb{R}^m} C(\mathcal{A}, s, y)\, dy \geq \int_{y \in \mathbb{R}^m} C(\mathcal{B}, s, y)\, dy \tag{3}$$

for $\mathcal{A} \subseteq \mathcal{B}$, $s \notin \mathcal{B}$.

To this end, we will show that for all $y \in \mathbb{R}^m$ the inequality $C(\mathcal{A}, s, y) \geq C(\mathcal{B}, s, y)$ holds. As $C(\cdot, \cdot, \cdot)$ can only assume the values 0 and 1, we have to show that for all $y \in \mathbb{R}^m$, $s \notin \mathcal{B}$ we have

$$C(\mathcal{A}, s, y) = 0 \quad \Rightarrow \quad C(\mathcal{B}, s, y) = 0$$

Following the definition of C, there are the following three cases where $C(\mathcal{A}, s, y) = 0$:

1. ($\nexists r \in \mathcal{R} : y \leqslant r$): In this case, we also have $C(\mathcal{B}, s, y) = 0$ as the condition is the same for $C(\mathcal{A}, s, y)$ and $C(\mathcal{B}, s, y)$.
2. ($f(s) \nleqslant r$): Again, we find $C(\mathcal{B}, s, y) = 0$ as the condition is the same for $C(\mathcal{A}, s, y)$ and $C(\mathcal{B}, s, y)$.
3. ($\exists p \in \mathcal{A} : f(p) \leqslant y$): In other words, there exists a solution $p \in \mathcal{A}$ in $\mathcal{A}$ which weakly dominates y. But as $\mathcal{A} \subseteq \mathcal{B}$, we also have $p \in \mathcal{B}$ and therefore, ($\exists p \in \mathcal{B} : f(p) \leqslant y$). Therefore, we find $C(\mathcal{B}, s, y) = 0$.

As a result, (3) holds and the hypervolume indicator is submodular. □

3 Upper Bound on the Approximate Effectiveness

In this section, we will provide quality guarantees on the hypervolume achieved by an EA in terms of the α-approximate effectiveness, i.e. we will provide an upper bound on α for all population sizes μ and offspring set sizes λ.

In the previous section, we showed that the hypervolume is non-decreasing submodular. Nemhauser, Wolsey and Fisher [3] have investigated interchange heuristics for non-decreasing set functions and showed approximation properties in case of submodular set functions. We will first show that the interchange heuristic in [3] is execution-equivalent to the previously defined (μ+λ)-EA framework. Then, the approximation properties for the R-interchange heuristics are used to determine upper bounds on α.

The heuristic described in [3] is shown in Algorithm 2 where we deliberately changed the variable names to make them fit to the notations introduced so far. It makes use of the difference between sets, which is defined as follows: Given two sets $\mathcal{A}$ and $\mathcal{B}$, the difference between $\mathcal{A}$ and $\mathcal{B}$ is $\mathcal{A} - \mathcal{B} = \{x : x \in A \wedge x \notin B\}$, i.e. the set of all solutions which are contained in $\mathcal{A}$ but not in $\mathcal{B}$.

The heuristic in Algorithm 2 is of a very general nature. No assumptions are made about the starting population $\mathcal{P}^0$, and the method of searching for $\mathcal{P}^t$. For example, we can set the function $z(\mathcal{P}) = H(\mathcal{P})$ and then choose the following strategy for Line 5:

1. Determine a set $\mathcal{O}^t$ of offspring of size λ.
2. Select μ solutions from $\mathcal{P}^{t-1} \cup \mathcal{O}^t$ using an archiving algorithm A, i.e. $\mathcal{S} = A(\mathcal{P}^{t-1}, \mathcal{O}^t)$.
3. Execute the above two steps until $H(\mathcal{S}) > H(\mathcal{P}^{t-1})$ and then set $\mathcal{P}^t = \mathcal{S}$, or until no such $\mathcal{S}$ can be found.

