H.J. Otten

Global optimization for Lipschitz continuous expensive black box functions

Master thesis

Supervisors:
dr. S.C. Hille
dr. M.T.M. Emmerich

Date master exam: 08 June 2018

Mathematisch Instituut, Universiteit Leiden

Abstract

When optimizing a multi-objective function, it is interesting to find the points where the best trade-off is reached between the different objectives, i.e. to find the points where no improvement in one of the objective functions can be made without degrading some of the other objective values. To find these optimal points, we suggest a new algorithm that uses the expected hypervolume improvement of a point. Subsequently this algorithm is considered for the single objective case in one dimension and compared to the existing algorithm for this single objective problem called Shubert's Algorithm. Moreover, some problems with the existing proof of the convergence rate of Shubert's Algorithm are resolved. Finally, the bi-objective case is investigated and a method is constructed to calculate the expected hypervolume improvement in a point in an explicit way.

Acknowledgements

I would first like to thank my Mathematics thesis supervisor dr. S.C. Hille. The door to his office was always open whenever I encountered a problem or had a question about my research or writing. Moreover, I would like to thank dr. M.T.M. Emmerich of the computer science department for introducing me to this very interesting research field and guiding me along the way.

I would also like to thank Wout Gevaert, one of my fellow Mathematics students, who helped by discussions when I was constructing one of the most difficult proofs of this thesis. Moreover, I would like to thank Rosa Schwarz, another fellow Mathematics student, who took the time to read my complete thesis and point out all the small (grammar) mistakes and who was always there for me when I needed some extra support. Finally, I would like to thank my family and other friends for their support.

Contents

1 Preliminaries **7**
- 1.1 Multi-objective optimization 7
- 1.2 Lipschitz continuous functions 9
- 1.3 Optimization for expensive black box functions 9

2 The expected hypervolume improvement approach **15**
- 2.1 The hypervolume improvement 15
- 2.2 The proposed EI-algorithm 16
- 2.3 Calculating the EI 17
- 2.4 An alternative: hypervolume-based DIRECT 19

3 A special case: single objective optimization in one dimension **21**
- 3.1 Shubert's Algorithm 21
  - 3.1.1 Convergence 23
  - 3.1.2 Error bounds 25
- 3.2 The expected hypervolume improvement 34
- 3.3 EI-method in comparison with Shubert's Algorithm 39
- 3.4 A comment on single objective optimization in two dimensions . 39

4 The expected hypervolume improvement for bi-objective optimization **41**
- 4.1 Introduction: the expected hypervolume improvement for the Gaussian approach 41
- 4.2 General expression for the expected hypervolume improvement . 42
- 4.3 Computation by HVI-tent structures 44
- 4.4 Using the HVI-tent structures for the proposed algorithm 48

5 Discussion **51**

6 Conclusion **53**

A Some calculations for the hypervolume improvement **55**

B Lipschitz estimates for the hypervolume improvement **59**

B.1 Difference in dominated hypervolume between two sets with equal cardinality 59

B.2 Difference in dominated hypervolume between a set and the same set with one more element 62

Introduction

Multi-objective global optimization is a subject that is extensively investigated by computer scientist and a wide variety of algorithms have been introduced to optimize multi-objective functions. Here we will look at the subject from a more mathematical point of view. We will consider only Lipschitz continuous objective functions of which the Lipschitz constant is known. Moreover, we assume that these objective functions are computationally expensive to evaluate, which makes it very important to do as few function evaluations as possible in the process of optimizing the objectives. In this thesis, we will construct a new algorithm to find the global minimum of such objective functions, or the so-called Pareto front in multi-objective optimization.

At first we meant to focus on multi-objective optimization, but to understand better how optimization algorithms worked for Lipschitz continuous functions, we first took a closer look at single objective optimization in one dimension, i.e. we consider objective functions f of the form $f : \mathbb{R}^1 \to \mathbb{R}^1$. However, the single objective case appeared to be more interesting than we expected and turned out to be an important part of this work.

In Chapter 3 this single objective case is treated extensively. Firstly, a new proof is constructed for the fact that a constant function generates the worst convergence for Shubert's Algorithm. Furthermore, we found that our proposed algorithm using the expected hypervolume improvement for single objective optimization in one dimension generally chooses the same points to evaluate as Shubert's Algorithm but occasionally deviates from Shubert's Algorithm.

Chapter 1 introduces some important terminology and definitions and describes the field of multi-criteria optimization. Chapter 2 takes a closer look at the expected hypervolume improvement, as described in Emmerich et al. [3], which we want to use in our new algorithm in which as little statistical assumptions as possible are used. In Section 2.2, a general formulation of this new algorithm is given.

Chapter 4 treats multiple approaches to find a useful expression for the expected hypervolume in two dimensions which can be used in a new algorithm to find an optimal point in which to evaluate the objective function.

Finally, in Appendix B, various bounds are given for the hypervolume improvement. However, these bounds are not very strict which makes it harder to use them. Nevertheless, as a result of the research performed for this master project, we have included them in this thesis, as appendix.

Chapter 1

Preliminaries

1.1 Multi-objective optimization

In multi-objective optimization our aim is to minimize a function $f : A \subset \mathbb{R}^d \to \mathbb{R}^n$, which is called *the (vector-valued) objective function*. The set $A \subset \mathbb{R}^d$ is called the *decision space*; $\mathbb{R}^n$ is called the *objective space*. We say that an element $\mathbf{x} \in \mathbb{R}^n$ is *dominated* by an element $\mathbf{y} \in \mathbb{R}^n$, written as $\mathbf{y} \prec \mathbf{x}$, if $y_i \leqslant x_i$ for all $i \in \{1, \ldots, n\}$ and $y_i < x_i$ for at least one $i \in \{1, \ldots, n\}$. Note that in this terminology the dominating element is 'the smallest' of the two, intuitively speaking. Now we are interested in the solutions for which none of the objective function values can be improved without degrading some of the other objective values, i.e. the non-dominated solutions.

Definition 1.1.1. The set of *Pareto optimal solutions* of $f : A \subset \mathbb{R}^d \to R^n$ on a set A, denoted by $P(f, A)_O$, consists of all $(y_1, \ldots, y_n) \in f(A)$ with:

$$\begin{aligned} \text{For all } \ (z_1, \ldots, z_n) \in f(A) : \text{if there exists an } i \in \{1, \ldots, n\} \text{ such that } z_i < y_i \\ \text{then there exists a } j \in \{1, \ldots, n\} \text{ such that } z_j > y_j. \end{aligned}$$

This set is called the *Pareto front* and thus consists of all the elements with the best trade-off between the different variables.

Definition 1.1.2. The set of *Pareto optimal decisions* in A, denoted by $P(f, A)_D$, consists of all $\mathbf{a} \in A$ with:

$$\begin{aligned} \text{For all } \ \mathbf{b} \in A : \text{if there exists an } i \in \{1, \ldots, n\} \text{ such that } f^i(\mathbf{b}) < f^i(\mathbf{a}) \\ \text{then there exists a } j \in \{1, \ldots, n\} \text{ such that } f^j(\mathbf{b}) > f^j(\mathbf{a}). \end{aligned}$$

Hence, $f(P(f, A)_D) = P(f, A)_O$. So we are looking for the set of Pareto optimal decisions $P(f, A)_D$, which gives us the Pareto front $P(f, A)_O = f(P(f, A)_D)$.

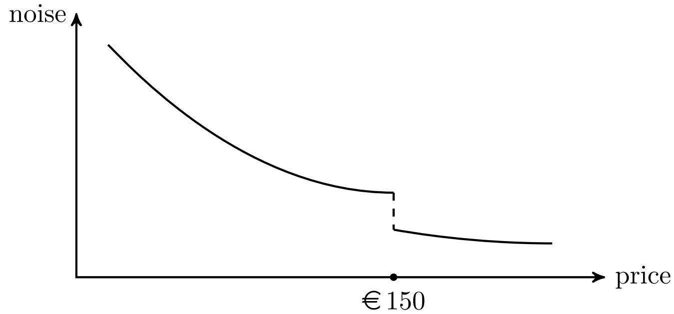

Figure 1.1: *Example of the Pareto front of the trade-off between price and noise of a vacuum cleaner.*

Definition 1.1.3. A set is called the *dominated set of* S, relative to a reference point $\mathbf{r} \in \mathbb{R}^n$ if it contains all elements that are dominated by a point in S and that dominate the reference point $\mathbf{r}$:

$$\mathrm{Dom}_{\mathbf{r}}(S) := \{\mathbf{y} \in \mathbb{R}^n \mid \exists\ \mathbf{z} \in S : \mathbf{z} \prec \mathbf{y} \text{ and } \mathbf{y} \prec \mathbf{r}\}$$

In multi-objective optimization the different variables can often not be optimized all at once. For example, imagine that you want to find a television set with the highest quality, best size and lowest price. Probably the television with the highest quality will not be the cheapest, so you search for the best trade-off between those variables. In some cases, the preferences of the decision maker can be used to decide that one of these options where the trade-off is optimal, a Pareto optimal solution, is better than another. However, here we attempt to make as few assumptions as possible about the function and the preferences of the decision maker while optimizing the function.

For example, suppose we want to buy a new vacuum cleaner that is not too expensive, say around € 140, and makes as little noise as possible. Because of a new technology which is a bit more expensive, the vacuum cleaners in the Pareto front that are € 150 or more are much more quiet than those which cost less than € 150, which is illustrated in Figure 1.1. One can now select the vacuum cleaner that is most quiet for the price one had in mind, namely € 140. But as can be seen in the figure, a slight increase in price beyond € 150 may result in a huge increase in quality. Because the budget is close to € 150, it is better to increase the budget a little.

Moreover, in most cases it is not desirable to save all the approximations of the Pareto optimal solutions that are found, because that will slow down the algorithm. In that case we can use the hypervolume indicator as described in Emmerich et al. [2]. It compares approximation sets to the Pareto front $P(f, A)_O$.

It is not always easy to compare the quality of different algorithms for multi-criteria optimization. In Zitzler et al. [9], a mathematical method is introduced to compare different algorithms by comparing the corresponding approximations of the Pareto front that every algorithm generates. They introduce a strict preference order on approximation sets. It appears hard to compare these approximations by a finite set of criteria such as diversity and distance between the points in the set. However, certain ways are proposed to use unary and binary quality indicators to compare the performance of different multi-criteria algorithms.

1.2 Lipschitz continuous functions

Definition 1.2.1. A function $f : \mathbb{R}^d \to \mathbb{R}^n$, with $f(\mathbf{x}) = (f^1(\mathbf{x}), \ldots, f^n(\mathbf{x}))$ for any $\mathbf{x} \in \mathbb{R}^d$ is called *Lipschitz continuous* or is said to satisfy the Lipschitz condition with constant $\mathbf{L} = (L^1, \ldots, L^n)$ if there exist $L^1, \ldots, L^n > 0$ such that for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have

$$|f^k(\mathbf{x}) - f^k(\mathbf{y})| \leqslant L^k \|\mathbf{x} - \mathbf{y}\|, \ k = 1, \ldots, n$$

where $\|\mathbf{x} - \mathbf{y}\| = \sum_{i=1}^{d} |x_i - y_i|$ is the l^1 or so-called Manhattan metric.

Note that f^k and L^k are not powers of f and $\mathbf{L}$. This notation with superscripts has been chosen to keep it in conformity with the literature (e.g. Žilinskas and Žilinskas [10]).

Definition 1.2.2. The *Lipschitz constant of* f, $|f|_{\text{Lip}}$, is the vector $\mathbf{L} = (L^1, \ldots, L^n)$ such that for all $k = 1, \ldots, n$ we have that L^k is the smallest value such that

$$|f^k(\mathbf{x}) - f^k(\mathbf{y})| \leqslant L^k \|\mathbf{x} - \mathbf{y}\| \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d;$$

in other words, we have

$$L^k = \sup_{\mathbf{x} \neq \mathbf{y}} \frac{|f^k(\mathbf{x}) - f^k(\mathbf{y})|}{d(\mathbf{x}, \mathbf{y})} \text{ for all } k \in \{1, \ldots, n\}.$$

In this thesis we work with a vector of Lipschitz constants $\mathbf{L} = (L^1, \ldots, L^n)$, for each coordinate function f^k, where $k = 1, \ldots, n$, since this will give more accuracy than working with a single Lipschitz constant L that is valid for all coordinate functions.

1.3 Optimization for expensive black box functions

In many practical situations it is expensive to evaluate the objective function at a point. For example, the computation of the objective function could be solving a system of differential equations or simulating an engineering design as described in Jones et al. [5]. In such cases, it is also often not possible to have full knowledge of the function's equations, but it is possible to evaluate the function at points. Moreover, limited knowledge might be available about the function's behavior, such as Lipschitz constants. In the literature, the resulting problem is termed black-box optimization (with expensive objective functions).

Thus, in the context of multi-objective optimization, the goal is to find an approximation of the Pareto front while using as few function evaluations as possible. That is the reason that we want to be able to predict in what region of the decision space the most progress can be made, i.e. the region where it is possible that a solution is found with a much better trade-off between the variables, or the other way around, to early-on prune regions in which it is certain that a global Pareto optimum cannot be found. In this section we will discuss the way this is done in Žilinskas and Žilinskas [10] for objective functions that satisfy a Lipschitz bound. This section will thus provide the notion of most of the arguments and notation that will be used subsequently. The reader gets acquainted with the various concepts by reading this section.

Let $A \subset \mathbb{R}^d$ be the domain of a function f and let $\mathbf{A}_{[\mathbb{R}]}$ be the set of hyper-rectangles which constitute a partition of A. Let $A_r \in \mathbf{A}_{[\mathbb{R}]}$ be a hyper-rectangle with $\mathbf{a}(r) = (a_1, \ldots, a_d)$ and $\mathbf{b}(r) = (b_1, \ldots, b_d)$ the different endpoints of a diagonal $\tilde{A}_r$. In Žilinskas and Žilinskas [10] it is proven that to find the Pareto front of the lower bound of f we only need the function values on the endpoints $\mathbf{a}(r)$ and $\mathbf{b}(r)$ of the diagonal with $a_i(r) < b_i(r)$ for all $i = 1, \ldots, d$ and only need to consider the points on the diagonal $\tilde{x} \in \tilde{A}_r$. We will prove this for any diagonal $\tilde{A}_r$, so we will not assume that $a_i(r) < b_i(r)$ for all $i = 1, \ldots, d$.

Define for $x \in A_r$ and $k = 1, 2, \ldots, n$

$$g^k(x, A_r) := \max\left(f^k(\mathbf{a}(r)) - L^k \sum_{i=1}^{d} |x_i - a_i(r)|, f^k(\mathbf{b}(r)) - L^k \sum_{i=1}^{d} |b_i(r) - x_i|\right)$$

Lemma 1.3.1. *$g^k(x, A_r)$ is a lower bound for $f^k(x)$ for all $x \in A_r$ and for all $k = 1, 2, \ldots, n$.*

Proof. Take an element $\mathbf{x} = (x_1, \ldots, x_d) \in A_r$. By the fact that f is Lipschitz continuous, we have that for all $k = 1, \ldots, n$ and for all $\mathbf{y} = (y_1, \ldots, y_d) \in \mathbb{R}^d$

$$\left|f^k(\mathbf{x}) - f^k(\mathbf{y})\right| \leqslant L^k \sum_{i=1}^{d} |x_i - y_i|.$$

This gives us the following inequalities:

$$f^k(\mathbf{x}) \leqslant f^k(\mathbf{a}(r)) + L^k \sum_{i=1}^{d} |x_i - a_i(r)| \quad \wedge \quad f^k(\mathbf{x}) \geqslant f^k(\mathbf{a}(r)) - L^k \sum_{i=1}^{d} |x_i - a_i(r)|,$$

$$f^k(\mathbf{x}) \leqslant f^k(\mathbf{b}(r)) + L^k \sum_{i=1}^{d} |b_i(r) - x_i| \quad \wedge \quad f^k(\mathbf{x}) \geqslant f^k(\mathbf{b}(r)) - L^k \sum_{i=1}^{d} |b_i(r) - x_i|.$$

Hence we find

$$f^k(\mathbf{x}) \geqslant \max\left(f^k(\mathbf{a}(r)) - L^k \sum_{i=1}^{d} |x_i - a_i(r)|, f^k(\mathbf{b}(r)) - L^k \sum_{i=1}^{d} |b_i(r) - x_i| \right),$$

and so $f^k(\mathbf{x}) \geqslant g^k(x, A_r)$ for all $\mathbf{x} \in A_r$. □

In Žilinskas and Žilinskas [10] some results, such as Lemma 1.3.2, are proven specially for the bi-objective case. That is why we consider, for now, the bi-objective problem

$$\min_{x \in A_r} g(x, A_r), \quad g(x, A_r) = \left(g^1(x, A_r), g^2(x, A_r)\right)^T$$

Lemma 1.3.2. *No element of the Pareto front $P(g, A_r)_O$ is dominated by a vector $f(\mathbf{x})$ with $\mathbf{x} \in A_r$.*

Proof. Assume there is a $\mathbf{x} \in A_r$ such that $f(\mathbf{x}) \prec \mathbf{y}$ with $\mathbf{y} \in P(g, A_r)_O$. Because $g^k(\mathbf{x})$ is a lower bound for $f^k(\mathbf{x})$ for all $k = 1, \ldots, n$, there exists a $\mathbf{v} \in P(g, A_r)_O$ such that $\mathbf{v} \preceq f(\mathbf{x})$. Now we have $\mathbf{v} \preceq f(\mathbf{x}) \prec \mathbf{y}$, but this is a contradiction with the fact that $\mathbf{y} \in P(g, A_r)_O$ and thus $\mathbf{y}$ is not dominated by any element in $P(g, A_r)_O$. So there exists no element $\mathbf{x} \in A_r$ such that $f(\mathbf{x})$ dominates an element in $P(g, A_r)_O$. □

The following lemma can be found in Žilinskas and Žilinskas [10], where this lemma is Lemma 2. This is proven for the diagonal $\tilde{A}_r$ with endpoints $\mathbf{a}(r)$ and $\mathbf{b}(r)$ for which holds $a_i(r) < b_i(r)$ for all $i = 1, \ldots, d$, but here we will prove it for any diagonal of A_r.

