Microsoft Research Technical Report MSR-TR-2011-42

# Activity Report for 2009-2011

## Microsoft-CNRS Chair
## "Optimization for Sustainable Development"

*Scientific co-directors*:

- Youssef Hamadi: *Microsoft Research Cambridge, UK*
  Email:youssefh@microsoft.com
- Leo Liberti: *LIX, École Polytechnique, F-91128 Palaiseau, France*
  Email:liberti@lix.polytechnique.fr

March 31, 2011


http://chaire-osd.polytechnique.fr

*This report details the first two years of activity of the Microsoft-CNRS OSD Chair at Ecole Polytechnique. The main objective for this first period was to identify promising research themes where optimization techniques were applied to real-world applications in sustainable development; two promising application fields were identified: transportation and energy. Another important goal, that of making the Chair known to the scientific community, was fulfilled by ranking first (over more than fifty participating research teams from all around the world) in the prestigious "EURO/ROADEF Challenge 2010" sponsored by* Electricité de France.

# Contents

# Chapter 1

# Introduction

*This chapter gives a general overview of the OSD chair, its history, its scientific achievements, and the roadmap for the next three years.*

## 1.1 Chairs at Ecole Polytechnique

Ecole Polytechnique (X) hosts a certain number of research/teaching chairs[1] in its teaching departments and research laboratories, called X-chairs in the rest of this document. As a general guideline, an X-chair is a financial endowments over a certain number of years (usually 5), dedicated to support research and teaching on a specific scientific topic, under the joint responsibility of a senior researcher (the chair's director) possessing an X affiliation, its research lab, and X. These endowments are usually supported by industry and other institutions. They are not to be seen for a provision of services in exchange of money, but rather as a form of patronage. "Teaching" usually means the organization of a new master course or the support of an existing master course. "Research" usually means the recruitment of nonpermanent researchers (typically PhDs and postdoctoral fellows, but also visiting staff permanently affiliated to other institutions) to pursue fundamental research. The "scientific topic" is decided jointly by the responsible researcher and the industrial sponsors. To date, there are 19 X-chairs[2], five of which list the word "sustainable" in their name.

## 1.2 The OSD chair

The Microsoft-CNRS chair for Optimization for Sustainable Development (OSD) is an X-chair based on a contract involving three partners: Microsoft (MS), CNRS and X. MS actually involves two separate branches: Microsoft Research (MSR) and Microsoft France (MSF). The level of the scientific contributions is guaranteed by two scientific co-directors (L. Liberti and Y. Hamadi) and a scientific councillor (D. Krob) which is part of the steering committee.

[1] `http://www.polytechnique.edu/accueil/entreprises/chaires-enseignement-recherche/`.

[2] In economics: Insurance and Major Risks; Business Economics; Sustainable Development; Sustainable Finance and Responsible Investment; Health, Risk, Insurance. In computer science: Engineering of Complex Systems; *Optimization for Sustainable Development*. In applied mathematics: Derivatives of the Future; Finance and Sustainable Development: Quantitative Aspects; Mathematical Modeling and Biodiversity; Mathematical Modeling and Digital Simulation; Financial Risks. In physics and mechanics: Sustainable Energies; Nanoscience and Nanotechnology; Science of Materials for Sustainable Construction; Science of Materials and Active Surfaces. In humanities and social sciences: Innovation and Regulation of Digital Services; Innovation Management; Multicultural Management and Business Performance.

The OSD chair is hosted by the *Laboratory* (LIX) and the *Department* (DIX) of computer science at X. LIX has a hierarchical structure organized in several research teams: the OSD chair is part of the *System Modelling and Optimization* (SYSMO) team, headed by Leo Liberti. As such, the OSD chair has privileged links with the SYSMO team, and the members of SYSMO often contribute research time to the chair's topics.

### 1.2.1 Governance

The chair's governance is ensured by the director and a steering committee where all the participating institutions are represented. The steering committee convenes at least once a year. Currently, this committee includes: Leo Liberti (LIX), Youssef Hamadi (MSR), Pierre-Louis Xech (MSF), Philippe Guédon (DRIP-X), Daniel Krob (LIX), Florence Sedes (CNRS).