Algorithm 2 Interchange heuristic: μ is the size of the final set; λ the maximum number of elements which can be exchanged.

```
function HEURISTIC($\mu$, $\lambda$)
    $\mathcal{P}^0 \leftarrow$ initialize with an arbitrary set of size $\mu$
    $t \leftarrow 1$
    while true do
        determine a set $\mathcal{P}^t$ of size $\mu$ with $|\mathcal{P}^t - \mathcal{P}^{t-1}| \leq \lambda$ such that $z(\mathcal{P}^t) > z(\mathcal{P}^{t-1})$
        if no such a $\mathcal{P}^t$ exists then
            break
        end if
        $t \leftarrow t + 1$
    end while
    return $\mathcal{P}^G \leftarrow \mathcal{P}^{t-1}$
end function
```

Following Algorithm 2, the above steps need to guarantee that a set $\mathcal{P}^t$ with $H(\mathcal{P}^t) > H(\mathcal{P}^{t-1})$ is found if it exists. For example, we can use an exhaustive offspring generation, i.e. every subset of size λ of the decision space $\mathcal{X}$ can be determined with a probability larger than zero. Moreover, the archiving algorithm A must be able to determine an improved subset of $\mathcal{P}^{t-1} \cup \mathcal{O}^t$ if it exists. In other words, we require from A that $H(A(\mathcal{P}, \mathcal{O})) > H(\mathcal{P})$ if there exists a subset of $\mathcal{P} \cup \mathcal{O}$ of size μ with a larger hypervolume than $H(\mathcal{P})$. For example, A may in turn remove all possible subsets of size λ from $\mathcal{P}^{t-1} \cup \mathcal{O}^t$ and return a set that has a better hypervolume than $\mathcal{P}^{t-1}$. Note that this instance of the interchange heuristic can be easily rephrased in the general $(\mu+\lambda)$-EA framework of Algorithm 1 with an unbounded number of generations.

Nemhauser et al. [3] have proven the following result for the interchange heuristic:

Theorem 2. *Suppose z is non-decreasing and submodular. Moreover, define the optimization problem $z^* = \max_{\mathcal{P} \subseteq \mathcal{X}, |\mathcal{P}| \leq \mu} z(\mathcal{P})$. If $\mu = q \cdot \lambda - p$ with q a positive integer, and p integer with $0 \leq p \leq \lambda - 1$, then*

$$\frac{z^* - z(\mathcal{P}^G)}{z^* - z(\emptyset)} \leq \frac{\mu - \lambda + p}{2\mu - \lambda + p}$$

where $z(\mathcal{P}^G)$ is the value of the set obtained by Algorithm 2 and $z(\emptyset)$ is the value of the empty set.

We have shown that the hypervolume indicator is non-decreasing submodular. Therefore, if we set the function $z(\mathcal{P}) = H(\mathcal{P})$ and note that $H(\emptyset) = 0$, we can easily obtain the following bound on the approximation quality of Algorithm 2:

Proposition 1. *If $\mu = q \cdot \lambda - p$ with an integer $0 \leq p \leq \lambda - 1$, then*

$$H(\mathcal{P}^G) \geq \frac{1}{2 - \frac{\lambda - p}{\mu}} \cdot H_\mu^{\max}(\mathcal{X}) \tag{4}$$

This bound can be compared to the definition of the approximate effectiveness, see Definition 5, i.e. it bounds the achievable optimization quality in terms of the hypervolume if a certain algorithm structure is used. But whereas Definition 5 and the corresponding value of $\alpha = 2+\epsilon$ from [1] is related to Algorithm 1, the above bound with $\alpha = 2 - \frac{\lambda-p}{\mu}$ is related to Algorithm 2.

We will now show that the improved approximation bound of $\alpha = 2 - \frac{\lambda-p}{\mu}$ is valid also in the case of Algorithm 1, thereby improving the results in [1].

Theorem 3. *Suppose a non-decreasing $(\mu + \lambda)$-archiving algorithm which satisfies in addition*

$$\exists \mathcal{S} \,:\, (\mathcal{S} \subset \mathcal{P} \cup \mathcal{O}) \wedge (|\mathcal{S}| = \mu) \wedge (H(\mathcal{S}) > H(\mathcal{P})) \quad \Rightarrow \quad H(A(\mathcal{P},\mathcal{O})) > H(\mathcal{P})$$