Lemma 1.3.3. *The Pareto front $P(g, A_r)_O$ of g on A_r coincides with the Pareto front $P(g, \tilde{A}_r)_O$ of g on $\tilde{A}_r$ where $\tilde{A}_r$ is one of the diagonals of A_r.*

Proof. Let $\mathbf{z}$ be such that $\mathbf{z} = g(\mathbf{x}, A_r)$ for some $\mathbf{x} \in A_r$. Then we have

$$z_k = \max\left(f^k(a(r)) - L^k \sum_{i=1}^{d} |x_i - a_i(r)|, f^k(b(r)) - L^k \sum_{i=1}^{d} |b_i(r) - x_i| \right)$$

for $k = 1, 2$. Let $\tau \in [0, 1]$ be defined by the equality

$$\tau = \frac{\sum_{j=1}^{d} |x_j - a_j(r)|}{\sum_{j=1}^{d} |b_j(r) - a_j(r)|}.$$

By the definition of a diagonal with endpoints $a(r)$ and $b(r)$, we have that

$$\tilde{x} := a(r) + \tau(b(r) - a(r)) \in \tilde{A}_r.$$

Now we find that

$$\begin{aligned}\sum_{i=1}^{d}|\tilde{x}_i - a_i(r)| &= \sum_{i=1}^{d}|a_i(r) + \tau(b_i(r) - a_i(r)) - a_i(r)| \\ &= |\tau|\sum_{i=1}^{d}|b_i(r) - a_i(r)| \\ &= \sum_{i=1}^{d}|x_j - a_j(r)|\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{d}|b_i(r) - \tilde{x}_i| &= \sum_{i=1}^{d}|b_i(r) - (a_i(r) + \tau(b_i(r) - a_i(r)))| \\ &= |1-\tau|\sum_{i=1}^{d}|b_i(r) - a_i(r)| \\ &= \frac{\left|\sum_{j=1}^{d}|b_j(r) - a_j(r)| - \sum_{j=1}^{d}|x_j - a_j(r)|\right|}{\sum_{j=1}^{d}|b_j(r) - a_j(r)|}\sum_{i=1}^{d}|b_i(r) - a_i(r)| \\ &= \sum_{j=1}^{d}\Big||b_j(r) - a_j(r)| - |x_j - a_j(r)|\Big|.\end{aligned}$$

Note that if $b_j(r) > a_j(r)$ it follows that $x_j > a_j(r)$ and if $b_j(r) < a_j(r)$ it follows that $x_j < a_j(r)$, because $\mathbf{x}$ lies in the hyperrectangle with the line between $a(r)$ and $b(r)$ as diagonal, so

$$\begin{aligned}&\sum_{j=1}^{d}\Big||b_j(r) - a_j(r)| - |x_j - a_j(r)|\Big| \\ &= \begin{cases}\sum_{j=1}^{d}|b_j(r) - a_j(r) - (x_j - a_j(r))| & \text{if } b_j(r) > a_j(r) \\ \sum_{j=1}^{d}|a_j(r) - b_j(r) - (a_j(r) - x_j)| & \text{if } b_j(r) < a_j(r)\end{cases} \\ &= \sum_{j=1}^{d}|b_j(r) - x_j|.\end{aligned}$$

Now we find that

$$\begin{aligned}z_k &= \max\left(f^k(a(r)) - L^k\sum_{i=1}^{d}|x_i - a_i(r)|, f^k(b(r)) - L^k\sum_{i=1}^{d}|b_i(r) - x_i|\right) \\ &= \max\left(f^k(a(r)) - L^k\sum_{i=1}^{d}|\tilde{x}_i - a_i(r)|, f^k(b(r)) - L^k\sum_{i=1}^{d}|b_i(r) - \tilde{x}_i|\right).\end{aligned}$$

So for every value $z = g(x, A_r)$ for $x \in A_r$ there exists a $\tilde{x} \in \tilde{A}_r$ with $g(\tilde{x}, A_r) = z$.

□

So we have found that to find the Pareto front of g on the hyperrectangle A_r, we only need the function values on the endpoints of the diagonal $\mathbf{a}(r)$ and $\mathbf{b}(r)$ and we only have to consider the points $\tilde{x} \in \tilde{A}_r$ on any diagonal of A_r.

Chapter 2

The expected hypervolume improvement approach

Recall that the goal of this thesis is to find a new algorithm for optimizing a multi-objective Lipschitz continuous function on $A \subset \mathbb{R}^d$ using the expected hypervolume improvement. The expected hypervolume improvement method is designed to make maximum use of the a priori assumptions (Lipschitz constants in our setting) of the objective functions, such that the least number of (expensive) evaluations of the objective function need to be made. The precise form of this proposed algorithm can be found in section 2.2, but first we will introduce some new concepts including the precise definition of the expected hypervolume improvement.

2.1 The hypervolume improvement

As mentioned in Emmerich et al. [3], the expected hypervolume improvement is a useful tool for global optimization. To understand how this works, we first need to introduce the hypervolume indicator. Assume we have evaluated the Lipschitz continuous objective function f, which we want to minimize, at k points $\mathbf{x}^{(1)}, \ldots \mathbf{x}^{(k)} \in \mathbb{R}^d$ and we denote the corresponding function values by $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(k)} \in \mathbb{R}^n$. Let S_k be the set of evaluated points and Y_k the set of corresponding function values.

Definition 2.1.1. The *hypervolume indicator* is the n-dimensional Lebesgue measure λ_n of the dominated subspace limited from above by a reference point $\mathbf{r} \in \mathbb{R}^n$. In formula this gives

$$hv(Y_k) = \lambda_n \left(\{ \mathbf{u} \in \mathbb{R}^n \mid \exists\, \mathbf{y} \in Y_k : \mathbf{y} \prec \mathbf{u} \text{ and } \mathbf{u} \prec \mathbf{r} \} \right) = \lambda_n \left(\bigcup_{\mathbf{y} \in Y_k} [\mathbf{y}, \mathbf{r}] \right).$$

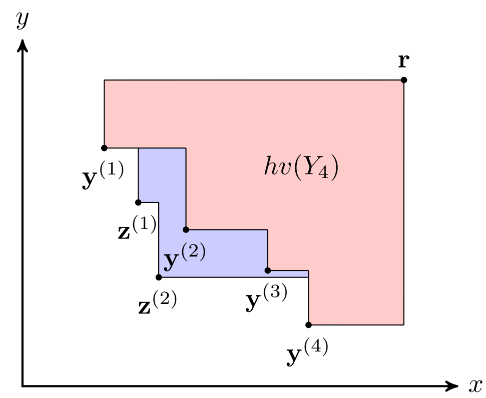

Figure 2.1: *The pink area is the hypervolume indicator of a set* $Y_4 = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(4)}\} \subset \mathbb{R}^2$ *with respect to reference point* $\mathbf{r} \in \mathbb{R}^2$*. The blue area is the hypervolume improvement of* $z = \{z^{(1)}, z^{(2)}\}$ *relative to* Y_k*.*

Here

$$[\mathbf{y}, \mathbf{r}] := \{(u_1, \ldots, u_n) \in \mathbb{R}^n \mid y_i \leqslant u_i \leqslant r_i \text{ for all } i \in \{1, \ldots, n\}\}.$$

So $hv(Y_k)$ is the amount of hypervolume dominated by a set Y_k. The hypervolume improvement of a point $\mathbf{y} \in \mathbb{R}^n$ indicates the amount of hypervolume that is dominated by $\mathbf{y}$ that was not yet dominated by the set $Y_k = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(k)}\}$. Analogously, the hypervolume improvement of a set Z given Y_k is the amount of hypervolume that is dominated by one of the points in Z that was not yet dominated by Y_k. That is, the hypervolume improvement $HVI(Z \mid Y_k)$ of a set Z given Y_k is

$$\begin{aligned} HVI(Z \mid Y_k) &= \lambda_n\Big(\{\mathbf{y}' \in \mathbb{R}^n \mid \exists\, \mathbf{y} \in Z \text{ such that } \mathbf{y} \prec \mathbf{y}' \\ &\qquad\qquad \text{and not } \mathbf{y}^{(j)} \prec \mathbf{y}' \text{ for any } j = 1, \ldots, k\}\Big). \\ &= hv(Z \cup Y_k) - hv(Y_k) \end{aligned}$$

2.2 The proposed EI-algorithm

The expected hypervolume improvement (EI) of a point $\mathbf{x} \in \mathbb{R}^d$ is the extra amount of dominated hypervolume we 'expect' to get given a predictive distribution on $f(x)$ if we evaluate the function f in $\mathbf{x}$. Using the Lipschitz constant and the previously evaluated function values $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(k)}$, we know that $f(\mathbf{x})$ must lie in a certain area $E_{\mathbf{x}} \subset \mathbb{R}^n$. Based on the assumptions, there is no further information where $f(\mathbf{x})$ will be within $E_{\mathbf{x}}$, so we assume the possible value Y to be a random variable, homogeneously distributed over $E_{\mathbf{x}}$. The expected hypervolume improvement is then the expected value of the hypervolume improvement of Y relative to $\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(k)}\}$.

A formal definition of the expected hypervolume improvement will follow in Section 2.3. The idea is to use this expected hypervolume improvement to minimize the amount of function evaluations that is needed to get the improvement threshold below a certain value. With $EI(\mathbf{x}_t \mid \mathscr{D})$ we denote the expected hypervolume improvement when we evaluate the objective function f in $\mathbf{x}_t$ given a set $\mathscr{D} \subset \mathbb{R}^d$ of already evaluated points. A sketch of the proposed algorithm using the EI is given in Algorithm 1.

Algorithm 1 Lipschitz-hv

Require:

vectorial Lipschitz function f to be minimized that satisfies the Lipschitz condition with constant $\mathbf{L}$.
search space X with the function values at the edges $\mathbf{x_i}$
evaluation budget v
improvement threshold σ

for $t = 1, 2, \ldots, v$ **do**
  $\mathscr{P} \leftarrow ND(\{f(\mathbf{x}_i) : i \in H_t, \sigma_i \geqslant \sigma_t\})$ and $\mathscr{D} \leftarrow \{\mathbf{x}_i : f(\mathbf{x}_i) \in \mathscr{P}\}$
  Find an $\mathbf{x}_t$ for which $EI(\mathbf{x}_t \mid \mathscr{D})$ is maximized
  if $EI(\mathbf{x}_t \mid \mathscr{D}) < \sigma$ **then**
    Stop
  end if
  Evaluate f at $\mathbf{x}_t$.
  $t = t + 1$
end for
return $ND\left(\{f(\mathbf{x}_t)\}_{t=1,2\ldots}\right)$

2.3 Calculating the EI

To make the proposed algorithm faster than previous algorithms, it is important to be able to easily determine the $\mathbf{x} \in \mathbb{R}^d$ for which the EI is maximized. In this section we derive an exact expression for computing the EI.

Since f is Lipschitz continuous, we know that if we evaluate f in $\mathbf{x} \in \mathbb{R}^d$, the corresponding $\mathbf{y}$-value $\mathbf{y} = f(\mathbf{x}) \in \mathbb{R}^n$ satisfies for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, k\}$:

$$f^i\left(\mathbf{x}^{(j)}\right) - L^i\|\mathbf{x} - \mathbf{x}^{(j)}\| \leqslant y_i \leqslant f^i\left(\mathbf{x}^{(j)}\right) + L^i\|\mathbf{x} - \mathbf{x}^{(j)}\|.$$

So $\mathbf{y}$ has to be in the hyper-rectangle $E_{\mathbf{x}}$ that is the Cartesian product of intervals:

$$E_{\mathbf{x}} := \bigtimes_{i=1}^{n} \left[\max_j \left\{ f^i\left(\mathbf{x}^{(j)}\right) - L^i \|\mathbf{x} - \mathbf{x}^{(j)}\| \right\}, \right.$$
$$\left. \min_j \left\{ f^i\left(\mathbf{x}^{(j)}\right) + L^i \|\mathbf{x} - \mathbf{x}^{(j)}\| \right\} \right] \tag{2.1}$$

It follows that

$$\mathrm{Vol}(E_x) = \prod_{i=1}^{d} \left(\min_j \left\{ f^i\left(\mathbf{x}^{(j)}\right) + L^i \|\mathbf{x} - \mathbf{x}^{(j)}\| \right\} \right.$$
$$\left. - \max_j \left\{ f^i\left(\mathbf{x}^{(j)}\right) - L^i \|\mathbf{x} - \mathbf{x}^{(j)}\| \right\} \right) \tag{2.2}$$

Because f is Lipschitz continuous, this hyper-rectangle cannot be empty. Now we want to find an $\mathbf{x}$ to evaluate for which the expected hypervolume improvement is maximal. For all $\mathbf{x} \in \mathbb{R}^d$ this hypervolume improvement is bounded by $\lambda\left(E_{\mathbf{x}} \backslash \mathrm{Dom}_r(Y_k)\right)$. This is illustrated by the 2-dimensional example in Figure 2.2.

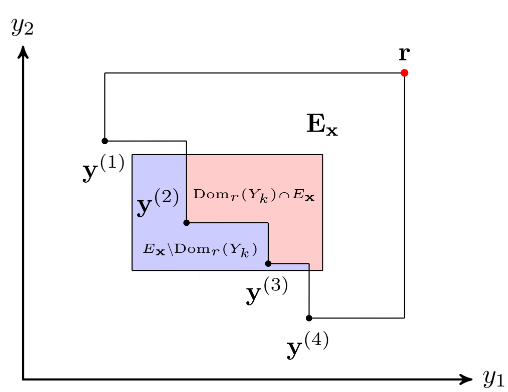

Figure 2.2: *An example of the area $E_{\mathbf{x}}$ in the case $n = 2$ given some set of evaluated points $Y_k = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(4)}\}$.*

There is no specific further information about where the corresponding function value $\mathbf{y}$ of $\mathbf{x}$ will lie in $E_{\mathbf{x}}$, so let Y be a random variable that models the value of $\mathbf{y}$, the outcome of the evaluation $f(\mathbf{x})$. With the information that is available, i.e. S_k and the corresponding values in Y_k, and the Lipschitz constants $\mathbf{L}$ of f, the value of Y is assumed to be homogeneously distributed in $E_{\mathbf{x}}$. The hypervolume improvement of Y will be 0 if $Y \in \mathrm{Dom}_r(Y_k) \cap E_{\mathbf{x}}$. Otherwise it will be $HVI(Y \mid Y_k)$. This yields a precise definition for the expected hypervolume improvement.

Definition 2.3.1. The *expected hypervolume improvement (EI)* of a point $\mathbf{x} \in \mathbb{R}^d$ relative to the set S_k of previously evaluated points is

$$EI(\mathbf{x} \mid S_k) = \mathbb{E}\left[HVI(Y \mid Y_k)\right] = \frac{1}{\text{Vol}(E_{\mathbf{x}})} \int_{E_{\mathbf{x}} \setminus \text{Dom}_r(Y_k)} HVI(\mathbf{y} \mid Y_k) d\mathbf{y}.$$

The volume of $E_{\mathbf{x}}$ is given by (2.2). The key issue is to have a convenient expression for the factor $\int_{E_{\mathbf{x}} \setminus \text{Dom}_r(Y_k)} HVI(\mathbf{y} \mid Y_k) d\mathbf{y}$. In Appendix A, some calculation rules are derived for the hypervolume improvement. Moreover, some explicit calculations of the hypervolume improvement are given for the case that $n = 1$ (Chapter 3) and $n = 2$ (Chapter 4). In the proposed EI-algorithm, we want to select in the $(k+1)$-th step:

$$\mathbf{x}^{(k+1)} \in \arg\max EI(\mathbf{x} \mid S_k).$$

So we need to determine those $\mathbf{x}$ for which $EI(\mathbf{x} \mid S_k)$ is maximized. The goal is to find an expression for this expected hypervolume improvement that is easy to maximize.

2.4 An alternative: hypervolume-based DIRECT

In Wong et al. [8] the algorithm MO-DIRECT-hv is described. This algorithm uses the hypervolume improvement too. It is created with the idea that it is important to have a good balance between diversification, exploration and exploitation. This algorithm is stated in Algorithm 2.

It uses the next equation, Eq. (1):

$$w_j = \frac{1}{\min_{k \in \{-1,1\}} \|f(\mathbf{c}_i + k \cdot \delta \cdot \mathbf{e}_j) - f(\mathbf{c_j})\|}.$$

where $\mathbf{c}_i$ is the center of hyper-rectangle H_i (since the algorithm splits the domain in t hyperrectangle $H_1, \ldots, H_t$). Moreover, δ is is the maximal side length of the hyper-rectangles H_i and $\mathbf{e}_i$ is the i-th unit vector.

However, for this algorithm there is no proof that it converges globally. Our goal is to create an algorithm for which it can be proved that it converges globally.

The biggest difference between MO-Direct-hv and the algorithm of Žilinskas and Žilinskas [10] is that it uses the function values at the center of the hyper-rectangles instead of the function values at the endpoints of the diagonal. Moreover, it uses the hypervolume indicator to select new hyper-rectangles to partition where Žilinskas and Žilinskas [10] use another performance indicator.

Algorithm 2 MO-DIRECT-hv

Require:

vectorial function f to be minimized
search space X
evaluation budget v
hyper-rectangle threshold σ_t
initial $H_1 = \{X\}$

for $t = 1, 2, \ldots, v$ **do**
  Evaluate all centers $\mathbf{c}_i$ of new hyper-rectangles in H_t
  $\mathscr{P} \leftarrow ND(\{f(\mathbf{c}_i) : i \in H_t, \sigma_i \geqslant \sigma_t\})$
  if size $(\mathscr{P} > 2)$ **then**
    $hv_j = hypervolume(\{f(\mathbf{c}_j) : j \in \mathscr{P}\})$
    $P^t_{hv} = \sum(hv_j)$
    $I_t \leftarrow ND(\{(-hv_j, \sigma_j) : j \in \mathscr{P})$
    if $P^t_{hv} - P^{t-1}_{hv} < 0.0001$ and $\max(\mathbf{hv}) > 0.001 \cdot P^t_{hv}$ **then**
      $rank_i = nondominatedsort(\{f(\mathbf{c}_i) : i \in H_t, \sigma_i \geqslant \sigma_t\})$
      $I_t = ND(\{(rank_i, \sigma_i) : i \in H_t, \sigma_i \geqslant \sigma_t\})$
    end if
  else
    $I_t = \mathscr{P}$
  end if
  return I_t
  Partition the hyper-rectangles in I_t using Eq. (1) (which determines in what dimension to split)
  $H_{t+1} = (H_t \backslash I_t) \cup \{I_t\text{'s newly generated hyper-rectangles}\}$
  $t = t + 1$
end for
return $ND\left(\{f(\mathbf{c}_i)\}_{i \in H_t}\right)$

Chapter 3

A special case: single objective optimization in one dimension

Before looking at multi-objective optimization of Lipschitz continuous functions, we first take a closer look at single objective optimization. In the case of an expensive black-box and Lipschitz continuous function, an algorithm has already been introduced: Shubert's Algorithm.

3.1 Shubert's Algorithm

In 1972 Bruno O. Shubert introduced a new algorithm to search for a global minimum of a Lipschitz continuous function $f : [a, b] \to \mathbb{R}$ that satisfies the Lipschitz condition with constant L. The algorithm as described in Shubert [7] is formulated for a maximization problem. However, to be consistent with other literature, we have decided to look at minimization. The algorithm as described below, is thus rewritten to an algorithm that works for the minimization of some function f, which is the same as maximizing $-f$.