### 1.2.2 History

The OSD chair was conceived by Youssef Hamadi and the original director (Philippe Baptiste), with the active efforts of Pierre-Louis Xech at Microsoft France. The contract, initially for two years, was signed on 30 march 2009. Two recruitments were made during the first year of existence of the OSD chair, under the responsibility of Ph. Baptiste: Dr. Shmelev and Dr. Touati. The former left after only a couple of months of employments, wheras the latter has been with the chair for two years.

In February 2010, Ph. Baptiste accepted to be the director of the INS2I institute within the CNRS, and therefore had to quit the directorship of the OSD chair, leaving this responsibility to Leo Liberti. Almost immediately, the chair started a politic of candidate searchq, which eventually led to recruit three postdoctoral fellows (A. Fialho, D. Brockhoff and the renewal of N. Touati's contract) and a Ph.D. student (F. Roda, co-sponsored by another X-chair).

In July 2010, a team formed by Vincent Jost, David Savourey (researchers in the SYSMO team who often work on OSD topics) and others managed to win the first prize of the prestigious EURO/ROADEF yearly challenge, out of around fifty participating teams all around the world. This marvellous feat has been instrumental in expanding the awareness of the OSD chair within the optimization community.

### 1.2.3 Mission

The mission of the OSD chair is to identify and perform research and teaching actions on topics concerning the application of optimization techniques to real-world problems and issues involving sustainable development.

### 1.2.4 Past goals and objectives

The initial goals of the OSD chair were essentially research related:

- produce and divulge knowledge related to sustainable development;
- work with real-world actors to keep research from becoming too abstract;
- identify, model and solve important problems in the domain of sustainable development;
- develop theoretical knowledge in some specific optimization related fields.

After the 2010 steering committee meeting, it was decided to also pursue a teaching action, i.e. supporting the new *Master Parisien en Recherche Operationnelle* (MPRO).

The OSD chair employs the following type of actions in order to pursue its goals:

- recruitment of nonpermanent staff for research and teaching purposes;
- funding participation of the chair members to conferences, workshops and meetings on relevant topics;
- organization of scientific events (e.g. seminars and workshops) on relevant topics;
- co-funding the organization of international conferences on relevant topics;
- funding of fundamental and applied research (pursued by agents external to the chair itself) within the scope of the chair's scientific mission;
- acquisition of material necessary to pursue research (e.g. computers, books, software).

The OSD chair is up for renewal on 30th march 2010 (or as soon as possible after this date). After two years of work, all the objectives listed in Sect. 1.2.3 have been partially attained: we are currently starting to publish papers in optimization topics applied to sustainable development problems, we have been establishing contacts with several industrial partners working in sustainable development (such as EDF, ALSTOM Power and others), we have identified two very promising research themes (i.e. energy and transportation), and we have been making important contributions in optimization theory and methodology. To validate all this, the chair was able to secure a comfortable success status by winning the EURO/ROADEF challenge in 2010.

### 1.2.5 Scientific achievements

The detailed account of all the scientific output of the first two years of existence is detailed in the main body of this report. Here follows a résumé.

- **Sustainable energy**
  1. *The EURO/ROADEF Challenge*. Our most prestigious scientific achievement is the first prize in this industrially sponsored computational competition open to all research teams worldwide, and supported by the European and French Operations Research Societies. The OSD team ranked 1st (in the senior category) over around 50 participants. The competition concerned the planning of outages of nuclear power plants at EDF.
  2. *Survey on optimization and Smart Grids*. There is a lot of talk about Smart (electrical) Grids, and several informal definitions as to what impediments a smart grid should be able to cope with; the literature fails to some extent to provide a formal definition concerning what a smart grid should be able to actually *do*. Moreover, from the point of view of optimization techniques, no satisfactory survey of the state of the art was written yet. This survey is preparatory for working with prospective sustainable energy partners to the OSD chair.
  3. *Combinatorial optimization for electric vehicles management*. Growing concerns about environmental quality of cities are calling for sustainable road transportation technologies. Electric Vehicles (EV), for public and private transport, can contribute significantly to the lowering of the current pollution levels. However, the EV use is currently facing several weaknesses among which are: limited driving range, high cost and overall limited efficiency. This report aims at specifying some key contributions of combinatorial optimization for an efficient electric vehicles management.