Then for all finite sets $\mathcal{X}$, all objective functions f and all μ-populations $\mathcal{P}^0 \subseteq \mathcal{X}$ the following holds: For any run of an instance of Algorithm 2, one can determine a sequence of λ-populations $\mathcal{O}^1, ..., \mathcal{O}^N$ such that

$$H(A(\mathcal{P}^0, \mathcal{O}^1, ..., \mathcal{O}^N)) = H(\mathcal{P}^G)$$

Proof. The proof uses the special instance of Algorithm 2 that has been introduced above. Line 5 is implemented as follows: (1) Determine a set $\mathcal{O}^t$ of offspring of size λ using an exhaustive generation, i.e. each subset of $\mathcal{X}$ is determined with non-zero probability. (2) Use the archiving algorithm A to determine a set $\mathcal{S} = A(\mathcal{P}^{t-1}, \mathcal{O}^t)$. (3) Repeat these two steps until $H(\mathcal{S}) > H(\mathcal{P}^{t-1})$ or no such $\mathcal{S}$ can be found. Due to the required property of A, no such $\mathcal{S}$ can be found if it does not exist.

Algorithm 2 yields as final population $\mathcal{P}^G = \mathcal{P}^{t-1}$ which can be rewritten as $\mathcal{P}^{t-1} = A(\mathcal{P}^0, \mathcal{O}^1, ..., \mathcal{O}^{t-1})$ The sets of offspring $\mathcal{O}^i$ are generated as described above. Using $N = t-1$ yields the required result $H(A(\mathcal{P}^0, \mathcal{O}^1, ..., \mathcal{O}^N)) = H(\mathcal{P}^G)$. □

As a direct consequence of the execution equivalence between Algorithm 1 and Algorithm 2 according to the above theorem, the Definition 5 and (4), we can state the following result:

Proposition 2. *A non-decreasing $(\mu + \lambda)$-archiving algorithm $A(\mathcal{P}, \mathcal{O})$, which yields a subset of $\mathcal{P} \cup \mathcal{O}$ of size μ with a better hypervolume than that of $\mathcal{P}$ if there exists one, is $(2 - \frac{\lambda-p}{\mu})$-approximate where $\mu = q \cdot \lambda - p$ with an integer $0 \leq p \leq \lambda - 1$.*

It is interesting to note two special cases of the above proposition:

1. $\mu = \lambda$: In this case, we have a $(\mu + \mu)$-EA. It holds that $p = 0$ and therefore, the formula evaluates to $\alpha = 1$, which means that this algorithm actually is effective. This corresponds to the obvious result mentioned in the introduction.
2. $\lambda = 1$: In this case, we have a $(\mu + 1)$-EA. It holds that $p = 0$ and $q = \mu$ and therefore, the formula evaluates to $\alpha = 2 - \frac{1}{\mu}$, which is tighter than the bound of Bringmann and Friedrich [1].

Figure 1 shows the relation between λ and α for several settings of μ. As can be seen, it is a zigzag line which corresponds to the modulo-like definition of p and q. The local maxima of each line are located where μ is an integer multiple of λ.

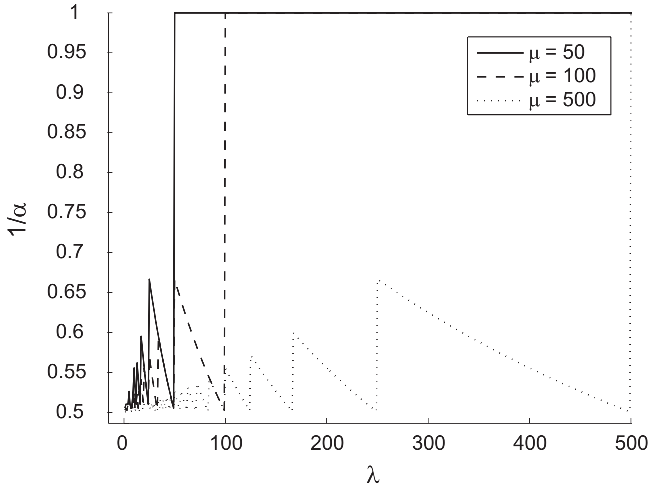

Fig. 1. Quality guarantees for the hypervolume achieved by a $(\mu + \lambda)$-EA. For a given μ and a given λ, there is a sequence of offspring such that at least $\frac{1}{\alpha} \cdot H_{\mu}^{max}(\mathcal{X})$ can be achieved, irrespective of the optimization problem and the chosen initial population.