Let $\phi = \min_{x \in [a,b]} f(x)$ be the global minimum of the function f, and denote by $\Phi = \{x \in [a, b] : f(x) = \phi\}$ the set of all x for which this minimum is attained. Shubert's Algorithm uses the following sample rule.

We define the sampling sequence $x_0, x_1, x_2, \ldots$ of points from $[a, b]$ recursively as follows:

$$\begin{aligned} &x_0 \in [a, b] \\ &x_{n+1} \text{ such that } F_n(x_{n+1}) = M_n,\ n = 0, 1, 2, \ldots \end{aligned}$$

where

$$F_n(x) = \max_{k=0,\ldots,n} \left(f(x_k) - L|x - x_k|\right) \text{ and}$$
$$M_n = \min_{x \in [a,b]} F_n(x).$$

So $F_n(x)$ is the height of the Lipschitz lower bound for $x \in [a, b]$ and M_n is the lowest point of the Lipschitz lower bound on $[a, b]$. We will call any discontinuity of the derivative of F_n (informally) a *kink* in the lower bound. So any place where the maximum $\max_{k=0,\ldots,n} (f(x_k) - L|x - x_k|)$ switches from holding for x_l to x_m with $x_l \neq x_m$ is called a kink.

In the formulation of Shubert's Algorithm above, an arbitrary x_0 is evaluated in the first iteration. By definition of the algorithm one of the bounds a or b is then evaluated immediately after, as x_1. However, in practice one usually starts with evaluating f at a and b in the first two iterations (see for example Brise [1]). So this means that x_0 is not randomly chosen but set to a and then by definition of the algorithm b is evaluated in the second iteration. We will refer to this version of the algorithm as the *Canonical Shubert Algorithm (CSA)*, e.g. in Subsection 3.1.2. Note that there may be no unique minimum of the lower bound, so that one of the minima can be chosen. Denote by B_L all the Lipschitz continuous functions $f : \mathbb{R} \to \mathbb{R}$ with a Lipschitz constant smaller than L. Now define a CSA-sequence as follows.

Definition 3.1.1. A sequence $\xi = (x_n)_{n \in \mathbb{N}}$ is a *CSA-sequence* in $[a, b]$ for $f \in B_L$ if $x_n \in [a, b]$ satisfy for all $n \in \{0, 1, \ldots\}$

i.) $x_0 = a$,

ii.) $x_{n+1} \in \arg\min_{x \in [a,b]} F_n(x)$.

Note that for any CSA-sequence (x_n), we have $x_1 = b$. Let ξ be a CSA-sequence for an objective function $f \in B_L$. Then $F_n[f, \xi]$ and $M_n[f, \xi]$ denote the corresponding sequences of dented Lipschitz lower bounds and the minimum of the lower bounds in iteration n.

One of the main results of this thesis is the next theorem which will be proved at the end of this section.

Theorem 3.1.2. *The sampling sequence of the Expected Hypervolume Improvement Algorithm as described in Section 2.2 applied to single objective optimization ($n = 1$) of a Lipschitz continuous objective function on $[a, b] \subset \mathbb{R}$ ($d = 1$) will generally follow that of Shubert's Algorithm, but may deviate at steps, occasionally.*

The recursive sampling of the Canonical Shubert Algorithm is visualized in Figure 3.1. In the first two iterations a and b are evaluated. Based on the lower bounds dependent on $f(a)$ and $f(b)$, x_2 is found and evaluated. After these three evaluations, the algorithm looks at the dented lower Lipschitz bounds $F_n(x)$, the grey lines, and searches for the lowest points on these bounds. If there are multiple minima, it chooses one. On this iteration it chooses the point indicated by $F_3(x_3)$. After the evaluation of the objective function at x_3, the algorithm finds new dented lower bounds in between $f(a)$ and $f(x_2)$. These are indicated by the orange lines. Now the minimum of the lower bounds is the point indicated by $F_4(x_4)$. So now the function will be evaluated at x_4. By repeating this, on every iteration a new point is found where the dented lower Lipschitz bound is the lowest and at its x-value the function is evaluated. By doing so the difference between the lowest point and the minimal evaluated function value found so far, decreases. Thus the algorithm gets closer and closer to the global minimal function value. In the next section we will proof that the algorithm indeed converges to the global minimum of the objective function on $[a, b]$.

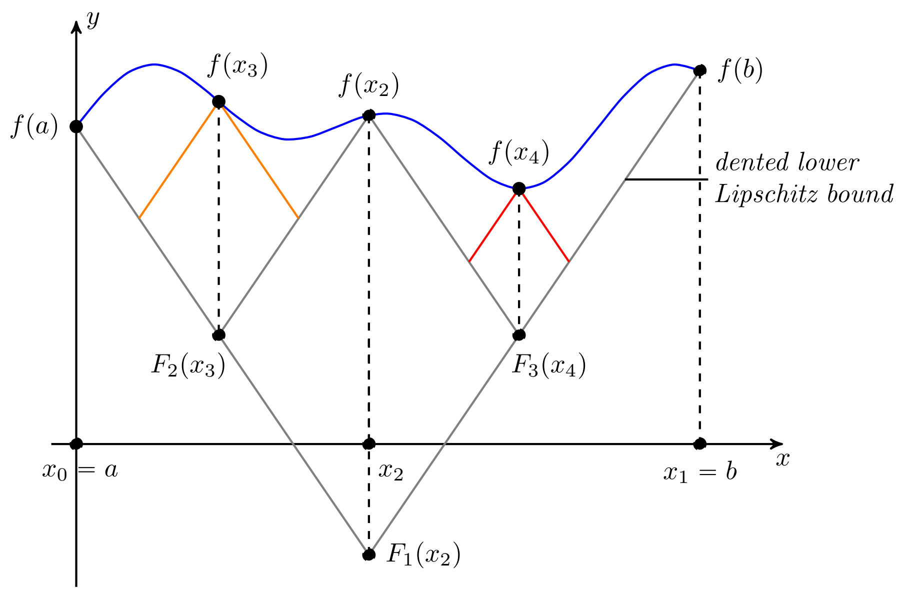

Figure 3.1: *Visualization of the sampling of Shubert's Algorithm.*

3.1.1 Convergence

Shubert's Algorithm as defined above, indeed converges to the global minimum of the function it is used for. The following lemma is found in Shubert [7], Section 2.

Lemma 3.1.3. *For any choice of CSA-sequence ξ for $f \in B_L$, Shubert's Algorithm converges to the global minimum ϕ:*

$$\lim_{n\to\infty} M_n[f, \xi] = \phi.$$

Proof. Let $(x_n)_{n\in\mathbb{N}}$ be a CSA-sequence for f which originated from the Canonical Shubert Algorithm and let $X = \{x_0, x_1, \dots\}$. Assume that there exists $m > n$ such that $x_m = x_n$. Then

$$M_{m-1} = F_{m-1}(x_m) = \max_{k=0,\dots,m-1}\{f(x_k) - C|x_m - x_k|\} = f(x_n),$$

since the maximum is reached for $k = n$ (by the Lipschitz continuity of f). By the definition of the lower bound, we now have that

$$f(x) \geqslant \max_{k=0,\dots,m-1}\{f(x_k) - C|x_m - x_k|\} = f(x_n)$$

for all $x \in [a,b]$, so $f(x_n)$ is a the minimum value of f on $[a,b]$. So then Shubert's Algorithm has already converged to the minimum value. Note that the minimum value can be attained at multiple global minimum points, so that the algorithm will still switch between these points. However, the corresponding minimum value $f(x_n)$ will then be constant.

So suppose that $x_m \neq x_n$ for all $m \neq n$. Because $M_n \leqslant \phi$ for all $n \in \mathbb{N}$ and $M_n \geqslant M_m$ for all $n > m$, the sequence $M_0, M_1, \dots$ is non-decreasing and bounded, so it has a limit point M according to the Monotone Convergence Theorem.

Since X had infinite cardinality, it has at least one limit point in $[a,b]$. Let $\epsilon > 0$ and assume there is a sequence $x_{n_0}, x_{n_1}, \dots$ of points in X with limit point z such that

$$f(z) \geqslant M + \epsilon.$$

Let $k(\epsilon)$ be an index such that

$$k \geqslant k(\epsilon) \Rightarrow |x_{n_k} - z| < \frac{\epsilon}{2L}.$$

Now we have for $k \geqslant k(\epsilon)$:

$$f(x_{n_k}) \geqslant f(z) - L|x_{n_k} - z| > f(z) - \tfrac{1}{2}\epsilon \geqslant M + \tfrac{1}{2}\epsilon.$$

Moreover, we have for all $x \in [a,b]$ that

$$F_{n_k}(x) \geqslant f(x_{n_k}) - L|x - x_{n_k}|,$$

because $F_{n_k}(x) = \max_{i=0,\dots,n_k}\{f(x_i) - L|x - x_i|\}$. Now we have for $k \geqslant k(\epsilon)$ that $|x - x_{n_k}| \leqslant \frac{\epsilon}{2L}$ implies

$$F_{n_k}(x) \geqslant f(x_{n_k}) - L|x - x_{n_k}| > M + \tfrac{1}{2}\epsilon - \tfrac{1}{2}\epsilon = M.$$

It follows from the definition of M that $M_{n_k} \leqslant M$ for all $n \in \mathbb{N}$, so we have that $F_{n_k}(x) > M \geqslant M_{n_k}$ for all $n \in \mathbb{N}$. However, by the definition of M_{n_k} this is impossible, since $M_{n_k} = \max_{x\in[a,b]} F_{n_k}(x)$. So we must have that $|x - x_{n_k}| > \frac{\epsilon}{2L}$ for all $x \in [a,b]$. But by the choice of $k(\epsilon)$ we have $|z - x_{n_k}| \leqslant \frac{\epsilon}{2L}$ and z is a limit point of the sequence $x_{n_0}, x_{n_1}, \dots$ in $X \subset [a,b]$, so this yields a contradiction.

Thus the assumption that there is a limit point z with $f(z) \geqslant M + \epsilon$ for any $\epsilon > 0$ is false. So for all limit points we have $f(z) < M + \epsilon$. Moreover, we have by the definition of M as the limit point of $M_0, M_1, \ldots$ that $f(z) \geqslant M_n \geqslant M$, so we must have that

$$f(z) = M$$

for all limit points z of X. Since $f(x_n) \geqslant \phi \geqslant M$ for all $n \in \mathbb{N}$, it follows that

$$\lim_{n\to\infty} M_n = \phi \text{ and } \lim_{n\to\infty} f(x_n) = \phi.$$

so Shubert's Algorithm converges to the minimum value of the function. □

3.1.2 Error bounds

It is not possible to do infinitely many iterations to find the global minimum. Hence it is interesting to know what the difference is between the best point found with the Canonical Shubert Algorithm after k iterations, $\min_{0\leqslant i\leqslant k} f(x_i)$, and the global minimum. Intuitively it would seem logical to look at the difference between the derivative of the function and its Lipschitz constant. Because the objective function is Lipschitz continuous, it is differentiable almost everywhere. However, even a function of which the derivative is equal to the Lipschitz constant almost everywhere can be a 'worst case scenario' for the Canonical Shubert Algorithm. Namely, a function with $|f'(x)| = L$ for almost every x, but with an enormous amount of kinks between $f'(x) = L$ and $f'(x) = -L$ is very hard to minimize, since all the local minima lie really close to each other. This means the algorithm cannot decide easily on which minimum it has to focus, so it spreads its focus over all the minima which causes a bad performance.

That is why we will look at the deviation between the minimum in the lower bound at iteration k and the minimum of the function found until iteration k to quantify the performance of the Canonical Shubert Algorithm. Note that the minimum of the lower bound is always smaller than the global minimum of the function. We get

$$M_k[f;\xi] \leqslant \min_{x\in[a,b]} f(x) \leqslant \min_{0\leqslant n\leqslant k} f(x_n). \tag{3.1}$$

Here ξ is a CSA-sequence for f in $[a,b]$. We choose to consider the deviation between the minimum in the lower bound and the minimum that has been found until iteration k, because we cannot say much about the real minimum, and this deviation is the maximum deviation between the minimum found by the algorithm after k iterations and the real minimum. We denote this deviation for a function $f : [a,b] \to \mathbb{R}$ by

$$\Delta_k[f,\xi] := \phi_k[f,\xi] - M_k[f,\xi], \tag{3.2}$$

where $\phi_k[f,\xi] = \min\{f(x_i) : x_i \in \xi, 0 \leqslant i \leqslant k\}$. So $\Delta_k[f,\xi]$ is the length of the interval in which the true global minimum point lies, given by the lower and upper estimate in Equation (3.1). Note that $\Delta_k[f,\xi]$ is bounded for a Lipschitz continuous function that satisfies the Lipschitz condition with constant L, because the lower bound for f at iteration k, M_k, is bounded from below by $f(a) - L|b-a|$. This also holds for the minimum of f, $\phi_k[f,\xi]$, and they are both bounded from above by $f(a)$.

Moreover, note that $\Delta_k[f,\xi]$ is a non-increasing sequence in k:

$$\Delta_{k+1}[f,\xi] \leqslant \Delta_k[f,\xi], \tag{3.3}$$

since $\phi_k[f,\xi]$ is a non-increasing sequence by definition and $M_k[f,\xi]$ is non-decreasing. Now denote by $\overline{\Delta}_k[f]$ the worst convergence for this specific $f \in B_L$:

$$\overline{\Delta}_k[f] = \sup\{\Delta_k[f,\xi] \mid \xi \text{ is a CSA-sequence for } f \text{ on } [a,b]\}.$$

Note that $\Delta_k[f,\xi]$ depends on ξ only through $x_0, x_1, \ldots, x_k$. So if $\xi = (x_n)_{n\in\mathbb{N}}$ and $\xi' = (x'_n)_{n'\in\mathbb{N}}$, then $\Delta_k[f,\xi] = \Delta_k[f,\xi']$ if $x_n = x'_n$ for all $n \in \{0,1,\ldots,k\}$. Consequently

$$\overline{\Delta}_k[f] = \max\{\Delta_k[f,\xi] \mid \xi \text{ is a CSA-sequence for } f \text{ on } [a,b]\}.$$

Let B_L and B_L^0 be the following sets

$$\begin{aligned} B_L &:= \{f : [a,b] \to \mathbb{R} \mid f \text{ is Lipschitz continuous with} |f|_{\mathrm{Lip}} \leqslant L\}, \\ B_L^0 &:= \{f \in B_L \mid f(a) = 0\}. \end{aligned}$$

Note that $\Delta_k[f + c\mathbb{1}, \xi] = \Delta_k[f,\xi]$ for every $k \in \mathbb{N}$, $c \in \mathbb{R}$ and f a Lipschitz continuous function that satisfies the Lipschitz condition with constant L, i.e. $\Delta_k[f,\xi]$ is independent of vertical translations of f since then ϕ_k and M_k will be shifted the same amount. Hence, if ξ is a CSA-sequence for f, then it is a CSA-sequence for $f + c\mathbb{1}$ too, for any $c \in \mathbb{R}$. We are looking for

$$\Delta_k^* := \sup_{f\in B_L} \overline{\Delta}_k[f] = \sup_{f\in B_L^0} \overline{\Delta}_k[f],$$

the worst convergence of this algorithm for functions $f \in B_L$. We shall show below that the supremum is attained for a constant function, so it is a maximum. In fact, the main result that we will proof in this section is the following proposition.

Proposition 3.1.4. *For $k = 1, 2, \ldots,$ we have*

$$\Delta_k^* = \overline{\Delta}_k[0] \leqslant \frac{L(b-a)}{k+1}.$$

Remark 3.1.5. In Shubert [7] a proof of Proposition 3.1.4 is given, but this proof appears to be incomplete and some of the assumptions seem to be incorrect.

Let us discuss these issues before giving our proof of the proposition. Here we will use the notation that is used in Shubert [7], where the evaluated points $y_0, \ldots, y_k$ corresponding to ξ are given in parentheses for Δ_k:

We note $\Delta_k[f, \xi](y_0, \ldots, y_k)$ if $f(x_0) = y_0, \ldots, f(x_k) = y_k$ with $\xi = (x_n)_{n \in \mathbb{N}}$.

Firstly, it is stated that because $\Delta_k[f, \xi](y_0, \ldots, y_k)$ can depend only on the first $k + 1$ samples and is clearly a constant with respect to the first sample y_0, the value y_0 does not need to be taken into account when maximizing $\Delta_k[f, \xi](y_0, \ldots, y_k)$. However, $\Delta_k[f, \xi](y_0, \ldots, y_k)$ does not need to be constant with respect to the first sample. As is shown in Figure 3.2, for y_1 constant, $\Delta_1[f, \xi](y_0, y_1)$ varies if y_0 is decreased or increased.

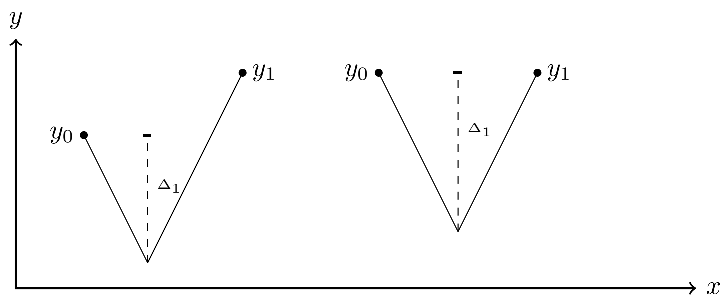

Figure 3.2: *Let $k = 1$ and y_1 be a fixed value. Then $\Delta_k[f, \xi]$ changes in size for different values of y_0.*

It is true that $\Delta_k[f, \xi]$ is independent of vertical translations of the function, i.e.

$$\begin{aligned}\Delta_k[f, \xi](y_0, y_1, \ldots, y_k) &= \Delta_k[f + c, \xi](y_0 + c, y_1 + c, \ldots, y_k + c) \\ &= \Delta_k[f - y_0, \xi](0, y_1 - y_0, y_2 - y_0, \ldots, y_k - y_0).\end{aligned}$$

However, then the notation should be

$$\Delta_k^* = \sup_{y_1 \in Y_1} \cdots \sup_{y_k \in Y_k} \Delta_k[f, \xi](0, y_1, \ldots, y_k).$$

Moreover, in Shubert [7] the function $G_{n,k}$ is defined as $F_{n,k}$ with the $(n-k)$-th term $y_{n-k} + L|x - x_{n-k}|$ left out from the set. So

$$G_{n,k} := \min_{i \in \{0,\ldots,n\} \setminus \{n-k\}} \left(f(x_i) + L|x - x_i| \right).$$

In the proof it is stated that if g, defined as $g := \max_{x \in [a,b]} G_{n,k}(x)$ is smaller than $G_{n,k}(x_{n-k})$, it follows that $M_{n,k} = g$. However, in the example of Figure 3.3 this is not the case. We have $g < G_{5,1}(x_4)$, but also $g < M_{5,1}$.