4. *Optimal running of biomass production*. This work actually pre-dates the chair by one year, but is nonetheless interesting insofar as: (a) it was mostly carried out by the current scientific Polytechnique co-director; (b) it is gaining visibility through citations. It concerns the optimized production of energy via a set of facilities for transforming biomasses into energy.
5. *Optimizing buildings towards higher energy efficiency*. We aim at contributing to the issue of energy consumption by proposing tools to automatically define some aspects of the architectural and structural design of buildings. Our architecture starts with a building design, and automatically generates and searches a space of acceptable design variations. It outputs a variation which maximizes energy efficiency, and respect cost constraints. The optimization stage is done by the combination of an energy consumption simulation program, EnergyPlus [336], with a state-of-the-art multi-objective evolutionary algorithm.

- **Sustainable transportation**

1. *On green routing and scheduling problem*. The vehicle routing and scheduling problem has been studied with much interest within the last four decades. In this report, some of the existing literature dealing with routing and scheduling problems with environmental issues is reviewed, and a description is provided of the problems that have been investigated and how they are treated using combinatorial optimization tools.
2. *The price of equity in the Hazmat Transportation Problem*. "Equity" is not an easy term to define. In the context of transportation of hazardous materials, different definitions of equity lead to different situations as concerns transportation costs and the risk connected to loss of (quality of) life following catastrophic accidents. Specifically, regional fairness costs lives with respect to individual fairness. We discuss a methodology for trying to peg a monetary cost to this trade-off.
3. *Routing of hazardous materials*. The routing of vehicles represents an important component of many distribution and transportation systems and has been intensively studied in the operations research literature. In this report, particular consideration is given to routing models for the transportation of hazardous materials. This problem has received a large interest in recent years, this results from the increase in public awareness of the dangers of hazardous materials and the enormous amount of hazardous materials being transported. We describe here some major differences between routing of hazardous materials and the classical vehicle routing problems. We review some general models and optimization techniques and propose several direction for future research.
4. *Evolutionary optimization for the multiobjective risk-equity constrained routing problem*. We propose a methodology to find a routing plan for transporting hazardous materials that attempts to balance costs with safety and green requirements.
5. *A branch-and-price algorithm for the risk-equity constrained routing problem*. We study a multicriteria variant of the problem of routing hazardous material on a geographical area subdivided in regions. The two objective functions are given by a generally defined routing cost and a *risk equity* equal to the maximum, over each region, of the risk *perceived* within a region. This is a multicommodity flow problem where integer variables are used to define the number of trucks used for the routing. This problem admits a straightforward path formulation, and we propose a branch-and-price problem where, for each node of the branch-and-bound tree, column generation is used to obtain a lower bound. Experimental results on a set of instances are reported.

- **Innovative methods for optimization**

1. *Multi-Objective Differential Evolution with Adaptive Control of Parameters and Operators*. Differential Evolution (DE) is a simple yet powerful evolutionary algorithm, whose performance highly depends on the setting of some parameters. In this report, we propose an adaptive DE algorithm for multi-objective optimization problems. Firstly, a novel tree neighborhood density estimator is proposed to enforce a higher spread between the non-dominated solutions, while the Pareto dominance strength is used to promote a higher convergence to the

Pareto front. These two metrics are then used by an original replacement mechanism based on a three-step comparison procedure; and also to port two existing adaptive mechanisms to the multi-objective domain, one being used for the autonomous selection of the operators, and the other for the adaptive control of DE parameters CR and F. Experimental results confirm the superior performance of the proposed algorithm, referred to as Adap-MODE, when compared to two state-of-the-art baseline approaches, and to its static and partially-adaptive variants.