4 Lower Bound on the Approximate Effectiveness

In the previous section we gave an upper bound on α. In this section, on the other hand, we will give a lower bound on α. This lower bound is tight for $\mu = 2$, i.e. is equal to the upper bound. To find this bound, we will show that there exist optimization problems and initial populations, such that any non-decreasing archiving algorithm will end up with a hypervolume that is at most $1/(1 + \frac{1}{2\lambda})$ of the optimal hypervolume. Whereas a first particular example has been shown in [4], a more general lower bound was shown in [1], where Bringmann and Friedrich found a problem where any non-decreasing archiving algorithm ends up with a hypervolume that is at most $1/(1 + 0.1338(1/\lambda - 1/\mu) - \epsilon)$ of the optimal hypervolume, for any $\epsilon > 0$. The new bound substantially tightens the result of [1], but relies on the general definition of the hypervolume indicator which uses a reference set $\mathcal{R}$ instead of a single reference point.

Theorem 4. *Let $\lambda < \mu$. There is no α-approximate non-decreasing $(\mu + \lambda)$-archiving algorithm for any $\alpha < 1 + \frac{1}{2\lambda}$.*

Proof. We proof this theorem by finding a population $\mathcal{P}^0 = \{s_0, ..., s_{\mu-1}\}$ whose hypervolume indicator $H(\mathcal{P}^0)$ can not be improved by any non-decreasing $(\mu + \lambda)$-archiving algorithm, i.e. it is locally optimal. At the same time, the optimal population $\mathcal{P}^* = \{o_0, ..., o_{\mu-1}\}$ has a hypervolume indicator value of $H(\mathcal{P}^*)$ which satisfies $H(\mathcal{P}^*) = (1 + \frac{1}{2\lambda} - \delta)H(\mathcal{P}^0)$ for any $\delta > 0$.

The setting we are considering for the proof is shown in Figure 2. There are $2 \cdot \mu$ points, where the initial population is set to $\mathcal{P}^0 = \{s_0, ..., s_{\mu-1}\}$ and the optimal population would be $\mathcal{P}^* = \{o_0, ..., o_{\mu-1}\}$. We consider a setting with multiple reference points $\{r_0, ..., r_{2\mu-2}\}$, such that the areas contributing to the hypervolume calculation are A_i (areas only dominated by the initial population), B_i (areas only dominated by the optimal population), and C_i and D_i (areas dominated by one solution of the initial population and one solution of the optimal population), see Figure 2. The objective space is the union of all points, i.e. $\mathcal{Y} = \mathcal{P}^0 \cup \mathcal{P}^*$.

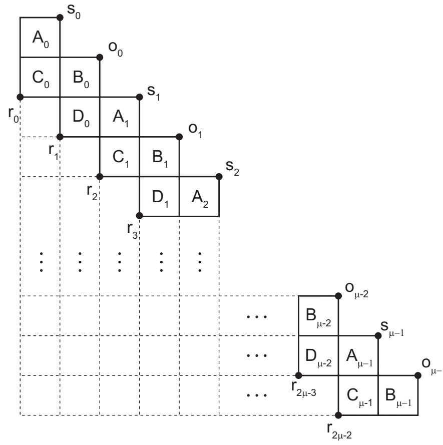

Fig. 2. Schematic drawing of the example setting in the proof of Theorem 4.

In our example, we set these areas as follows, assuming $\lambda < \mu$:

$$A_i = \epsilon \text{ for } 0 \leq i < \mu \quad , \quad B_i = \begin{cases} \epsilon & \text{for } 0 \leq i < \lambda \\ 1 & \text{for } \lambda \leq i < \mu \end{cases}$$

$$C_i = \sum_{i-\lambda \leq j < i} B_{j \bmod \mu} \quad , \quad D_i = \sum_{i+1 \leq j < i+1+\lambda} B_{j \bmod \mu}$$