Furthermore, after proving that $\Delta_n[f, \xi](y_0, \ldots, y_k) \leqslant \Delta_n[f, \xi](y_0, y_0, \ldots, y_0)$ for all $y_0, \ldots, y_k \in \mathbb{R}$, it is concluded that the supremum of Δ_n over $f \in B_L$

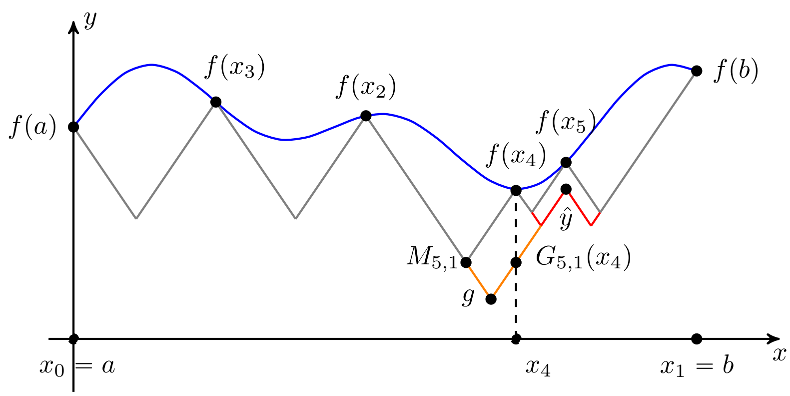

Figure 3.3: *Visualization of the values of* $g, M_{n,k}$ *and* $G_{n,k}(x_{n-k})$ *for some function* f *with* $n = 5$ *and* $k = 1$.

is attained for f a constant function. However, this does not follow from that fact, because Δ_n depends not only on the values of y_i, but also on the location of these values, x_i. Let $x'_0, x'_1, \ldots, x'_n$ be the sampling sequence $x_0, x_1, \ldots, x_n$ sorted by size. Now we have that

$$\Delta_n[f,\xi](y_0, y_0, \ldots, y_0) = \tfrac{1}{2}L \cdot \max_{i=1,\ldots,n}(x'_i - x'_{i-1})$$

which is probably and certainly possibly bigger than the biggest interval between x_i's for a sampling sequence of a constant function.

Because of this, we decided to follow a different approach to establish the validity of this key result, that the convergence rate of the Canonical Shubert Algorithm is bounded from above by $\frac{L(b-a)}{k+1}$. Firstly, we will proof the following lemma. Because $\overline{\Delta}_k[f + c\mathbb{1}] = \overline{\Delta}_k[f]$, instead of considering any constant function $f = c\mathbb{1}$, $c \in \mathbb{R}$, we can just look at the zero map $f(x) = 0$. However, the lemma holds for any other constant function $f = c\mathbb{1}$ too.

Lemma 3.1.6. *Consider the zero map* $f(x) = 0$ *for all* $x \in [a, b]$ *and a CSA-sequence* ξ^0 *corresponding to* f*. Let* $N \in \mathbb{N}_{>1}$*. It follows that*

(i) $\Delta_{2^N}[0,\xi^0] = \frac{1}{2}\Delta_{2^{N-1}}[0,\xi^0]$*;*

(ii) $\Delta_{2^N+j}[0,\xi^0] = \Delta_{2^N}[0,\xi^0]$ *for* $j = 1,\ldots,2^N-1$.

As a consequence

$$\Delta_k[0,\xi^0] \leqslant \frac{L(b-a)}{k+1} \text{ for } k = 1, 2, \ldots. \tag{3.4}$$

Proof. We consider the zero map $f(x) = 0$ with corresponding CSA-sequence ξ^0. Note that we only have to look at $M_k[f,\xi^0]$ to determine $\Delta_k[f,\xi^0]$ since

$\phi_k[f,\xi^0] = 0$ for all $k \in \mathbb{N}_{>0}$. After the evaluation of a and b, which will result in $f(a) = 0$ and $f(b) = 0$, the middle m will be evaluated. Then in the fourth and fifth iteration the middle of $[a, m]$ and of $[m, b]$ will be evaluated and so on. So the minimum of the lower bound after evaluating a is $-L(b-a)$. After the first iteration, where b is evaluated, $-\frac{1}{2}L(b-a)$ will be the minimum of the lower bound and after two and three iterations $-\frac{1}{4}L(b-a)$ will be the minimum, because the lower bound is after two iterations on two points equal to $-\frac{1}{4}L(b-a)$ and after three iterations there is still one point left where it equals $-\frac{1}{4}L(b-a)$. This pattern continues for a bigger amount of iterations. So for $N \in \mathbb{N}_{>1}$, after 2^N iterations there will be 2^N different intervals with minimum of the lower bound, F_{2^N}, equal to $-\frac{1}{2^{N+1}}L(b-a)$. We find that

$$\Delta_{2^N}[0,\xi^0] = \tfrac{1}{2}\Delta_{2^{N-1}}[0,\xi^0]. \tag{3.5}$$

Moreover after this 2^N-th iteration, the minimum of the lower bound will stay the same for $2^N - 1$ iterations, since there are 2^N different places where the lower bound equals $-\frac{1}{2^{N+1}}L(b-a)$ which all have to be evaluated before the global minimum will change, so

$$\Delta_{2^N+j}[0,\xi^0] = \Delta_{2^N}[0,\xi^0] \text{ for } j = 1,\ldots,2^N - 1. \tag{3.6}$$

The fact that $\Delta_1[0,\xi^0] = \frac{1}{2}L(b-a)$ combined with (3.5) and (3.6) now gives us for $k = 1, 2, \ldots$ that

$$\Delta_k[0,\xi^0] \leqslant \frac{L(b-a)}{k+1}. \tag{3.7}$$

□

Now we will look again at all $f \in B_L$ and prove that for all $k = 1, 2, \ldots$ the value of $\Delta_k[f,\xi]$ is at least halved from k to $2k$ iterations.

Lemma 3.1.7. *For all $f \in B_L$, ξ a CSA-sequence for f and $k = 1, 2, \ldots$, it holds that*

$$\Delta_{2k}[f,\xi] \leqslant \tfrac{1}{2}\Delta_k[f,\xi].$$

Proof. For $i \in \mathbb{Z}_{\geqslant 0}$, let x_i be the x-value in which f is evaluated in the i-th iteration according to the CSA-sequence ξ, i.e. $\xi = (x_i)_{i\in\mathbb{N}}$. Now, let $\bar{x}_i^{(k)}$ be the x-value in which f is evaluated before or in the k-th iteration that is closest to x_i on the left side, i.e.

$$\bar{x}_i^{(k)} = \max_j\{x_j : 0 \leqslant j \leqslant k, x_j < x_i\},\ i = 1,\ldots,k.$$

Clearly, the intervals $[\bar{x}_i^{(k)}, x_i]$, $i = 1,\ldots,k$, cover the interval $[x_0, x_1] = [a, b]$, i.e. $[a,b] = \bigcup_{i=1}^k [\bar{x}_i^{(k)}, x_i]$. Let M_k^i be the minimum of the lower bound $F_k(x)$ in the interval $[\bar{x}_i^{(k)}, x_i]$:

$$M_k^i = \min_{x\in[\bar{x}_i^{(k)}, x_i]} F_k(x).$$

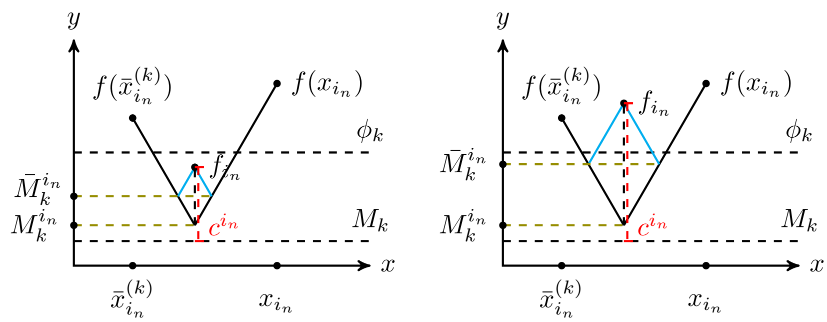

Figure 3.4: *Visualization of the different variables in an interval* $[\bar{x}_{i_n}^{(k)}, x_{i_n}]$. *On the left side the case that* $f_{i_n} < \phi_k$ *and thus* $l_{i_n} = f_{i_n}$. *On the right side the case that* $f_{i_n} > \phi_k$ *and thus* $l_{i_n} = \phi_k$.

Recall that M_k is the global minimum of the lower bound F_k, so it follows that $M_k = \min_{i\in\{1,\ldots,k\}} M_k^i$. Define for $i = 1, \ldots, k$, i.e. for all the intervals $[\bar{x}_i^{(k)}, x_i]$, the set

$$T_i^{(k)} := \{j \in \mathbb{N} \mid k < j \leqslant 2k{+}1, x_j \in [\bar{x}_i^{(k)}, x_i]\}.$$

Note that $T_i^{(k)}$ could be empty if the lower bound M_k^i in the interval $[\bar{x}_i^{(k)}, x_i]$ is too high, for example bigger than the minimum found up to iteration k. Note that there exists an $i^* \in \{1, \ldots, k\}$ for which $|T_{i^*}^{(k)}| \geqslant 2$, because from the $k+1$-th up to and including the $2k+1$-th iteration $k+1$ evaluations take place in the interior of the k intervals. In the "worst case scenario" an evaluation takes place in all the different intervals, i.e. $|T_i^{(k)}| \geqslant 1$, but since j is considered for $k+1$ iterations, at least one set $T_i^{(k)}$ contains 2 or more elements. For $|T_i^{(k)}| \geqslant 2$, let $\check{T}_i^{(k)} = T_i^{(k)} \backslash \{\min T_i^{(k)}\}$ be the set $T_i^{(k)}$ without its minimal element. Define

$$\rho_k := \min_{\substack{1\leqslant i\leqslant k\\ |T_i^{(k)}|\geqslant 2}} (\min \check{T}_i^{(k)}).$$

Because $\check{T}_i^{(k)}$ is the set $T_i^{(k)}$ without its minimal element, it immediately follows from the definition of $T_i^{(k)}$ that $k{+}1 < \rho_k \leqslant 2k{+}1$. Now let

$$n_k := \rho_k - k - 1.$$

So after $n_k + 1$ new iterations after the k-th iteration, f is evaluated for the first time in an interval $[\bar{x}_i^{(k)}, x_i]$, $i \in \{1, \ldots, k\}$, in which f was already evaluated since the k-th iteration. By the definition of n_k, it follows that $1 \leqslant n_k \leqslant k$.

In iteration $n = k + 1, \ldots, k + n_k$ an evaluation takes place in x_n. Now let $i_n \in \{1, \ldots, k\}$ be such that $x_n \in [\bar{x}_{i_n}^{(k)}, x_{i_n}]$. Hence, the global minimum M_n is attained in the interval $[\bar{x}_{i_n}^{(k)}, x_{i_n}]$. That is, $M_k^{i_n} = M_n$. Moreover, by construction of n_k, an evaluation in the interval $[\bar{x}_{i_n}^{(k)}, x_{i_n}]$ takes place for the first time since iteration k. Let f_{i_n} be the f-value corresponding to x_n. Denote by $c^{i_n} := f_{i_n} - M_k$ the distance between the global minimum of the lower bound after k iterations and the new function value f_{i_n}. Moreover, let $\bar{M}_k^{i_n}$ be the new minimum of the lower bound in the interval $[\bar{x}_{i_n}^{(k)}, x_{i_n}]$ when taking f_{i_n} into account. Now we find

$$\bar{M}_k^{i_n} = M_k^{i_n} + \tfrac{1}{2}(f_{i_n} - M_k^{i_n}) = \frac{M_k^{i_n} + f_{i_n}}{2} \tag{3.8}$$

$$\begin{aligned} \text{and } \bar{M}_k^{i_n} - M_k &= \frac{M_k^{i_n} + f_{i_n}}{2} - M_k = \frac{M_k^{i_n} - M_k}{2} + \frac{f_{i_n} - M_k}{2} \\ &\geqslant \frac{f_{i_n} - M_k}{2} = \frac{c^{i_n}}{2}. \end{aligned} \tag{3.9}$$

Define $l_{i_n} := \min\{f_{i_n}, \phi_k\}$. Hence, $\phi_{k+n_k} = \min_{n\in\{k+1,\ldots,k+n_k\}} l_{i_n}$.

$$\begin{aligned} \text{Moreover, } l_{i_n} - M_k &= \min\{f_{i_n}, \phi_k\} - M_k = \min\{f_{i_n} - M_k, \phi_k - M_k\} \\ &= \min\{c^{i_n}, \Delta_k[f,\xi]\}. \end{aligned} \tag{3.10}$$

$$\begin{aligned} \text{Then } \Delta_k[f,\xi] - \tfrac{1}{2}c^{i_n} &= \phi_k - M_k + \frac{M_k - f_{i_n}}{2} \\ &= \phi_k - \tfrac{1}{2}M_k - \tfrac{1}{2}f_{i_n} = \frac{\phi_k - M_k}{2} + \frac{\phi_k - f_{i_n}}{2}. \end{aligned} \tag{3.11}$$

Finally, consider $l_{i_n} - \bar{M}_k^{i_n}$. Suppose that $f_{i_n} \leqslant \phi_k$. Then

$$\begin{aligned} l_{i_n} - \bar{M}_k^{i_n} &= f_{i_n} - \bar{M}_k^{i_n} = f_{i_n} - M_k - (\bar{M}_k^{i_n} - M_k) \\ &= c^{i_n} - (\bar{M}_k^{i_n} - M_k) \overset{(3.9)}{\leqslant} c^{i_n} - \tfrac{1}{2}c^{i_n} = \tfrac{1}{2}c^{i_n} \\ &\leqslant \tfrac{1}{2}\Delta_k[f,\xi]. \end{aligned} \tag{3.12}$$

Else, we have $f_{i_n} > \phi_k$ and thus

$$\begin{aligned} l_{i_n} - \bar{M}_k^{i_n} &= \phi_k - \bar{M}_k^{i_n} = \phi_k - M_k - (\bar{M}_k^{i_n} - M_k) = \Delta_k[f,\xi] - (\bar{M}_k^{i_n} - M_k) \\ &\overset{(3.9)}{\leqslant} \Delta_k[f,\xi] - \tfrac{1}{2}c^{i_n} \overset{(3.11)}{=} \tfrac{1}{2}\Delta_k[f,\xi] + \frac{\phi_k - f_{i_n}}{2} \\ &\leqslant \tfrac{1}{2}\Delta_k[f,\xi]. \end{aligned} \tag{3.13}$$

Hence, we find

$$l_{i_n} - \bar{M}_k^{i_n} \leqslant \tfrac{1}{2}\Delta_k[f,\xi]. \tag{3.14}$$

We now observe, because $n_k \leqslant k$, that

$$\begin{aligned} \Delta_{2k}[f,\xi] &\leqslant \Delta_{k+n_k}[f,\xi] = \phi_{k+n_k} - M_{k+n_k} \\ &= \phi_{k+n_k} - \min_{j\in\{1,\ldots,k+n_k\}} M_{k+n_k}^j \end{aligned} \tag{3.15}$$

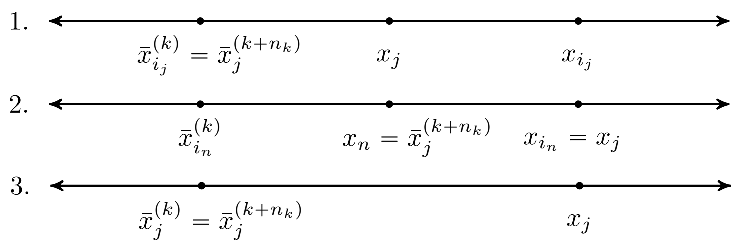

Figure 3.5: *The three options: 1.* $j \in \{k+1, \ldots, k+n_k\}$*, 2.* $j \in \{1, \ldots, k\}$ *and* $j = i_n$ *for some* $n \in \{k+1, \ldots, k+n_k\}$*, and 3.* $j \in \{1, \ldots, k\}$ *and* $j \neq i_n$ *for all* $n \in \{k+1, \ldots, k+n_k\}$.

The cover of $[a, b]$ defined by the intervals $[\bar{x}_j^{k+n_k}, x_j]$, $j = 1, 2, \ldots, k + n_k$ can be described in further detail. The three cases we will distinguish are illustrated in Figure 3.5.

If $k + 1 \leqslant j \leqslant k + n_k$, then $[\bar{x}_j^{(k+n_k)}, x_j] = [\bar{x}_{i_j}^{(k)}, x_j]$, because only at most one evaluation occurred in each interval $[\bar{x}_i^{(k)}, x_i]$. Moreover, $M_{k+n_k}^j = \bar{M}_k^{i_j}$.

If $j = i_n \in \{1, \ldots, k\}$ for some $n \in \{k + 1, \ldots, k + n_k\}$, then $[\bar{x}_j^{(k+n_k)}, x_j] = [x_n, x_{i_n}]$ and $M_{k+n_k}^j = \bar{M}_k^{i_n}$.

In both these cases, for some $n \in \{k + 1, \ldots, k + n_k\}$, we have

$$\phi_{k+n_k} - M_{k+n_k}^j = \phi_{k+n_k} - \bar{M}_k^{i_n} \leqslant l_{i_n} - \bar{M}_k^{i_n} \overset{(3.14)}{\leqslant} \tfrac{1}{2}\Delta_k[f, \xi], \tag{3.16}$$

because $\phi_{k+n_k} = \min_{n' \in \{k+1,\ldots,k+n_k\}} l_{i_{n'}} \leqslant l_{i_n}$.

If $j \in \{1, \ldots, k\}$ and $j \neq i_n$ for all $n \in \{k + 1, \ldots, k + n_k\}$, then $[\bar{x}_j^{(k+n_k)}, x_j] = [\bar{x}_j^{(k)}, x_j]$ and $M_{k+n_k}^j = M_k^j$. Moreover, it follows that $M_k^j \geqslant M_{k+n_k}^i$ for any $i \in \{k + 1, \ldots, k + n_k\} \cup \{i_n : k + 1 \leqslant n \leqslant k + n_k\}$, otherwise an evaluation should have taken place in one of the iterations $k + 1, \ldots, k + n_k$ in the interval $[\bar{x}_j^{(k)}, x_j]$ and thus there would be a $n \in \{k + 1, \ldots, k + n_k\}$ such that $j = i_n$.

Now we find that for this case, for any $n \in \{k + 1, \ldots, k + n_k\}$, we have

$$\phi_{k+n_k} - M_{k+n_k}^j \leqslant \phi_{k+n_k} - \bar{M}_k^{i_n} \leqslant l_{i_n} - \bar{M}_k^{i_n} \overset{(3.14)}{\leqslant} \tfrac{1}{2}\Delta_k[f, \xi] \tag{3.17}$$

too. So, using Equations (3.15), (3.16) and (3.17), we find

$$\Delta_{2k}[f, \xi] \leqslant \tfrac{1}{2}\Delta_k[f, \xi]. \tag{3.18}$$

□

Recall that Proposition 3.1.4 stated that for all $k = 1, 2, \ldots$ it follows that

$$\Delta_k^* \leqslant \frac{L(b-a)}{k+1}.$$

We will now proof this fact.