2. *Mirrored Sampling and Sequential Selection in Evolution Strategies*. This report presents a refined single parent evolution strategy that is derandomized with mirrored sampling and/or uses sequential selection. The paper analyzes some of the elitist variants of this algorithm. We prove, on spherical functions with finite dimension, linear convergence of different strategies with scale-invariant step-size and provide expressions for the convergence rates as the expectation of some known random variables. In addition, we derive explicit asymptotic formulae for the convergence rate when the dimension of the search space goes to infinity. Convergence rates on the sphere reveal lower bounds for the convergence rate of the respective step-size adaptive strategies. We prove the surprising result that the (1+2)-ES with mirrored sampling converges at the same rate as the $(1 + 1)$-ES without and show that the tight lower bound for the $(1 + \lambda)$-ES with mirrored sampling and sequential selection improves by 16% over the $(1 + 1)$-ES reaching an asymptotic value of about $-0.235$.

3. *Mirrored Sampling in Evolution Strategies With Weighted Recombination*. This report introduces mirrored sampling into evolution strategies with weighted multi-recombination. *Pairwise selection* selects at most one of two mirrored vectors in order to avoid a bias due to recombination. *Selective mirroring* only mirrors the originally worst solutions of the population. Convergence rates on the sphere function are derived also yielding lower bounds. The optimal ratio of mirrored offspring is 1/2 (maximal) for randomly selected mirrors and about 1/6 for selective mirroring, where the convergence rate reaches a value of 0.390. This is an improvement of more than 50% compared to the best known convergence rate of 0.25 with positive recombination weights. Selective mirroring is combined with CMA-ES and benchmarked on unimodal functions and on the COCO/BBOB-2010 testbed.

4. *Optimal $\mu$-Distributions for Linear Fronts*. To simultaneously optimize multiple objective functions, several evolutionary multiobjective optimization (EMO) algorithms have been proposed. Nowadays, often set quality indicators are used when comparing the performance of those algorithms or when selecting "good" solutions during the algorithm run. Hence, characterizing the solution sets that maximize a certain indicator is crucial—complying with the optimization goal of many indicator-based EMO algorithms. If these optimal solution sets are upper bounded in size, e.g., by the population size $\mu$, we call them *optimal $\mu$-distributions*. Recently, optimal $\mu$-distributions for the well-known hypervolume indicator have been theoretically analyzed, in particular, for bi-objective problems with a linear Pareto front. Although the exact optimal $\mu$-distributions have been characterized in this case, not all possible choices of the hypervolume's reference point have been investigated. Moreover, some of the results rely on a lower bound for the reference point in order to ensure the extremes of the front in the optimal $\mu$-distributions. In this report, we revisit the previous results and rigorously characterize the optimal $\mu$-distributions also for all other reference point choices. In this sense, our characterization is now exhaustive as the result holds for any linear Pareto front and for any choice of the reference point and the optimal $\mu$-distributions turn out to be always unique in those cases. We also prove a tight lower bound (depending on $\mu$) such that choosing the reference point above this bound ensures the extremes of the Pareto front to be always included in optimal $\mu$-distributions.

5. *Hypervolume-based Multiobjective Optimization*. In recent years, indicator-based evolutionary algorithms, allowing to implicitly incorporate user preferences into the search, have become widely used in practice to solve multiobjective optimization problems. When using this type of methods, the optimization goal changes from optimizing a set of objective functions simultaneously to the single-objective optimization goal of finding a set of $\mu$ points that maximizes

the underlying indicator. Understanding the difference between these two optimization goals is fundamental when applying indicator-based algorithms in practice. On the one hand, a characterization of the inherent optimization goal of different indicators allows the user to choose the indicator that meets her preferences. On the other hand, knowledge about the sets of $\mu$ points with optimal indicator values—so-called *optimal $\mu$-distributions*—can be used in performance assessment whenever the indicator is used as a performance criterion. However, theoretical studies on indicator-based optimization are sparse. In previous work, we theoretically investigated the *unweighted* hypervolume indicator in terms of a characterization of optimal $\mu$-distributions and the influence of the hypervolume's reference point for general bi-objective optimization problems. In this report, we generalize those results to the case of the *weighted* hypervolume indicator.