Note that for any choice of areas A_i, B_i, C_i, and D_i, corresponding coordinates can be found for all s_i and o_i and r_i by using the following recursions:

$$s_i^x = o_{i-1}^x + \frac{A_i}{s_i^y - o_i^y}, \quad o_i^x = s_i^x + \frac{B_i}{o_i^y - s_{i+1}^y}$$

$$s_i^y = o_{i-1}^y - \frac{C_{i-1}}{s_{i-1}^x - o_{i-2}^x}, \quad o_i^y = s_i^y - \frac{D_{i-1}}{o_{i-1}^x - s_{i-1}^x}$$

$$r_i^x = \begin{cases} o_{i/2-1}^x & i \text{ even} \\ s_{(i-1)/2-1}^x & i \text{ odd} \end{cases}, \quad r_i^y = \begin{cases} o_{i/2+1}^y & i \text{ even} \\ s_{(i-1)/2}^y & i \text{ odd} \end{cases}$$

where s_i^x, o_i^x, s_i^y, o_i^y and r_i^y, r_i^y are the x-axis and y-axis coordinates of s_i, o_i and r_i, respectively. While s_0^x, s_0^y, and r_0^x with $r_0^x < s_0^x$ can be chosen arbitrarily, the coordinates for o_0^y and s_1^y are set as follows:

$$o_0^y = s_0^y - \frac{A_0}{s_0^x - r_0^x}, \quad s_1^y = o_0^y - \frac{C_0}{s_0^x - r_0^x}$$

Furthermore, $r_{2\mu-2}^y$ and $o_{\mu-1}^x$ are set as follows:

$$r_{2\mu-2}^y = o_{\mu-1}^y - \frac{C_{\mu-1}}{s_{\mu-1}^x - o_{\mu-2}^x}, \quad o_{\mu-1}^x = s_{\mu-1}^x + \frac{B_{\mu-1}}{o_{\mu-1}^y - r_{2\mu-2}^y}$$

First, we want to show that for the example, $\mathcal{P}^0$ is a local optimum, i.e. $H(\mathcal{P}^0)$ can not be improved by any non-decreasing $(\mu+\lambda)$-archiving algorithm. To do so consider a λ-population $\mathcal{O} \subset \mathcal{Y}$ and a μ-population $\mathcal{P}^1 \subset \mathcal{P}^0 \cup \mathcal{O}$. In order for $\mathcal{P}^0$ to be a local optimum, we have to show that $H(\mathcal{P}^0) \geq H(\mathcal{P}^1)$.

Note that for the rest of the proof, we will always use the indices modulo μ without writing it explicitly. Put differently, we will write A_i,B_i,C_i, and D_i as a short form of $A_{i \bmod \mu}$,$B_{i \bmod \mu}$,$C_{i \bmod \mu}$, and $D_{i \bmod \mu}$.

The hypervolume of the initial population can be written as $H(\mathcal{P}^0) = H - \sum_{0 \leq i < \mu} B_i = H - (\mu - \lambda) - \lambda\epsilon$, where H is the hypervolume of all solutions, i.e. $H = H(\mathcal{P}^0 \cup \mathcal{P}^*)$. Similarly, we can write $H(\mathcal{P}^1) = H - \sum_{i:s_i,o_i \notin P^1} C_i - \sum_{i:s_{i+1},o_i \notin P^1} D_i - \sum_{i:o_i \notin \mathcal{P}^1} B_i - \sum_{i:s_i \notin \mathcal{P}^1} A_i$. Using these expressions, we get the following set of equivalent inequalities:

$$\begin{aligned} H(\mathcal{P}^0) &\geq H(\mathcal{P}^1) \\ H - (\mu-\lambda) - \lambda\epsilon &\geq H - \textstyle\sum_{i:s_i,o_i \notin P^1} C_i - \sum_{i:s_{i+1},o_i \notin P^1} D_i \\ &\quad - \textstyle\sum_{i:o_i \notin \mathcal{P}^1} B_i - \sum_{i:s_i \notin \mathcal{P}^1} A_i \\ (\mu-\lambda) + \lambda\epsilon &\leq \textstyle\sum_{i:s_i,o_i \notin P^1} C_i + \sum_{i:s_{i+1},o_i \notin P^1} D_i \\ &\quad + ((\mu-\lambda) + \lambda\epsilon - \textstyle\sum_{i:o_i \in \mathcal{P}^1} B_i) + \sum_{i:s_i \notin \mathcal{P}^1} A_i \end{aligned}$$

$$\sum_{i:o_i \in \mathcal{P}^1} B_i \leq \sum_{i:s_i,o_i \notin P^1} C_i + \sum_{i:s_{i+1},o_i \notin P^1} D_i + \sum_{i:s_i \notin \mathcal{P}^1} A_i \tag{5}$$

To prove this inequality (5), we need to consider all possible μ-populations $\mathcal{P}^1 \subset \mathcal{P}^0 \cup \mathcal{O}$, i.e. the results of all possible $(\mu+\lambda)$-archiving algorithms. To go from $\mathcal{P}^0$ to $\mathcal{P}^1$, λ solutions s_i of the initial set are discarded and the same number of solutions o_i from the optimal set are added. We call these discarded s_i and added o_i *affected* solutions.