Proof. [Proposition 3.1.4] Let $k = 1$, so the points a and b are evaluated with corresponding function values $f(a)$ and $f(b)$. As discussed before Δ_k is independent of vertical translations of f, so we can choose without loss of generality that $f(a) = 0$. Now we find for the lower bound that it is the maximum of the lines $y = -Lx + La$ and $y = Lx + f(b) - L(b)$. These two lines intersect at

$$\begin{aligned} x &= \tfrac{1}{2}(a+b) - \frac{f(b)}{L}, \\ y &= \tfrac{1}{2}(L(a-b) + f(a) + f(b)), \end{aligned} \tag{3.19}$$

which gives us the minimum of the lower bound. Now assume without loss of generality that $f(a) \leqslant f(b)$ (if $f(a) > f(b)$ we have a symmetric situation). Then we find

$$\begin{aligned} \Delta_1[f,\xi] = \phi_1 - M_1 &= f(a) - \tfrac{1}{2}(L(a-b) + f(a) + f(b)) \\ &= \tfrac{1}{2}(L(b-a) + f(a) - f(b)) \\ &\leqslant \tfrac{1}{2}L(b-a) = \Delta_1[0,\xi^0]. \end{aligned}$$

Assume that the inequality $\Delta_k[f,\xi] \leqslant \Delta_k[0,\xi^0]$ holds for all $k \in \{1,\ldots,2^N-1\}$. Because of Lemma 3.1.7, we have for $k = 2^N$ that $\Delta_{2^N}[f,\xi] \leqslant \frac{1}{2}\Delta_{2^{N-1}}[f,\xi]$. With the induction hypothesis and Lemma 3.1.6.*(i)*, it follows that

$$\begin{aligned} \Delta_{2^N}[f,\xi] &\leqslant \tfrac{1}{2}\Delta_{2^{N-1}}[f,\xi] \\ &\leqslant \tfrac{1}{2}\Delta_{2^{N-1}}[0,\xi^0] \\ &= \Delta_{2^N}[0,\xi^0]. \end{aligned}$$

Moreover, using Lemma 3.1.6.*(ii)* and the fact that $\Delta_k[f,\xi]$ is non-increasing, it follows for all $j = 1,\ldots,2^N-1$ that

$$\Delta_{2^N+j}[f,\xi] \leqslant \Delta_{2^N}[f,\xi] \leqslant \Delta_{2^N}[0,\xi^0] = \Delta_{2^N+j}[0,\xi^0].$$

So if $\Delta_k[f,\xi] \leqslant \Delta_k[0,\xi^0]$ for all $k = 1,\ldots,2^N - 1$, it follows that

$$\Delta_{2^N+j}[f,\xi] \leqslant \Delta_{2^N+j}[0,\xi^0] \text{ for all } j = 0,\ldots,2^N - 1.$$

Because $\Delta_1[f,\xi] \leqslant \Delta_1[0,\xi^0]$, it follows that

$$\Delta_k[f,\xi] \leqslant \Delta_k[0,\xi^0] \text{ for all } k \in \mathbb{N},$$

for all $f \in B_L$. Now we find that

$$\Delta_k^* = \Delta_k[0,\xi^0] \leqslant \frac{L(b-a)}{k+1}.$$

□

So the convergence rate of Shubert's Algorithm is equal to

$$\frac{L(b-a)}{k+1}.$$

3.2 The expected hypervolume improvement

Let $f : [a, b] \to \mathbb{R}$ be a single objective function and let $S_k \subset [a, b]$ be the set of evaluated points with $Y_k \subset \mathbb{R}$ the set of corresponding function values.

Define for any $x \in [a, b]$: $x^- := \arg\min_{z \in S_k, z < x} x - z$, the evaluated point closest to x on its left side, and $x^+ := \arg\min_{z \in S_k, z > x} z - x$, the evaluated point closest to x on its right side. Moreover, define $y_{\min} := \min_{y \in Y_k} y$, the minimal value found so far. The Lipschitz continuity of f gives us for $x \in [a, b]$

$$\begin{aligned} f(x^-) - L(x - x^-) \leqslant f(x) \leqslant f(x^-) + L(x - x^-) \\ \text{and } f(x^+) + L(x - x^+) \leqslant f(x) \leqslant f(x^+) - L(x - x^+). \\ \text{So } \max\{f(x^-) - L(x - x^-), f(x^+) + L(x - x^+)\} \leqslant f(x) \\ \leqslant \min\{f(x^-) + L(x - x^-), f(x^+) - L(x - x^+)\}. \end{aligned}$$

For the readability define $LB_x := \max\{f(x^-) - L(x - x^-), f(x^+) + L(x - x^+)\}$ and $UB_x := \min\{f(x^-)+L(x-x^-), f(x^+)-L(x-x^+)\}$. Note that $UB_x \geqslant y_{\min}$, since $L(x - x^-) \geqslant 0$, $L(x - x^+) \leqslant 0$ and $f(x^-), f(x^+) \geqslant y_{\min}$.

Recall that E_x is the region of possible f-values when sampling at x based on the Lipschitz continuity of f. Moreover, the expected improvement of choosing x as new sample value, given the samples S_k with corresponding y-values Y_k, is the average hypervolume improvement of every f-value in E_x. By taking the average, we assume that the possible y-values are distributed homogeneously over E_x.

To calculate $EI(x \mid S_k)$, we first note that $\text{Vol}(E_x) = UB_x - LB_x$ and for $y \in E_x \backslash \text{Dom}_r(Y_k)$ we have $HVI(y \mid Y_k) = y_{\min} - y$. We find

$$\begin{aligned} EI(x \mid S_k) = \mathbb{E}(HVI(Y \mid Y_k)) &= \frac{1}{\text{Vol}(E_x)} \int_{E_x \backslash \text{Dom}_r(Y_k)} HVI(y \mid Y_k) dy \\ &= \frac{1}{UB_x - LB_x} \int_{\min\{LB_x, y_{\min}\}}^{y_{\min}} (y_{\min} - y) dy \\ &= \frac{1}{UB_x - LB_x} \left[y \cdot y_{\min} - \frac{1}{2} y^2 \right]_{\min\{LB_x, y_{\min}\}}^{y_{\min}}. \end{aligned}$$

Assume that $LB_x < y_{\min}$, since otherwise we have $EI(x \mid S_k) = 0$. We get

$$\begin{aligned}
EI(x \mid S_k) &= \frac{\left[y y_{\min} - \frac{1}{2}y^2\right]_{LB_x}^{y_{\min}}}{UB_x - LB_x} \\
&= \frac{(y_{\min})^2 - \frac{1}{2}(y_{\min})^2 - \left(LB_x \cdot y_{\min} - \frac{1}{2}LB_x^2\right)}{UB_x - LB_x} \\
&= \frac{\frac{1}{2}(y_{\min})^2 - LB_x \cdot y_{\min} + \frac{1}{2}LB_x^2}{UB_x - LB_x} \\
&= \frac{(y_{\min} - LB_x)^2}{2(UB_x - LB_x)}.
\end{aligned}$$

Now we take as next point to evaluate

$$\begin{aligned}
x_{n+1} &= \arg\max_{\mathbf{x}} \frac{(y_{\min} - LB_x)^2}{2(UB_x - LB_x)} \\
&= \arg\max_{\mathbf{x}} \frac{(\min_{y \in Y_k} y - \max\{f(x^-) - L(x - x^-), f(x^+) + L(x - x^+)\})^2}{2\left(\min\left\{\begin{array}{c} f(x^-) + L(x - x^-), \\ f(x^+) - L(x - x^+) \end{array}\right\} - \max\left\{\begin{array}{c} f(x^-) - L(x - x^-), \\ f(x^+) + L(x - x^+) \end{array}\right\}\right)}.
\end{aligned}$$

Assume without loss of generality that $f(x^-) \leqslant f(x^+)$. If $f(x^-) > f(x^+)$, we have a symmetric situation. To determine for which x the maximum EI is attained, divide the interval between x^- and x^+ in three parts: we call the part from x^- until the point where there is a kink in the lower bound *interval (1)*. We call the part from the kink in the lower bound until the kink in the upper bound *interval (2)* and finally the part from the second kink until x^+ *interval (3)*. In the middle part, the length of $E_{\mathbf{x}}$ is constant, as can be seen in Figure 3.6.

For the location x_L of the kink in the lower bound we have

$$\begin{aligned}
f(x^-) - L(x_L - x^-) &= f(x^+) + L(x_L - x^+) \\
\text{So } x_L &= \tfrac{1}{2}\left(x^- + x^+ + \frac{f(x^-) - f(x^+)}{L}\right) \\
y_L &= f(x^-) - L\left(\tfrac{1}{2}\left(x^- + x^+ + \frac{f(x^-) - f(x^+)}{L}\right) - x^-\right) \\
&= \tfrac{1}{2}\left(f(x^-) + f(x^+) + L(x^- - x^+)\right).
\end{aligned}$$

For the location x_U of the kink in the upper bound we have

$$\begin{aligned}
f(x^-) + L(x_U - x^-) &= f(x^+) - L(x_U - x^+) \\
\text{So } x_U &= \tfrac{1}{2}\left(x^- + x^+ + \frac{-f(x^-) + f(x^+)}{L}\right) \\
y_U &= f(x^-) - L\left(\tfrac{1}{2}\left(x^- + x^+ + \frac{-f(x^-) + f(x^+)}{L}\right) - x^-\right) \\
&= \tfrac{1}{2}\left(f(x^-) + f(x^+) - L(x^- - x^+)\right).
\end{aligned}$$

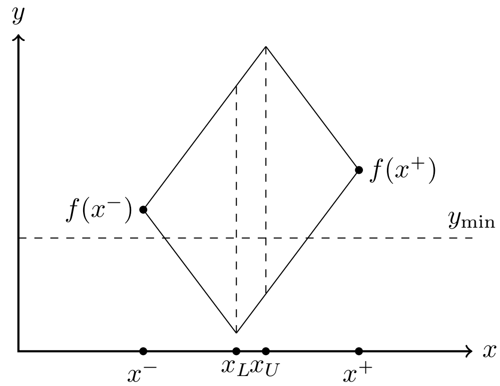

Figure 3.6: *The upper and lower bound for the values of $f(x)$ in between two evaluated points x^- and x^+. For the x-values x_L and x_U, the set E_x of possible values for $f(x)$ is denoted by the vertical dashed line.*

To calculate the length of $E_{\mathbf{x}}$ at (and in between) these kinks, we determine the height of the lower bound at the point of the kink in the upper bound and vice versa. The x-value of the kink in the lower bound, gives us the following corresponding point in the upper bound:

$$\begin{aligned} y = \min\Big\{& f(x^-) + L\left(\tfrac{1}{2}\left(x^- + x^+ + \frac{f(x^-) - f(x^+)}{L}\right) - x^-\right), \\ & f(x^+) - L\left(\tfrac{1}{2}\left(x^- + x^+ + \frac{f(x^-) - f(x^+)}{L}\right) - x^+\right)\Big\} \\ = \min\Big\{& \tfrac{3}{2}f(x^-) - \tfrac{1}{2}f(x^+) - \tfrac{1}{2}L(x^- - x^+), \\ & -\tfrac{1}{2}f(x^-) + \tfrac{3}{2}f(x^+) - \tfrac{1}{2}L(x^- - x^+)\Big\}. \end{aligned}$$

So the length of $E_{\mathbf{x}}$ in the interval $[x_L, x_U]$ is constant and equal to

$$\begin{aligned} \mathscr{L} = \min\Big\{& \tfrac{3}{2}f(x^-) - \tfrac{1}{2}f(x^+) - \tfrac{1}{2}L(x^- - x^+), \\ & -\tfrac{1}{2}f(x^-) + \tfrac{3}{2}f(x^+) - \tfrac{1}{2}L(x^- - x^+)\Big\} \\ & - \tfrac{1}{2}\left(f(x^-) + f(x^+) + L(x^- - x^+)\right) \\ = \min\{& f(x^-) - f(x^+) - L(x^- - x^+), -f(x^-) + f(x^+) - L(x^- - x^+)\} \\ = & L(x^+ - x^-) - \left|f(x^-) - f(x^+)\right|. \end{aligned}$$

Note that $L(x^+ - x^-) \geqslant |f(x^-) - f(x^+)|$ because f is Lipschitz continuous. So we have found the length of $E_{\mathbf{x}}$ in between the two kinks. Because this length is constant, the maximal expected hypervolume improvement,

$$\max \frac{(y_{\min} - LB_x)^2}{2(UB_x - LB_x)},$$

in this interval is found for x on the kink of the lower bound, where $y_{\min} - LB_x$ is maximal. So $x_L = \frac{1}{2}\left(x^- + x^+ + \frac{f(x^-)-f(x^+)}{L}\right)$.

We get for the expected hypervolume improvement of this x that

$$\begin{aligned}
EI(x_L \mid S_k) &= \int_{E_{x_L}\setminus \mathrm{Dom}_r(Y_k)} I_y dy \cdot \frac{1}{\mathrm{Vol}(E_{\mathbf{x}})} \\
&= \frac{\int_{\frac{1}{2}(f(x^-)+f(x^+)+L(x^- - x^+))}^{y_{\min}} (y_{\min} - y)dy}{L(x^+ - x^-) - |f(x^-) - f(x^+)|} \\
&= \frac{\left[y_{\min} \cdot y - \frac{1}{2}y^2\right]_{\frac{1}{2}(f(x^-)+f(x^+)+L(x^- - x^+))}^{y_{\min}}}{L(x^+ - x^-) - |f(x^-) - f(x^+)|} \\
&= \frac{\left(y_{\min} - \frac{1}{2}\left(f(x^-) + f(x^+) + L(x^- - x^+)\right)\right)^2}{2\left(L(x^+ - x^-) - |f(x^-) - f(x^+)|\right)}.
\end{aligned}$$

Now we take a closer look at interval (1), the interval $[x^-, x_L]$. By the definition of $y_{\min} = \min_{x \in S_k} f(x)$, we have that $f(x^-), f(x^+) \geqslant y_{\min}$. In this area the lower limit of $E_{\mathbf{x}}$ lies on the line $y = f(x^-) - L(x - x^-)$ and the upper limit lies on $y = f(x^-) + L(x - x^-)$, so the length of $E_{\mathbf{x}}$ is $2L(x - x^-)$. For the expected hypervolume improvement we get for all $x \in [x^-, x_L]$:

$$\begin{aligned}
EI(x \mid S_k) &= \int_{E_{\mathbf{x}}\setminus \mathrm{Dom}_r(Y_k)} I_y dy \cdot \frac{1}{\mathrm{Vol}(E_{\mathbf{x}})} \\
&= \frac{\int_{f(x^-)-L(x-x^-)}^{y_{\min}} (y_{\min} - y)dy}{2L(x - x^-)} \\
&= \frac{\left[y_{\min} y - \frac{1}{2}y^2\right]_{f(x^-)-L(x-x^-)}^{y_{\min}}}{2L(x - x^-)} \\
&= \frac{\frac{1}{2}y_{\min}^2 - (f(x^-) - L(x - x^-))y_{\min} + \frac{1}{2}(f(x^-) - L(x - x^-))^2}{2L(x - x^-)} \\
&= \frac{\left(y_{\min} - (f(x^-) - L(x - x^-))\right)^2}{4L(x - x^-)}.
\end{aligned}$$

Analogously we find for $x \in [x_U, x^+]$ that

$$EI(x \mid S_k) = \frac{(y_{\min} - (f(x^+) + L(x - x^+)))^2}{-4L(x - x^+)}.$$

The upper bound in interval (1) is $f(x^-) + L(x - x^-)$. Because $x - x^- > 0$ for all x in interval (1), we have that $f(x^-) + L(x - x^-)$ is dominated for all x in interval (1). Since $f(x^-) \geqslant y_{\min}$, we have that the lower bound is dominated from $f(x^-)$ to $f(x^-) - L(x - x^-) = y_{\min}$. As is illustrated in Figure 3.6, the part of E_x for x in interval (1) lies as much above $f(x^-)$ as below. Moreover, $y_{\min}$ lies on the same height or below $f(x^-)$. So the fraction of the non-dominated part of E_x is bigger than or at least as big as on all the other places in interval (1) for $x = x_L$. So the fraction $\frac{y_{\min}-(f(x^-)-L(x-x^-))}{4L(x-x^-)}$ is the biggest for this x. Moreover, the numerator of this fraction is the biggest at this point, because the the non-dominated part of E_x is the biggest for this x. Since the numerator of the expected hypervolume improvement $EI(x \mid S_k)$ is the square of $y_{\min} - (f(x^-) - L(x - x^-))$, its value is maximized in the kink of the lower bound for x in interval (1). As is mentioned before, for interval (2) the expected hypervolume improvement is also maximized at the kink. Finally, interval (3) is symmetrically the same as interval (1), except that it lies higher because $f(x^+) > f(x^-)$ and thus the non-dominated part is smaller, because $y_{\min}$ is constant. So in interval (3) there is no point where the expected hypervolume improvement is higher than on the kink in the lower bound. If $f(x^+)$ would be smaller than $f(x^-)$, we would have a symmetric case, so again $EI(x \mid S_k)$ would be the biggest for x on the place where the kink in the lower bound is.

Assume we have sorted all the evaluated function points, $x^{(i)}$, from smallest x-value to biggest, with corresponding function value $y^{(i)}$. Note by $x_{L,i}$ the x-value of the kink in the lower bound in between $x^{(i)}$ and $x^{(i+1)}$:

$$x_{L,i} := \tfrac{1}{2}\left(x^- + x^+ + \frac{f(x^-) - f(x^+)}{L}\right).$$

Moreover, put $y_{L,i} := \min\{f(x^{(i)} - L(x_{L,i} - x^{(i)}), f(x^{(i+1)} + L(x_{L,i} - x^{(i+1)}\}$, i.e. $y_{L,i}$ is the height of the lower bound at $x_{L,i}$. Let $z_i := y_{\min} - y_{L,i}$ and $w_i := \min\{y^{(i)}, y^{(i+1)}\} - y_{\min}$. Because the gradient of the upper bound and lower bound are respectively L and $-L$, we have for $x_{L,i}$ that

$$UB_{x_{L,i}} - LB_{x_{L,i}} = 2(\min\{y^{(i)}, y^{(i+1)}\} - y_{L,i}).$$

Now we find that after n iterations we have

$$\begin{aligned}
i_{\min,n} &= \underset{x\in\{1,n-1\}}{\arg\min} \frac{(y_{\min} - LB_x)^2}{2(UB_x - LB_x)} = \underset{i\in\{1,\dots,n-1\}}{\arg\min} \frac{(y_{\min} - LB_{x_{L,i}})^2}{2(UB_{x_{L,i}} - LB_{x_{L,i}})} \\
&= \underset{i\in\{1,\dots,n-1\}}{\arg\min} \frac{(y_{\min} - y_{L,i})^2}{4(\min\{y^{(i)}, y^{(i+1)}\} - y_{L,i})} \\
&= \underset{i\in\{1,\dots,n-1\}}{\arg\min} \frac{z_i^2}{4(z_i + w_i)} \\
x_{n+1} &= \underset{x\in[a,b]}{\arg\max}\, EI(x \mid S_k) = \underset{x\in[a,b]}{\arg\max} \frac{(y_{\min} - LB_x)^2}{2(UB_x - LB_x)} \\
&= \tfrac{1}{2}\left(x^{(i_{\min,n})} + x^{(i_{\min,n}+1)} + \frac{f(x^{(i_{\min,n})}) - f(x^{(i_{\min,n}+1)})}{L}\right).
\end{aligned}$$

3.3 EI-method in comparison with Shubert's Algorithm

In this section we will proof Theorem 3.1.2, i.e. that the sampling sequence of the Expected Hypervolume Improvement Algorithm as described in Section 2.2 applied to single objective optimization ($n = 1$) of a Lipschitz continuous objective function on $[a, b] \subset \mathbb{R}$ ($d = 1$) will generally follow that of Shubert's Algorithm, but may deviate at steps, occasionally.