In the following, we consider *blocks* of affected solutions. To this end, we first mark all solutions in $\mathcal{P}^0 \cup \mathcal{P}^*$ that are either removed from $\mathcal{P}^0$ or added to $\mathcal{P}^0$ when going from $\mathcal{P}^0$ to $\mathcal{P}^1$. This set of marked solutions is then partitioned into the minimal number of subsets, such that each subset contains all solutions in index range $[i, i+k]$. Depending on whether the first and last solutions in such a subset are from set $\mathcal{P}^0$ or $\mathcal{P}^*$ we call it an (s,s)-, (s,o)-, (o,s)- or (o,o)-block, respectively. For example, an (o,s)-block with index range $[2,5]$ contains solutions $\{o_2, s_3, o_3, s_4, o_4, s_5\}$. The rationale is that non-neighboring solutions do not influence each other, as they do not dominate any common area. As for the blocks, there are two cases which will be considered separately.

Blocks of even length: There are two types of blocks of even length: Those starting with an added solution from the optimal set, i.e. (o,s)-blocks, and those starting with a discarded solution from the initial set, i.e. (s,o)-blocks. The first case can be formalized as follows: The (o,s)-block with index range $[i, i+k]$ exists iff $(o_l \in \mathcal{P}^1, i \le l < i+k) \wedge (o_{i+k} \notin \mathcal{P}^1) \wedge (s_i \in \mathcal{P}^1) \wedge (s_l \notin \mathcal{P}^1, i+1 \le l < i+k+1)$. For this block, (5) evaluates to:

$$\begin{aligned}
\textstyle\sum_{i:o_i\in\mathcal{P}^1} B_i &\le \textstyle\sum_{i:s_i,o_i\notin P^1} C_i + \sum_{i:s_{i+1},o_i\notin P^1} D_i + \sum_{i:s_i\notin\mathcal{P}^1} A_i \\
\textstyle\sum_{i\le l<i+k} B_l &\le C_{i+k} + 0 + \textstyle\sum_{i+1\le l<i+k+1} A_l \\
\textstyle\sum_{i\le l<i+k} B_l &\le \textstyle\sum_{i+k-\lambda\le l<i+k} B_l + k\epsilon \\
0 &\le \textstyle\sum_{i+k-\lambda\le l<i} B_l + k\epsilon
\end{aligned}$$

The last step is true because we know that $k \le \lambda$. As all B_l as well as ϵ are larger than zero, (5) holds.

The second case can be formalized as follows: The (s,o)-block with index range $[i, i+k]$ exists iff $(o_{i-1} \notin \mathcal{P}^1) \wedge (o_l \in \mathcal{P}^1, i \le l < i+k) \wedge (s_l \notin \mathcal{P}^1, i \le l < i+k) \wedge (s_{i+k} \in \mathcal{P}^1)$. For this block, (5) evaluates to:

$$\begin{aligned}
\textstyle\sum_{i:o_i\in\mathcal{P}^1} B_i &\le \textstyle\sum_{i:s_i,o_i\notin P^1} C_i + \sum_{i:s_{i+1},o_i\notin P^1} D_i + \sum_{i:s_i\notin\mathcal{P}^1} A_i \\
\textstyle\sum_{i\le l<i+k} B_l &\le 0 + D_{i-1} + \textstyle\sum_{i\le l<i+k} A_l \\
\textstyle\sum_{i\le l<i+k} B_l &\le \textstyle\sum_{i\le l<i+\lambda} B_l + k\epsilon \\
0 &\le \textstyle\sum_{i+k\le l<i+\lambda} B_l + k\epsilon
\end{aligned}$$

Again, we can see that the last inequality holds, and therefore, (5) holds.

Blocks of odd length: Such blocks consist of either a set of discarded solutions that enclose a set of added solutions or vice versa, i.e. (s,s)- or (o,o)-blocks. Due to $|\mathcal{P}^0| = |\mathcal{P}^1|$, the number of added solutions from the optimal set must be equal to the number of discarded solutions from the initial set. Directly following this, we know that for each block of discarded solutions enclosing added solutions, there must be another block of added solutions enclosing discarded solutions and vice versa. These two types of blocks can be formalized as follows: The (s,s)-block with index range $[i, i+k]$ exists iff $(o_l \in \mathcal{P}^1, i \le l < i+k-1) \wedge (o_{i-1}, o_{i+k-1} \notin \mathcal{P}^1) \wedge (s_l \notin \mathcal{P}^1, i \le l < i+k)$. The (o,o)-block with index range $[j, j+p]$ exists iff $(o_l \in \mathcal{P}^1, j \le l < j+p) \wedge (s_l \notin \mathcal{P}^1, j+1 \le l < j+p) \wedge (s_j, s_{j+p} \in \mathcal{P}^1)$. Also, we know that $1 \le k, p \le \lambda$ and $k + p \le \lambda + 1$. Considering both of these blocks, (5) evaluates to:

$$\begin{aligned}
\textstyle\sum_{i:o_i\in\mathcal{P}^1} B_i &\le \textstyle\sum_{i:s_i,o_i\notin P^1} C_i + \sum_{i:s_{i+1},o_i\notin P^1} D_i \\
&\quad + \textstyle\sum_{i:s_i\notin\mathcal{P}^1} A_i \\
\textstyle\sum_{i\le l<i+k-1} B_l + \sum_{j\le l<j+p} B_l &\le C_{i+k-1} + D_{i-1} + \textstyle\sum_{i\le l<i+k} A_l \\
&\quad + \textstyle\sum_{j+1\le l<j+p} A_l \\
\textstyle\sum_{i\le l<i+k-1} B_l + \sum_{j\le l<j+p} B_l &\le \textstyle\sum_{i+k-1-\lambda\le l<i+k-1} B_l + \sum_{i\le l<i+\lambda} B_l \\
&\quad +(k+p-1)\epsilon \\
\textstyle\sum_{j\le l<j+p} B_l &\le p \le \textstyle\sum_{i+k-1-\lambda\le l<i+\lambda} B_l + (k+p-1)\epsilon \\
p &\le \lambda\epsilon + \lambda - k + 1 + (k+p-1)\epsilon
\end{aligned}$$

The second last step can be done because we know that at most λ of the B_l's are set to ϵ and therefore, at least $\lambda - k + 1 \ge p$ of the B_l's remain which are set to 1. Also, because of $p \le \lambda - k + 1$, the last inequality holds and with it (5) holds.

Combinations of blocks: As stated before, only neighboring solutions in $\mathcal{Y} = \mathcal{P}^0 \cup \mathcal{P}^*$ share a common dominated area. From the definition of the different types of blocks it can be seen that there are no adjacent blocks, because in this case, the two blocks would be combined into one. Therefore, each pair of blocks is separated by at least one solution from $\mathcal{Y}$ which is not affected by the transition from $\mathcal{P}^0$ to $\mathcal{P}^1$. As a result, the changes in hypervolume when going from $\mathcal{P}^0$ to $\mathcal{P}^1$ can be considered separately for each block. We have shown that for any block, (5) holds. From this we can conclude that $H(\mathcal{P}^0) \ge H(\mathcal{P}^1)$ and therefore, $\mathcal{P}^0$ is a local optimum.

Now that we've done the first part of the proof, i.e. showing that any non-decreasing $(\mu + \lambda)$-archiving algorithm will not be able to escape from $\mathcal{P}^0$, we would like to calculate how far the hypervolume of $\mathcal{P}^0$ is from the maximum achievable hypervolume. In other words, we would like to calculate $\frac{H(\mathcal{P}^*)}{H(\mathcal{P}^0)}$. The hypervolume of the initial population evaluates to:

$$\begin{aligned}
H(\mathcal{P}^0) &= \textstyle\sum_{0\le l<\mu} C_l + \sum_{0\le l<\mu} D_l + \sum_{0\le l<\mu} A_l \\
&= \textstyle\sum_{0\le l<\mu} \left(\sum_{l-\lambda\le j<l} B_j + \sum_{l+1\le j<l+1+\lambda} B_j \right) + \mu\epsilon \\
&= \textstyle\sum_{0\le l<\mu} \left(\sum_{l-\lambda\le j<l+1+\lambda} B_j - B_l \right) + \mu\epsilon \\
&= (2\lambda+1) \textstyle\sum_{0\le l<\mu} B_l - \sum_{0\le l<\mu} B_l + \mu\epsilon \\
&= 2\lambda \textstyle\sum_{0\le l<\mu} B_l + \mu\epsilon
\end{aligned}$$