Proof. [Theorem 3.1.2] Let $x^{(0)} < x^{(1)} < \cdots < x^{(k-1)}$ be the enumeration of X_k in increasing order and put $y^{(i)} := f(x^{(i)})$. In Shubert's Algorithm the next point x_k is chosen at a position where $F_{k-1}(x)$ is minimal. F_{k-1} is the minimum of the functions $\min\{f(x^{(i)} - L(x - x^{(i)}), f(x^{(i+1)} + L(x - x^{(i+1)})\}$, $x \in [x^{(i)}, x^{(i+1)}]$, $i \in \{0, 1, \ldots, k-2\}$. Then $x_k = x_{L,i*}$ for index i^* for which $y_{L,i*}$ is minimal. Hence, $z_{i*} = y_{\min} - y_{L,i*}$ is maximal.

In our EI-algorithm the next point x_k is chosen where $EI(x \mid X_k)$ is maximal. In the previous section it is proved that x_k is one of the points $x_{L,i}$. Namely the $x_{L,i}$ for which $\frac{z_i^2}{4(z_i+w_i)}$ is maximal.

The maximum value is not necessarily attained for i with maximal z_i (but that can be the case). The values w_i must be taken into account too. Thus, the EHVI algorithm may select a next point x_k different from that selected by Shubert's Algorithm. □

It remains to be investigated how this phenomenon affects convergence rates to the global minimum.

3.4 A comment on single objective optimization in two dimensions

To get a better idea of how optimization works for in multiple dimensions, we take a closer look at the case that $d = 2$. We will still look at single objective optimization. So for the objective function we have $f : A \subset \mathbb{R}^2 \to \mathbb{R}$. In Novak and Ritter [6], a method is suggested for single objective optimization in multiple dimensions. The evaluation sequence $x_0, x_1, \ldots, x_n$ is determined using hyperbolic cross points. However, the results on the speed of the algorithm are mostly achieved numerically. It is just proved in Novak and Ritter [6] that this method using hyperbolic cross points converges faster than methods using grid points for functions of type $\sum_{i,j} f(x_i, x_j)$. It is not directly clear how the speed will be influenced for a worst case scenario objective function. Thus we consider a method where we can use the hypervolume improvement to determine the evaluation sequence.

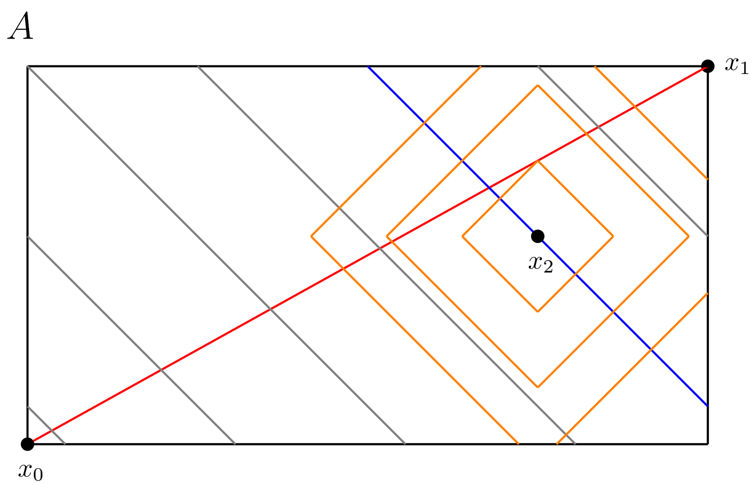

Figure 3.7: *Visualization of the level sets of the lower bounds in two dimensions.*

We will start by evaluating the endpoints of the diagonal of the domain. Since we use the Manhattan metric, there are level lines on which the distance of any point on the line to x_0 is constant and the distance to x_1 is constant too. If we determine the minimum of the lower Lipschitz bound on the diagonal, we find that the set of points where the lower Lipschitz bound is minimal is the level line through this minimum on the diagonal, as is depicted by the blue line in Figure 3.7. This is because all the points on these line have equal distance to x_0 and equal distance to x_1 in the Manhattan metric we use. Now one has a choice in which point to evaluate on the blue line. We suggest to evaluate the function in the middle of the line, since this will yield the most strict lower bounds at both endpoints of the blue line. This evaluation point is depicted by x_2 and the orange lines are the lower bounds that follow from the function value of f at x_2. However, another choice could be made on which point on the level line to evaluate. For example, one could choose to start with an evaluation on one of the endpoints or both the endpoints of the level line.

Because we are optimizing a single objective function, the set that describes the possible values that can be attained, $E_{\mathbf{x}}$, is an interval. Let $f_{\min} = \min_i f(\mathbf{x}_i)$, then we only need to take the $\mathbf{x}$'s into account for which the interval $E_{\mathbf{x}}$ contains values smaller than $f_{\min}$. The set $\{\mathbf{x} \in A : \text{for all } y \in E_{\mathbf{x}} : y > f_{\min}\}$ of all $\mathbf{x}$ for which $E_{\mathbf{x}}$ is fully dominated, does not need to be taken into account for the rest of the algorithm and will only grow for every iteration, because $f_{\min}$ will only get smaller or stay the same and the bounds of the interval $E_{\mathbf{x}}$ will also get stricter or stay the same. By excluding more and more of the domain, the algorithm will be faster in every iteration.

Chapter 4

The expected hypervolume improvement for bi-objective optimization

As we understand now how the expected hypervolume improvement behaves in one-dimensional cases, we make a first step in extending this to the bi-objective case. So in this chapter we will consider objective functions $f : \mathbb{R}^d \to \mathbb{R}^2$ and take a closer look at the expression for the expected hypervolume improvement and its behavior.

As discussed in Section 2.2 and 2.3, when we introduced the proposed EI-algorithm for Lipschitz continuous multi-objective functions, the term 'expected' relates to the mathematical expected value of a suitable random variable; Namely, the hypervolume improvement when a new evaluation results in a value Y, relative to the values $Y_k = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(k)}\}$ already obtained. Knowledge of the points $S_k = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}\}$ of evaluation and the Lipschitz constants, allow to single out a region $E_{\mathbf{x}}$ where Y must lie.

4.1 Introduction: the expected hypervolume improvement for the Gaussian approach

In Emmerich et al. [3] a different probabilistic model for the location Y of the new evaluation value is used than the one we introduce. There, a Gaussian random distribution is used. The hypervolume improvement (and its expected value) is computed using 'stripes'. The area where still improvement can be made, is partitioned in $k+1$ stripes where k is the number of points that have been evaluated already. The values of f at the evaluated points are denoted by $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(k)}$ and are ordered such that the first coordinates are decreasing. We assume that the set $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(k)}\}$ is minimal, in the sense that none of the elements dominates any other element. Moreover, we assume that $y_1^{(1)} < r_1$, the first coordinate of the reference point $\mathbf{r}$. So $y_1^{(1)} > y_1^{(2)} > \cdots > y_1^{(k)}$. Also $y_2^{(1)} < y_2^{(2)} < \cdots < y_2^{(k)} < r_2$. Stripe S^i is for $i = 2, \ldots, k$ the stripe with y_1-values between the y_1-value of evaluated point $\mathbf{y}^{(i)}$ and $\mathbf{y}^{(i-1)}$ and y_2-value lower than the y_2-value of $\mathbf{y}^{(i)}$. Moreover, stripe S^1 is the stripe with y_1-values between $y_1^{(1)}$ and r_1, the y_1-value of reference point $\mathbf{r}$, and y_2-value lower than $y_2^{(1)}$. Finally, stripe S^{k+1} is the stripe with y_1-value lower than the y_1-value of $\mathbf{y}^{(k)}$ and all y_2-values values less than that of the reference point $\mathbf{r} \in \mathbb{R}^2$. This is illustrated in Figure 4.1.

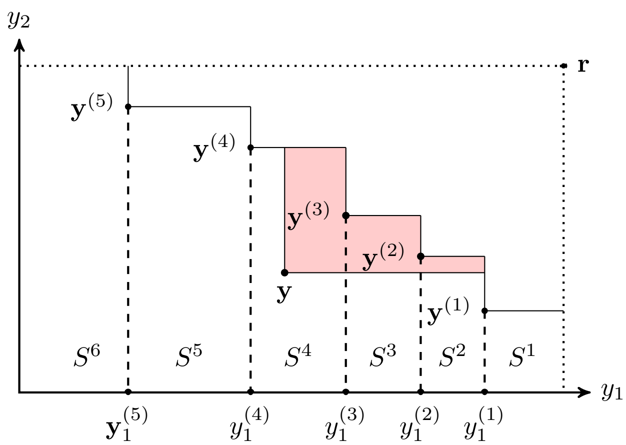

Figure 4.1: *The use of stripes S^i to calculate the hypervolume improvement of a point* $\mathbf{y}$.

The expected hypervolume is now a integral over the area where still improvement can be made, which is split in the sum of integrals over the stripes. By splitting the dominated area in this way, the integral over this area can be written as a sum of exponential functions and error functions, which can be calculated easier.

4.2 General expression for the expected hypervolume improvement

Let Y_k be the set of non-dominated function values found after k iterations and let $\mathbf{y} = (y_1, y_2) \in \mathbb{R}^2$ be the function value we will find in the next iteration, i.e. the next function evaluation. Recall that $Y_{\mathbf{y}}^-$ was defined in Section 2 to be $Y_{\mathbf{y}}^- := \{\mathbf{z} \in \mathbb{R}^2 \mid \mathbf{y} \prec \mathbf{z}\}$ and $Y_{\mathbf{y}}^0 := Y_k \backslash Y_{\mathbf{y}}^-$ to be the set of all function values in Y_k that are not dominated by $\mathbf{y}$. Assume we have an $\mathbf{y}$ that dominates m elements $\mathbf{z}_1, \ldots, \mathbf{z}_m$ in Y_k, and write $\mathbf{z}_j = (z_{j,1}, z_{j,2})$ for all $j \in \{1, \ldots, m\}$.

Define

$$\begin{aligned} \mathbf{z}_0 &:= \underset{\mathbf{y}^{(r)} \in Y_{\mathbf{y}}^0, y_1^{(r)} > y_1}{\arg\min} \quad y_1^{(r)} - y_1 \\ \mathbf{z}_{m+1} &:= \underset{\mathbf{y}^{(u)} \in Y_{\mathbf{y}}^0, y_2^{(u)} > y_2}{\arg\min} \quad y_2^{(u)} - y_2 \end{aligned} \tag{4.1}$$

to be respectively the non-dominated function value $\mathbf{y}^{(r)}$ closest to $\mathbf{y}$ on the right side and function value $\mathbf{y}^{(u)}$ closest to $\mathbf{y}$ above, as is illustrated in Figure 4.2. Now we have for any $\mathbf{y}$ that does not dominate any element of Y_k, i.e. $m = 0$, and is not dominated by any element in Y_k that

$$HVI(\mathbf{y} \mid Y_k) = (z_{0,1} - y_1)(z_{1,2} - y_2).$$

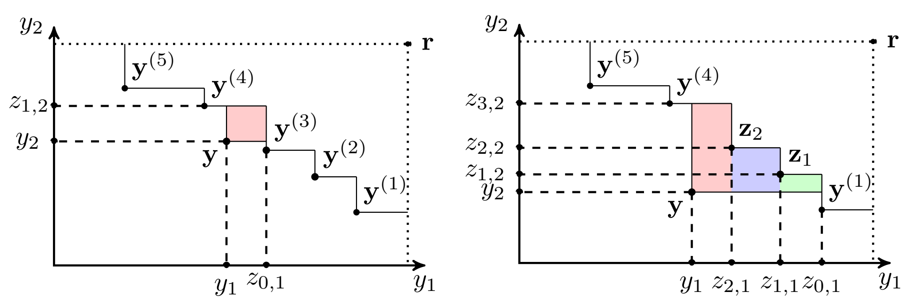

Figure 4.2: *Visualization of the hypervolume improvement of a point* $\mathbf{y}$ *given a set* $Y_k = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(i)}\}$ *of already found non-dominated function values. Left:* $\mathbf{y}$ *does not dominate any* $\mathbf{y}^{(l)} \in Y_k$. *Right:* $\mathbf{y}$ *dominates* $\mathbf{z_1}$ *and* $\mathbf{z_2}$. *The dominated area is divided in 3 stripes, the red, blue and green one.*

Now assume that $\mathbf{y}$ dominates $m > 0$ points in Y_k. Redefine $z_{m+1,1} := y_1$. We divide the dominated area in $m + 1$ vertical stripes, with width

$$z_{m,1} - z_{m+1,1}, z_{m-1,1} - z_{m,1}, \ldots, z_{0,1} - z_{1,1}.$$

These stripes have respectively height

$$z_{m+1,2} - y_2, z_{m,2} - y_2, \ldots, z_{1,2} - y_2.$$

An example for $m = 2$ is illustrated in Figure 4.2. So the hypervolume improvement of a point $\mathbf{y} \in \mathbb{R}^2$ that dominates m points is

$$HVI(\mathbf{y} \mid Y_k) = \sum_{j=1}^{m+1} (z_{j-1,1} - z_{j,1})(z_{j,2} - y_2).$$

We find that

$$\mathbf{x}^{(k+1)} = \arg\max \frac{1}{\mathrm{Vol}(E_{\mathbf{x}})} \int_{E_{\mathbf{x}} \setminus \mathrm{Dom}_r(Y_k)} HVI(\mathbf{y} \mid Y_k) d\mathbf{y}$$
$$= \arg\max \frac{\int_{E_{\mathbf{x}} \setminus \mathrm{Dom}_r(Y_k)} \sum_{j=1}^{m+1} (z_{j-1,1} - z_{j,1})(z_{j,2} - y_2) d\mathbf{y}}{\mathrm{Vol}(E_{\mathbf{x}})},$$

where

$$\mathrm{Vol}(E_{\mathbf{x}}) = \prod_{i=1}^{2} \min_j \left\{ f^i \left(\mathbf{x}^{(j)} \right) + L^i \|\mathbf{x} - \mathbf{x}^{(j)}\| \right\} - \max_j \left\{ f^i \left(\mathbf{x}^{(j)} \right) - L^i \|\mathbf{x} - \mathbf{x}^{(j)}\| \right\},$$

as was already given in Equation (2.2).

4.3 Computation by HVI-tent structures

Recall that $E_{\mathbf{x}}$ is the area in which $f(\mathbf{x})$ can lie for some $\mathbf{x} \in \mathbb{R}^d$, given that f satisfies the Lipschitz condition with constant L. To find an explicit way to calculate the expected hypervolume improvement for some $\mathbf{x} \in \mathbb{R}^d$ relative to the set Y_k of previous evaluations without an integral in the formula, we will split the area $E_{\mathbf{x}} \setminus \mathrm{Dom}_r(E_{\mathbf{x}})$ in three different categories of regions: the areas in which no point dominates any point in Y_k, called P_I-patches, the areas in which every point dominates exactly one point in Y_k, called P_{II}-patches, and the areas in which every point dominates multiple points in Y_k, called P_{III}-patches.

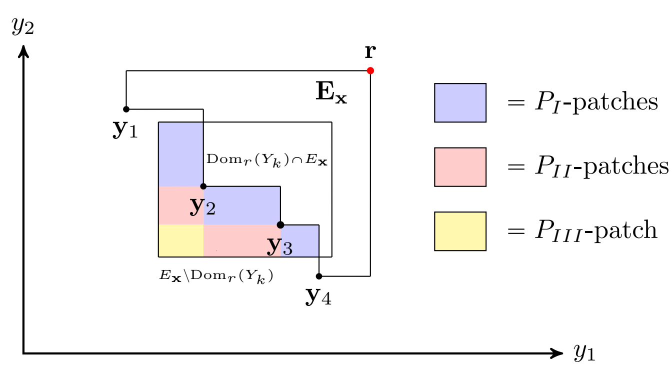

Figure 4.3: *The location of the P_I-patches, P_{II}-patches and P_{III}-patch.*

As can be seen in Figure 4.3, all the patches are of rectangular form. We will call the width of the rectangle Δy_1 and the height Δy_2. If there are places in the area $E_{\mathbf{x}}$ where more than two points are dominated, we split the P_{III}-patches in rectangles where an equal amount of points in Y_k is dominated by those in the rectangle, so that the P_{III}-patches are also of rectangular form. So actually these rectangles are P_{III}-, P_{IV}-, P_V-patches, etc., but because all these patches have corresponding corner heights $u, v, w > 0$, we call all these patches P_{III}-patches.

Now we will use '*HVI-tent structures*' which have as base the rectangular P_I-, P_{II}- or P_{III}-patches and as height the hypervolume improvement of the corresponding coordinate of the base. That is, for every patch $P_i \subset E_{\mathbf{x}}$, we put

$$T_i := \{(y_1, y_2, z) \in \mathbb{R}^3 \mid \mathbf{y} = (y_1, y_2) \in P_i, 0 \leqslant z \leqslant HVI(\mathbf{y} \mid Y_k)\}.$$

Thus T_i in $\mathbb{R}^3$ has the shape of halve a tent. Therefore we call it a '*HVI-tent structure*'. Call the hypervolume improvement of the upper right corner u, of the lower right corner v and of the upper left corner w. Necessarily one has $u \leqslant v$ and $u \leqslant w$. From the formula of the hypervolume improvement it follows that the height of the lower left corner is $v + w + \Delta y_1 \Delta y_2$. A schematic representation is given in Figure 4.4. Here the base EFGH corresponds to any rectangular patch P with E the lower left corner, F the lower right corner etc.