The hypervolume of the optimal population, on the other hand, can be calculated as follows:

$$\begin{aligned}
H(\mathcal{P}^*) &= \textstyle\sum_{0\le l<\mu} C_l + \sum_{0\le l<\mu} D_l + \sum_{0\le l<\mu} B_l \\
&= \textstyle\sum_{0\le l<\mu} \sum_{l-\lambda\le j<l+1+\lambda} B_j \\
&= (2\lambda+1) \textstyle\sum_{0\le l<\mu} B_l
\end{aligned}$$

Both sets of equations make use of $\sum_{0\le l<\mu} \sum_{l-\lambda\le j<l+1+\lambda} B_j = (2\lambda + 1) \sum_{0\le l<\mu} B_l$. This is due to the fact that the inner sum of the left-hand term consists of $2\lambda + 1$ summands. Because all indices are taken modulo μ, we see that each B_j is summed up $2\lambda + 1$ times in the whole term.

Finally, this leads us to the following result, which holds for any $\delta > 0$ if $\epsilon \to 0$ and $\lambda < \mu$:

$$\begin{aligned} \tfrac{H(\mathcal{P}^*)}{H(\mathcal{P}^0)} &= \tfrac{(2\lambda+1)\sum_{0\leq l<\mu} B_l}{2\lambda \sum_{0\leq l<\mu} B_l + \mu\epsilon} \\ &= 1 + \tfrac{1}{2\lambda} - \delta \end{aligned}$$

Note that in the case of $\lambda = \mu$, the equation evaluates to $\frac{H(\mathcal{P}^*)}{H(\mathcal{P}^0)} = 1$, which is very natural, since for $\mu = \lambda$, any non-decreasing $(\mu + \lambda)$-archiving algorithm is effective. □

We may also interpret the above result in terms of the more practical interchange heuristic shown in Algorithm 2. One can conclude that for $z(\mathcal{P}) = H(\mathcal{P})$, i.e. we use the hypervolume indicator for archiving, we may end up with a solution that is not better than $1/\alpha$ times the optimal hypervolume with $\alpha > 1 + \frac{1}{2\lambda}$, even after an unlimited number of iterations.

5 Conclusion

In this paper, we investigated the α-approximate effectiveness of $(\mu + \lambda)$-EAs that optimize the hypervolume. The value of α gives a lower bound on the hypervolume which can always be achieved, independent of the objective space and the chosen initial population. While it is obvious that for $\mu = \lambda$, α is equal to 1, Bringmann and Friedrich have shown that for $\lambda = 1$, α is equal to 2. This paper strictly improves the currently known bound and finds that for arbitrary λ, the approximation factor α is equal to $2 - \frac{\lambda - p}{\mu}$, where $\mu = q \cdot \lambda - p$ and $0 \leq p \leq \lambda - 1$.

Furthermore, we improve the available lower bound on α for the general definition of the hypervolume indicator, i.e. $\alpha > 1 + \frac{1}{2\lambda}$. Upper and lower bounds only match for a population size of $\mu = 2$. It might be possible to further tighten the lower bound by extending the worst case construction in the proof of Theorem 4 to higher dimensions of the objective space.

References

1. K. Bringmann and T. Friedrich. Convergence of hypervolume-based archiving algorithms i: Effectiveness. In *Genetic and Evolutionary Computation Conference (GECCO)*, pages 745–752, 2011.
2. J. Edmonds. Submodular functions, matroids and certain polyhedra. In R. Guy, editor, *Combinatorial Structures and their Applications*, pages 69–87. Gordon and Breach, New York, 1971.
3. G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions – i. *Mathematical Programming*, 14:265–294, 1978.
4. E. Zitzler, L. Thiele, and J. Bader. On set-based multiobjective optimization. *IEEE Trans. on Evolutionary Computation*, 14:58–79, 2010.