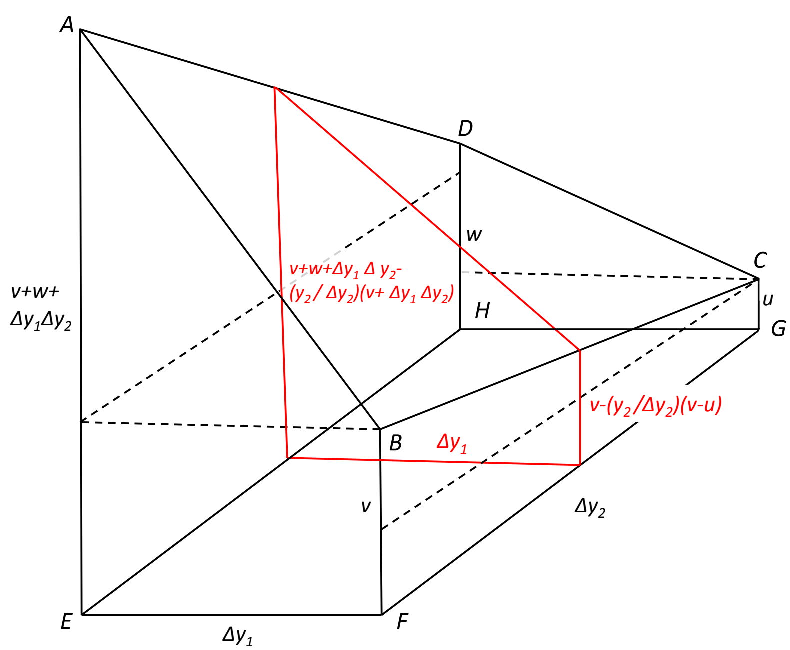

Figure 4.4: *A sketch of a HVI-tent structure with width Δy_1 and depth Δy_2. In red a cross-section of the tent structure for some $y_2 \in [0, \Delta y_2]$ is given (with the y_2-position relative to the y_2-value of the left lower corner E).*

Note that on the sides of a HVI-tent structure, the hypervolume improvement, i.e. the height of the tent structure, increases linearly. However, on the diagonal it increases quadratically. So the 'roof' of the HVI-tent structure, the surface ABCD, is actually a curved surface, not a plane. A representation of this 'roof' of a HVI-tent structure is given in Figure 4.5.

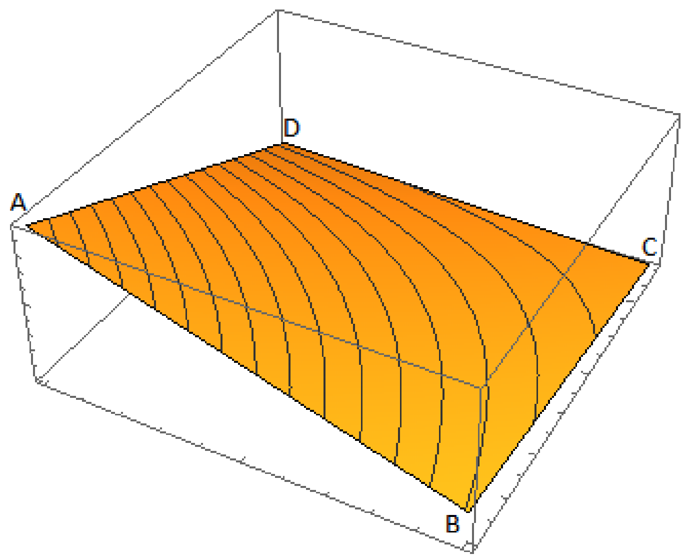

Figure 4.5: *The 'roof' of a P_{II}-tent structure, i.e. the hypervolume improvement for every point in the xy-area. On the sides the height increases linearly but on the diagonal from C to A it increases quadratically.*

Recall that

$$EI(\mathbf{x} \mid S_k) = \frac{1}{\text{Vol}(E_{\mathbf{x}})} \int_{E_{\mathbf{x}} \setminus \text{Dom}_r(Y_k)} HVI(\mathbf{y} \mid Y_k) d\mathbf{y}$$

and

$$HVI(\mathbf{y} \mid Y_k) = \sum_{j=1}^{m+1} (z_{j-1,1} - z_{j,1})(z_{j,2} - y_2),$$

where m is the amount of by $\mathbf{y}$ dominated points in Y_k. So we want to determine the integral over the non-dominated area of the hypervolume improvement which is the sum of the volume of all the tent structures. Therefore we are looking for a way to determine the volume of a single HVI-tent structure, given the width and depth of the base and height of the corners.

Lemma 4.3.1. *Let T be a HVI-tent structure with heights (u, v, w) and patch size $(\Delta y_1, \Delta y_2)$. Then*

$$Vol(T) = \tfrac{1}{4}\Delta y_1 \Delta y_2 (\Delta y_1 \Delta y_2 + u + 2v + 2w) \tag{4.2}$$

Remark 4.3.2. Note that the second factor is the sum of all 'poles' of the tent structure, i.e. the heights of the four corners of the base.

Remark 4.3.3. The hypervolume improvement that is represented on the vertical axis in the tent structure T has the dimension of an area (length-squared). This function is integrated over the two-dimensional patch P. Thus, the integral, i.e. Vol(T), is a four-dimensional hypervolume. This is consistent with expression (4.2).

Proof. To calculate the volume of a tent structure T, we need to integrate in both the y_1 and y_2 direction. Assume we are going to integrate in the y_1 direction. If we make a cross-section of the tent structure for a certain y_2-value $y_2 \in [0, \Delta y_2]$, it is in the form of a trapezium, which can be split into a rectangle and triangle. A representation is given in Figure 4.6, see also Figure 4.4. We take $y_2 = 0$ for the 'cross-section' AEFB and $y_2 = \Delta y_2$ for the 'cross-section' CDHG.

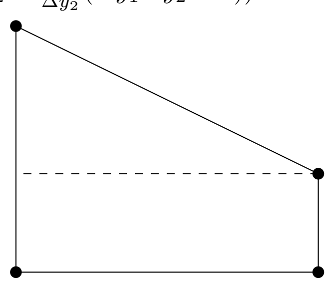

Figure 4.6: *A representation of a cross-section of a HVI-tent structure for a certain y-value with corresponding x- and z-values.*

The height of the rectangle is $v - \frac{y_2}{\Delta y_2}(v - u)$, because it changes linear from v to u over a distance of Δy_2. Now consider the height of the triangle, α. It is maximal for $y_2 = 0$ where it is the height of the lower left corner, $v + w + \Delta y_1 \Delta y_2$, minus the height of the rectangle, v. For $y_2 = \Delta y_2$ the height is w, the height of the upper left corner, minus u, the height of the upper right corner. So it changes linearly from $w + \Delta y_1 \Delta y_2$ to $w - u$. Thus we find that the height is equal to $\alpha = w + \Delta y_1 \Delta y_2 - \frac{y_2}{\Delta y_2}(\Delta y_1 \Delta y_2 + u)$. So the area of the trapezium shaped cross-section of tent structure T at the cross-section for y_2 is equal to

$$\begin{aligned} A_T(y_2) &= \Delta y_1 \left(v - \frac{y_2}{\Delta y_2}(v - u) \right) + \tfrac{1}{2} \Delta y_1 \left(w + \Delta y_1 \Delta y_2 - \frac{y_2}{\Delta y_2}(\Delta y_1 \Delta y_2 + u) \right) \\ &= \Delta y_1 (v + \tfrac{1}{2} w) - \frac{y_2 \Delta y_1}{\Delta y_2}(v - \tfrac{1}{2} u) + \tfrac{1}{2}(\Delta y_1)^2 (\Delta y_2 - y_2). \end{aligned} \tag{4.3}$$

Now we integrate over y_2. We find for the volume of the HVI-tent structure that

$$\begin{aligned}
\mathrm{Vol}(T) &= \int_0^{y_2} A_T(y_2)dy_2 \\
&= \int_0^{\Delta y_2} \Delta y_1(v+\tfrac{1}{2}w) - \frac{y_2\Delta y_1}{\Delta y_2}(v-\tfrac{1}{2}u) + \tfrac{1}{2}(\Delta y_1)^2(\Delta y_2 - y_2)dy_2 \\
&= \left[\Delta y_1(v+\tfrac{1}{2}w)y_2 - \frac{y_2^2\Delta y_1}{2\Delta y_2}(v-\tfrac{1}{2}u) + \tfrac{1}{2}(\Delta y_1)^2(y_2\Delta y_2 - \tfrac{1}{2}y_2^2)\right]_0^{\Delta y_2} \\
&= \Delta y_1\Delta y_2(v+\tfrac{1}{2}w) - \tfrac{1}{2}\Delta y_1\Delta y_2(v-\tfrac{1}{2}u) + \tfrac{1}{4}(\Delta y_1)^2(\Delta y_2)^2 \\
&= \tfrac{1}{4}(\Delta y_1\Delta y_2)\left(\Delta y_1\Delta y_2 + u + 2v + 2w\right). \qquad (4.4)
\end{aligned}$$

□

Note that for a tent structure T_I of a P_I-patch it holds that $u = 0$. Moreover $v = 0$ and/or $w = 0$. For a T_I tent structure for a P_{II}-patch we have that only $u = 0$. So we have found the following expressions for the different kinds of hypervolume tent structures:

$$\begin{aligned}
\mathrm{Vol}(T_I) &= \tfrac{1}{4}(\Delta y_1\Delta y_2)(\Delta y_1\Delta y_2 + 2w) \\
\text{or } \mathrm{Vol}(T_I) &= \tfrac{1}{4}(\Delta y_1\Delta y_2)(\Delta y_1\Delta y_2 + 2v) \\
\mathrm{Vol}(T_{II}) &= \tfrac{1}{4}(\Delta y_1\Delta y_2)(\Delta y_1\Delta y_2 + 2v + 2w) \\
\mathrm{Vol}(T_{III}) &= \tfrac{1}{4}(\Delta y_1\Delta y_2)\left(\Delta y_1\Delta y_2 + u + 2v + 2w\right).
\end{aligned}$$

So the volume of the hypervolume tent structures is a quarter of the area of the base of the tent structure times the sum of the height of all the corners of the tent structure (respectively u, v, w and $\Delta y_1\Delta y_2 + v + w$).

4.4 Using the HVI-tent structures for the proposed algorithm

For bi-objective optimization we can now use these HVI-tent structures to find a point $\mathbf{x} \in \mathbb{R}^d$ for which the expected hypervolume improvement is maximal. In the proposed algorithm of Section 2.2, the main issue is to find such a point. With the HVI-tent structures, we have reduced this problem to the calculation of the volume of $E_{\mathbf{x}}$ and the volume of the HVI-tent structures T_i corresponding to patches P_i covering the area $E_{\mathbf{x}} \backslash \mathrm{Dom}_r(E_{\mathbf{x}})$.

The calculation of the volume of all the HVI-tent structures can be easily done by starting with the patches P_I, where no points in Y_k are dominated. For such patches, where the upper left and lower right corner is a point in Y_k, i.e. $u = v = w = 0$, the volume is simply $\frac{1}{4}$ times the square of its area (see Figure 4.3). If one of these corner points is not in Y_k, then $v = 0$ or $w = 0$, not both, while $u = 0$. The volumes of the P_I-patches can then be used for the calculation of the volume of tent structures corresponding to P_{II}-patches, where only one point is dominated. The volumes of the tent structures of P_I-patches are now the heights of the different corners of the tent structures of the P_{II}-patches. In this way, the volumes that are calculated are only dependent on the coordinates of the previously evaluated non-dominated points in Y_k and the previously calculated volumes of other HVI-tent structures.

Maximizing the expected hypervolume improvement is now reduced to the maximization of the sum of the volumes of the HVI-tent structures divided by the volume of the area $E_{\mathbf{x}}$. One can now choose an existing single objective optimization algorithm to execute this maximization.

Chapter 5

Discussion

In Section 2.2 we proposed a new algorithm for finding approximations of the Pareto front of a multi-objective non-convex optimization problem where the objective function is Lipschitz continuous and a priori estimates of the Lipschitz constant are available. It exploits the idea of the expected hypervolume improvement introduced by Emmerich et al. [2] in another context. This thesis thereafter focused on finding explicit expressions for this improvement, which has to be maximized as part of the algorithm. We did not address the issue of proving that the proposed algorithm provides an approximation of the Pareto front with increasing accuracy and converges to this Pareto front. A rigorous proof is required, but seems to be quite involved.

In Section 3.3 it is proved that the proposed EI-method will generally choose to evaluate the same points to evaluate as Shubert's Algorithm. However, it will occasionally evaluate a different point. In future work it would be interesting to find out what the effect is of these different choices. In particular it would be interesting to see what it means for the convergence rate of the EI-algorithm.

In this thesis, we have mainly focused on how to use the expected hypervolume improvement to find useful points in which one can evaluate f. However, it can occur that there is a whole level set on which the expected hypervolume improvement is minimal instead of a single point (see Section 3.4). Then one still has to make a smart choice on where to evaluate the function. At the moment it is not known what will be the effect of particular choices. Therefore, examining and finding ways to make this smart choice could be very useful to help improve the speed of a new algorithm.

Moreover, in this work, we have found explicit ways to calculate the expected hypervolume improvement of a point for single objective and bi-objective optimization. In future work, we recommend to investigate how this can be extended to multi-objective optimization, i.e. for any $n \in \mathbb{N}_{>0}$. However, the method that was used here for bi-objective optimization will already become computationally much harder, because it will become much harder to calculate the hypervolume improvement of a point in more dimensions. Instead of dividing the area in $\frac{1}{2}(m+1)(m+2)$ HVI-tent structures, you need to divide the area in many more boxes. For $n > 3$ this will grow very fast. Thus, calculating the expected hypervolume improvement of a point will become computationally much more expensive.

In the proposed algorithm, to reduce the amount of function evaluations that has to be done, a point in chosen for which the expected hypervolume improvement is maximal. Therefore it is required that finding this point is easier than one function evaluation of the objective function. Hence, if calculating a maximal value and its location for the expected hypervolume improvement of a point will become computationally hard in multiple dimensions, the proposed algorithm will become too slow to be useful.

Chapter 6

Conclusion

In this thesis we have formulated a way to use the expected hypervolume improvement to find useful evaluation points to minimize an expensive black box function that is Lipschitz continuous. Subsequently, we took a closer look at how this method would work for single objective optimization in one dimension. We found that this method induced an algorithm that works similiar to Shubert's Algorithm but will occasionally choose a different point to evaluate.

Furthermore, a new approach was used to prove that Shubert's Algorithm converges the slowest for f a constant function. This yielded a convergence rate of order $O(\frac{1}{k})$.

Moreover, we took a closer look at the case for bi-objective optimization. For this case, we formulated a way to calculate the expected hypervolume improvement using HVI-tent structures. This resulted in an explicit expression for the expected hypervolume improvement that can be computed iteratively, and then maximized using existing techniques. This computation can unfortunately become expensive in higher dimensions ('curse of dimensionality').

Appendix A

Some calculations for the hypervolume improvement

In this appendix we present some calculation rules for the hypervolume improvement of a point or a set of points with regard to a set of function values that were derived during this thesis project. Although they were not used in the main part of this thesis, we discuss them here, since we believe them to be of separate interest. Rigorous proves have been omitted for this reason.

Throughout this appendix we fix a referent point $\mathbf{r} \in \mathbb{R}^n$. Suppose that the new point $\mathbf{y}$ dominates one or more points in Y_k. Define $Y_{\mathbf{y}}^- := \{\mathbf{z} \in Y_k \mid \mathbf{y} \prec \mathbf{z}\}$, the set of evaluated points that are dominated by $\mathbf{y}$, $Y_{\mathbf{y}}^0 := Y_k \backslash Y_{\mathbf{y}}^-$, the set of evaluated points that are not dominated by $\mathbf{y}$, and $Y_{\mathbf{y}}^+ := Y_{\mathbf{y}}^0 \cup \mathbf{y}$.

Suppose $Y_{\mathbf{y}}^- = \{\mathbf{y}^{(i)}\}$ for some $i \in \{1, \ldots, k\}$. We have:

$$HVI(\mathbf{y} \mid Y_k) = HVI(\mathbf{y} \mid Y_{\mathbf{y}}^0) - HVI(\mathbf{y}^{(i)} \mid Y_{\mathbf{y}}^0).$$

In Figure A.1 an example is given for the case $n = 2$. Here $Y_k = \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4\}$, $Y_{\mathbf{y}}^- = \{\mathbf{z}_2, \mathbf{z}_3\}$ and $Y_{\mathbf{y}}^0 = \{\mathbf{z}_1, \mathbf{z}_4\}$. As can be seen in the figure, the hypervolume improvement of $\mathbf{y}$ (the yellow area) is the hypervolume improvement of $\mathbf{y}$ in respect to the set $\{\mathbf{z}_1, \mathbf{z_4}\}$ (the rectangle with vertices $\mathbf{q}$ and $\mathbf{y}$) minus the hypervolume improvement $\{\mathbf{z}_2, \mathbf{z_3}\}$ in respect to the set $\{\mathbf{z}_1, \mathbf{z_4}\}$.

Let S be a set of points, not in $\mathrm{Dom}_{\mathbf{r}}(Y_k)$ and such that none of its elements dominates another element in S. Then if $\mathbf{z}$ is not in $\mathrm{Dom}_{\mathbf{r}}(Y_k \cup S)$ and does not dominate any element in S, it follows that

$$HVI(S \cup \{\mathbf{z}\} \mid Y_k) = HVI(\mathbf{z} \mid Y_k) + HVI(S \mid Y_k \cup \{\mathbf{z}\}).$$

Then if we want to calculate the hypervolume improvement of m points with respect to a set Y_k, where the m points do not dominate one of the points in

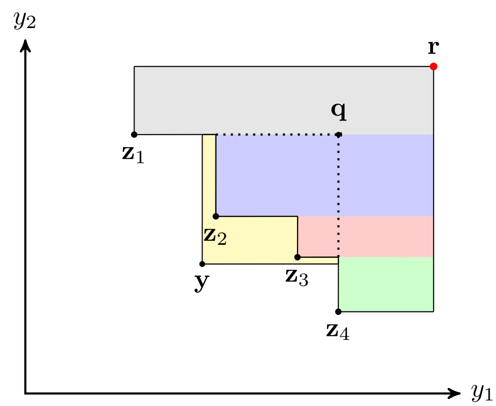

Figure A.1: *The hypervolume improvement of* $\mathbf{y}$ *with respect to a set* Y_k *where it dominates* $\mathbf{z}_2$ *and* $\mathbf{z}_3$ *and reference point* $\mathbf{r}$.

Y_k, we find:

$$\begin{aligned} HVI(\{\mathbf{z}_1,\ldots,\mathbf{z}_m\} \mid Y_k) &= HVI(\{\mathbf{z}_1\} \mid Y_k) + HVI(\{\mathbf{z}_2\} \mid Y_k \cup \{\mathbf{z}_1\}) + \\ &\quad \cdots + HVI(\{\mathbf{z}_m\} \mid Y_k \cup \{\mathbf{z}_1,\ldots,\mathbf{z}_{m-1}\}) \\ &= \sum_{i=1}^{m} HVI(\{\mathbf{z}_i\} \mid Y_k \cup \{\mathbf{z}_1,\ldots,\mathbf{z}_{i-1}\}). \end{aligned}$$

In Figure A.2 an example of the hypervolume improvement of $\mathbf{z}_2$ and $\mathbf{z}_3$ with respect to $\mathbf{z}_1$ is displayed. So here $HVI(\{\mathbf{z}_2,\mathbf{z}_3\} \mid \{\mathbf{z}_1\})$ is the sum of the hypervolume improvement of $\mathbf{z}_2$ with respect to $\mathbf{z}_1$ (the blue area) and the hypervolume improvement of $\mathbf{z}_3$ with respect to the set $\{\mathbf{z}_1,\mathbf{z}_2\}$ (the red area). It is easy to see that this also works if we do not look at the hypervolume improvement of two points with respect to one point, but the hypervolume improvement of m points with respect to a set of k points for any $m,k \in \mathbb{Z}_{>0}$, which confirms the formula we found in the equation above.

Suppose $Y_{\mathbf{y}}^{-} = \{\mathbf{z}_1,\ldots,\mathbf{z}_m\}$. Then we have

$$HVI(\mathbf{y} \mid Y_k) = HVI(\mathbf{y} \mid Y_{\mathbf{y}}^{0}) - HVI(Y_{\mathbf{y}}^{-} \mid Y_{\mathbf{y}}^{0}).$$

Because Y_k (hence $Y_{\mathbf{y}}^{-}$) is assumed to be reduced for dominance, we find

$$HVI(\mathbf{y} \mid Y_k) = HVI(\mathbf{y} \mid Y_{\mathbf{y}}^{0}) - \sum_{i=1}^{m} HVI(\mathbf{z}_i \mid Y_{\mathbf{y}}^{0} \cup \{\mathbf{z}_1,\ldots,\mathbf{z}_{i-1}\}).$$

So we have reduced the problem of calculating $HVI(\mathbf{y} \mid Y_k)$ to the calculation of a sum of terms of the shape $HVI(\mathbf{z} \mid A)$, where $\mathbf{z}$ is a point that does not dominate any point in the set A. In Wong et al. [8] an operator $hv(\cdot)$ is used to

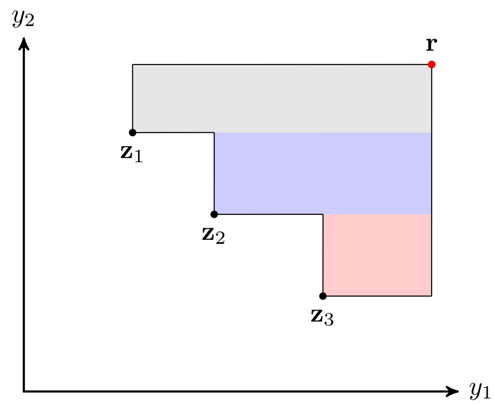

Figure A.2: *The hypervolume improvement of* $\mathbf{z}_2$ *and* $\mathbf{z}_3$ *with respect to* $\mathbf{z}_1$ *and reference point* $\mathbf{r}$

calculate the hypervolume contribution of every point in a set. So the problem of calculating $HVI(\mathbf{y} \mid Y_k)$ is reduced to the calculation of $hv(A \cup \mathbf{z})$ for $m+1$ different z and A, where $m = |Y_{\mathbf{y}}^{-}|$. A similar approach is used in Emmerich and Fonseca [4] to prove a lower bound for the time complexity of calculating the hypervolume improvement.

Appendix B

Lipschitz estimates for the hypervolume improvement

In this section, we are going to present Lipschitz bounds for the differences in hypervolume between two sets of evaluated points. Our focus is on sets with the same cardinality or those that differ by a single point. This results came out of initial investigations in this master project. Then we considered another concept of domination:

$$\mathbf{s} = (s_1, \ldots, s_n) \preceq \mathbf{x} = (x_1, \ldots, x_n) \Leftrightarrow s_i \leqslant x_i \text{ for all } i \in \{1, \ldots, n\},$$

related to maximization of objectives. Here we assume we want to maximize $f : \mathbb{R}^d \to [0, M]^n$. So we take $\mathbf{r} = \mathbf{0}$ as reference point. Note that for minimization we would get the same results. We write $H(S)$ for the amount of hypervolume dominated by a set S.

B.1 Difference in dominated hypervolume between two sets with equal cardinality

Let $S = \{\mathbf{s}^1, \ldots, \mathbf{s}^k\}$ be a set of points in the objective space. Now consider an other set $S' = \{\mathbf{t}^1, \ldots, \mathbf{t}^k\}$ of points in the objective space with $|S| = |S'| = k$. It follows that there is a bijection $\phi : S \to S'$. Define $D_{\mathbf{s}} = \{\mathbf{x} \in \mathbb{R}^n_+ : \mathbf{s} \preceq \mathbf{x}\}$. Then

$$H(S) = \int_{\mathbb{R}^n_+} \bigvee_{\mathbf{s} \in S} \mathbb{1}_{D_{\mathbf{s}}}(\mathbf{x}) d\mathbf{x}$$

(where $\bigvee_{\mathbf{s} \in S} \mathbb{1}_{D_{\mathbf{s}}} := \max_{\mathbf{s} \in S} \mathbb{1}_{D_s}$). To make it easier to read, we write $h_{\mathbf{s}} := \mathbb{1}_{D_{\mathbf{s}}}$. Moreover, write $\|\mathbf{z} - \mathbf{y}\|_\infty := \min_{1 \leqslant j \leqslant n} |z_j - y_j|$. Define $\alpha = (\alpha_j) \in \{0, 1\}^n$ as $\alpha_j = 1$ if $\min\{y_j, z_j\} = z_j$, and $\alpha_j = 0$ otherwise. Write $|\alpha| = \sum_{j=1}^n \alpha_j$.

Lemma B.1.1. *Let* $\mathbf{y}, \mathbf{z} \in [0, M]^n$. *Then*

$$\int_{\mathbb{R}^n_+} |h_{\mathbf{y}}(\mathbf{x}) - h_{\mathbf{z}}(\mathbf{x})| d\mathbf{x} \leqslant \max\left\{2^{|\alpha|-1}, 2^{n-|\alpha|-1}\right\} M^{n-1} \|\mathbf{z} - \mathbf{y}\|. \tag{B.1}$$

Proof. We have

$$\begin{aligned}\int_{\mathbb{R}^n_+} |h_{\mathbf{y}}(\mathbf{x}) - h_{\mathbf{z}}(\mathbf{x})| \, d\mathbf{x} &= \int_{\mathbb{R}^n_+} \left|\mathbb{1}_{D_{\mathbf{y}} \triangle D_{\mathbf{z}}}\right| d\mathbf{x} \\ &= \int_{\mathbb{R}^n_+} \mathbb{1}_{D_{\mathbf{z}} \setminus (D_{\mathbf{y}} \cap D_{\mathbf{z}})} + \mathbb{1}_{D_{\mathbf{y}} \setminus (D_{\mathbf{y}} \cap D_{\mathbf{z}})} d\mathbf{x}.\end{aligned}$$

Define $\mathbf{z'} \in [0, M]^n$ by $z'_j = \min\{y_j, z_j\}$. We get that

$$\begin{aligned}\int_{\mathbb{R}^n_+} |h_{\mathbf{y}}(\mathbf{x}) - h_{\mathbf{z}}(\mathbf{x})| d\mathbf{x} &= \int_{\mathbb{R}^n_+} h_{\mathbf{z}} - 2 \cdot h_{\mathbf{z'}} + h_{\mathbf{y}} d\mathbf{x} \\ &= \prod_{j=1}^n z_j - 2 \prod_{j=1}^n z'_j + \prod_{j=1}^n y_j \\ &= \prod_{j=1}^n z_j - 2 \prod_{j=1}^n z_j^{\alpha_j} y_j^{1-\alpha_j} + \prod_{j=1}^n y_j \\ &= \prod_{j=1}^n z_j - 2 \prod_{j:\alpha_j=0} y_j \cdot \prod_{j:\alpha_j=1} z_j + \prod_{j=1}^n y_j \\ &= \left(\prod_{j:\alpha_j=1} z_j \cdot \prod_{j:\alpha_j=0} (z_j - y_j)\right) + \left(\prod_{j:\alpha_j=0} y_j \cdot \prod_{j:\alpha=1} (y_j - z_j)\right) \\ &\leqslant M^{|\alpha|} \prod_{j:\alpha=0} |z_j - y_j| + M^{n-|\alpha|} \prod_{j:\alpha=1} |z_j - y_j|. \end{aligned} \tag{B.2}$$

If $|\alpha| \geqslant 1$, we have

$$\begin{aligned}(B.2) &\leqslant M^{|\alpha|} \|\mathbf{z} - \mathbf{y}\|_\infty (2M)^{n-|\alpha|-1} + M^{n-|\alpha|} \|\mathbf{z} - \mathbf{y}\|_\infty (2M)^{|\alpha|-1} \\ &\leqslant \max\left(2^{|\alpha|-1}, 2^{n-|\alpha|-1}\right) M^{n-1} \|\mathbf{z} - \mathbf{y}\|_\infty.\end{aligned}$$

If $|\alpha| = 0$, the same follows immediately from (B.2). Hence we find

$$|H(S) - H(S')| \leqslant C \cdot \max\left(2^{|\alpha|-1}, 2^{n-|\alpha|-1}\right) M^{n-1} \|\mathbf{z} - \mathbf{y}\|_1$$

for some $C \in \mathbb{R}_{\geqslant 0}$. □

Lemma B.1.2. *There exists a constant* $C \in \mathbb{R}_{\geqslant 0}$ *such that for all* $k \in \mathbb{N}$ *and* $S, S' \in [0, M]^n$ *with* $|S| = |S'| = k < \infty$:

$$|H(S) - H(S')| \leqslant C \cdot M^{n-1} \min_{\phi:\phi(S)=S'} \sum_{\mathbf{s} \in S, \mathbf{s'} \in S'} \|\mathbf{s} - \mathbf{s'}\|_1. \tag{B.3}$$

Proof. We have

$$\begin{aligned}\left|H(S) - H(S')\right| &= \left|\int_{\mathbb{R}^n_+} \bigvee_{\mathbf{s}\in S} \mathbb{1}_{D_{\mathbf{s}}}(x)dx - \int_{\mathbb{R}^n_+} \bigvee_{\mathbf{s}'\in S'} \mathbb{1}_{D_{\mathbf{s}'}}(x)dx\right| \\ &\leqslant \int_{\mathbb{R}^n_+} \left|\bigvee_{\mathbf{s}\in S} \mathbb{1}_{D_{\mathbf{s}}}(x) - \bigvee_{\mathbf{s}'\in S} \mathbb{1}_{D_{\mathbf{s}'}}(x)\right| dx.\end{aligned}$$

In the next calculation we use Birkhoff's Inequality:

$$|a \vee b - a \vee c| \leqslant |b - c| \text{ for any } a, b, c \in \mathbb{R}. \tag{B.4}$$

We get

$$\begin{aligned}\left|H(S) - H(S')\right| &\leqslant \int_{\mathbb{R}^n_+} \left|\bigvee_{\mathbf{s}\in S} h_{\mathbf{s}}(x) - \bigvee_{\mathbf{s}'\in S'} h_{\mathbf{s}'}(x)\right| dx \\ &= \int_{\mathbb{R}^n_+} \left|\bigvee_{\mathbf{s}\in S} h_{\mathbf{s}}(x) - \bigvee_{\mathbf{s}\in S} h_{\phi(\mathbf{s})}(x)\right| dx \\ &\overset{\triangle-\text{ineq}}{\leqslant} \int_{\mathbb{R}^n_+} \left|h_{\mathbf{s}^1}(x) \vee \bigvee_{i=2}^{k} h_{\mathbf{s}^i}(x) - h_{\mathbf{s}^1}(x) \vee \bigvee_{i=2}^{k} h_{\phi(\mathbf{s}^i)}(x)\right| \\ &\quad + \left|h_{\mathbf{s}^1}(x) \vee \bigvee_{i=2}^{k} h_{\phi(\mathbf{s}^i)}(x) - h_{\phi(\mathbf{s}^1)}(x) \vee \bigvee_{i=2}^{k} h_{\phi(\mathbf{s}^i)}(x)\right| dx \\ &\overset{(B.4)}{\leqslant} \int_{\mathbb{R}^n_+} \left|\bigvee_{i=2}^{k} h_{\mathbf{s}^i}(x) - \bigvee_{i=2}^{k} h_{\phi(\mathbf{s}^i)}(x)\right| + \left|h_{\mathbf{s}^1}(x) - h_{\phi(\mathbf{s}^1)}(x)\right| dx.\end{aligned}$$

By repeating the last two steps, the triangle inequality and Birkhoff's Inequality, we find the following:

$$\left|H(S) - H(S')\right| \leqslant \int_{\mathbb{R}^n_+} \sum_{i=1}^{k} \left|h_{\mathbf{s}^i}(x) - h_{\phi(\mathbf{s}^i)}(x)\right| dx.$$

Now apply Lemma B.1.1:

$$\begin{aligned}\left|H(S) - H(S')\right| &\leqslant \min_{\phi:\phi(S)=S'} 2^{n-1} M^{n-1} \sum_{\mathbf{s}\in S} \|\mathbf{s} - \phi(\mathbf{s})\|_\infty \\ &\leqslant C \cdot M^{n-1} \min_{\phi:\phi(S)=S'} \sum_{\mathbf{s}\in S} \|\mathbf{s} - \phi(\mathbf{s})\|_1\end{aligned}$$

which yields the required result. □

B.2 Difference in dominated hypervolume between a set and the same set with one more element

Now we consider the set S'' with $S'' = S \cup \{\mathbf{x}^*\}$ for some $\mathbf{x}^* \in \mathbb{R}^d$. We want to prove that

$$|H(S) - H(S'')| \leqslant M^{d-1} C d_H(S, S'')$$

for a good (small) C, where d_H is the Hausdorff distance:

$$d_H(S, S'') = \max(\sup_{x \in S} \inf_{y \in S''} d(x, y), \sup_{x \in S''} \inf_{y \in S} d(x, y)).$$

Let $\mathbf{s}^0 \in S$, then we have

$$\begin{aligned} |H(S) - H(S'')| &= \left| \int_{\mathbb{R}^n_+} \bigvee_{\mathbf{t} \in S''} h_{\mathbf{t}}(x) d\lambda - \int_{\mathbb{R}^n_+} \bigvee_{\mathbf{s} \in S} h_{\mathbf{s}}(x) d\lambda \right| \\ &\leqslant \int_{\mathbb{R}^n_+} \left| h_{\mathbf{x}^*} \vee \bigvee_{\mathbf{s} \in S} h_{\mathbf{s}} - \bigvee_{\mathbf{s} \in S} h_{\mathbf{s}} \right| d\lambda \\ &= \int_{\mathbb{R}^n_+} \left| h_{\mathbf{x}^*} \vee \bigvee_{\mathbf{s} \in S} h_{\mathbf{s}} - h_{\mathbf{s}^0} \vee \bigvee_{\mathbf{s} \in S} h_{\mathbf{s}} \right| d\lambda \\ &\leqslant \int_{\mathbb{R}^n_+} |h_{\mathbf{x}^*} - h_{\mathbf{s}^0}| \, d\lambda. \end{aligned}$$

Since this holds for every $\mathbf{s}^0 \in S$, we find

$$|H(S) - H(S'')| \leqslant \min_{\mathbf{s}^0 \in S} \int_{\mathbb{R}^n_+} |h_{\mathbf{x}^*} - h_{\mathbf{s}^0}| \, d\lambda \leqslant \min_{\mathbf{s}^0 \in S} C_S \|\mathbf{x}^* - \mathbf{s}^0\|_1.$$

So we have the following inequality.

$$|H(S) - H(S \cup \{\mathbf{x}^*\})| \leqslant C \cdot d(\mathbf{x}^*, S).$$

Now we found an inequality of the form we wanted, but this C is in most cases probably not as small as we would like.

Bibliography

[1] Yves Brise. *Lipschitzian Optimization, DIRECT Algorithm, and Applications.* pages 1–46, 2008. URL http://people.inf.ethz.ch/ybrise/data/talks/msem20080401.pdf.

[2] Michael Emmerich, Nicola Beume, and Boris Naujoks. *An EMO Algorithm Using the Hypervolume Measure As Selection Criterion*, pages 62–76. EMO'05. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 3-540-24983-4, 978-3-540-24983-2. URL http://dx.doi.org/10.1007/978-3-540-31880-4_5.

[3] Michael Emmerich, Kaifeng Yang, André Deutz, Hao Wang, and Carlos M Fonseca. *A Multicriteria Generalization of Bayesian Global Optimization*, pages 229–242. Springer International Publishing, 2016.

[4] Michael T. M. Emmerich and Carlos M. Fonseca. Computing hypervolume contributions in low dimensions: Asymptotically optimal algorithm and complexity results. In *Proceedings of the 6th International Conference on Evolutionary Multi-criterion Optimization*, EMO'11, pages 121–135, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-19892-2. URL http://dl.acm.org/citation.cfm?id=1987637.1987647.

[5] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998. ISSN 1573-2916. URL http://dx.doi.org/10.1023/A:1008306431147.

[6] Erich Novak and Klaus Ritter. *Global Optimization Using Hyperbolic Cross Points*, pages 19–33. Springer US, Boston, MA, 1996. ISBN 978-1-4613-3437-8. URL http://dx.doi.org/10.1007/978-1-4613-3437-8_2.

[7] Bruno O. Shubert. A sequential method seeking the global maximum of a function. *SIAM Journal on Numerical Analysis*, 9(3):379–388, 1972. URL https://doi.org/10.1137/0709036.

[8] Cheryl Sze Yin Wong, Abdullah Al-Dujaili, and Suresh Sundaram. *Hypervolume-Based DIRECT for Multi-Objective Optimisation*, pages 1201–1208. GECCO '16 Companion. ACM,

New York, NY, USA, 2016. ISBN 978-1-4503-4323-7. URL `http://doi.acm.org/10.1145/2908961.2931702`.

[9] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. G. da Fonseca. Performance assessment of multiobjective optimizers: An analysis and review. *Trans. Evol. Comp*, 7(2):117–132, April 2003. ISSN 1089-778X. URL `http://dx.doi.org/10.1109/TEVC.2003.810758`.

[10] Antanas Žilinskas and Julius Žilinskas. Adaptation of a one-step worst-case optimal univariate algorithm of bi-objective lipschitz optimization to multidimensional problems. *Communications in Nonlinear Science and Numerical Simulation*, 21(1–3):89 – 98, 2015. ISSN 1007-5704. URL `http://doi.org/10.1016/j.cnsns.2014.08.025`. Numerical Computations: Theory and Algorithms (NUMTA 2013), International Conference and Summer School.