- **Promising ideas**

  1. *Architecture evolutions of Information Systems*. In the normal lifespan of large enterprises, the strategic management of IT often evolves. Existing services must be replaced with new services without impairing operations. The problem of scheduling such replacement is of critical importance for the success of the operation. We analyze this problem from a quantitative point of view, underlining the trade-off nature of its solutions. We formalize this multi-objective optimization problem as a mathematical programming formulation. We discuss its theoretical properties and show that real-world instances can be solved by standard off-the-shelf tools.
  2. *Discretizable Molecular Distance Geometry Problem*. Identifying 3D conformation is a crucial step to synthesizing useful proteins; in particular, the conformation of some of the proteins linked to photosynthesis is still largely unknown. Since photosynthesis allows the production of clean energy from light, this line of research is relevant to sustainable energy, although our approach is more general than that. The Molecular Distance Geometry Problem (MDGP) consists in finding an embedding in $\mathbb{R}^3$ of a nonnegatively weighted simple undirected graph with the property that the Euclidean distances between embedded adjacent vertices must be the same as the corresponding edge weights. The Discretizable Molecular Distance Geometry Problem (DMDGP) is a particular subset of the MDGP which can be solved using a discrete search occurring in continuous space; its main application is to find three-dimensional arrangements of proteins using Nuclear Magnetic Resonance (NMR) data. The model provided by the DMDGP is too theoretical for practical exploitation. In the last five years we strove to adapt the DMDGP to be an ever closer model of the actual difficulties posed by the problem of determining protein structures from NMR data, whilst always keeping the discrete search property valid. This survey lists recent developments on DMDGP related research.
  3. *Sustainable public policies for natural resource access in Senegal*. Territorial laws in Senegal for regulating public access to natural resources are currently being debated. The debate concerns the trade-off between productivity, fairness, equitability and sustainability. This project aims to develop a modelling/simulation tool based on optimization techniques to evaluate the impact of different public policies concerning the assignment of natural resource access to different actors. Theoretically speaking, our proposal is based on a complex assignment problem with several objectives. This report concerns the first phase of the work where data are collected from domain actors (farmers and shepherds on the terrain).

### 1.2.6 Dissemination effort

The effort of the OSD chair members as concerns dissemination and communication during the period 2009-2011 included dedicated workshops, dedicated sessions in national and international conferences, and partial sponsorship of international conferences of high visibility.

- **Workshops**.

1. *Kick-off workshop*. This was organized on 3rd june 2009 at Ecole Polytechnique. Guest introductory presentations by Ms. N. Kosciusko-Morizet (state secretary for digital economy development and now minister for sustainable development), and Mr. Rick Rashid, Senior Vice President, Research, Microsoft. Scientific presentations by P. van Hentenryck (Brown University), L. Doyen (CNRS), C. Gollier (Univ. Toulouse I), M. Metaiche (CNRS), C. Le Pape (Schneider Electric), K. Pruhs (Univ. of Pittsburgh), L. Schmitt (ALSTOM Power). See `http://chaire-osd.polytechnique.fr/accueil/activites/inauguration/` for more details.
2. *OSD day on multi-objective programming*. This was organized on 22nd june 2010 with purely scientific goals. The day included 2 distinguished speakers (A. Tsoukias, CNRS and P. Perny, Univ. Paris 6) and two more "junior" speakers (S. Kaci, CRIL and F. Roda, LIX). The talks have been collected at `http://www.lix.polytechnique.fr/~liberti/multiobjective-100622/`.
3. *OSD workshop at Microsoft France: "L'innovation au secours de la planète"*. This took place on 27 january 2011 and had a three purposes: (a) to invite contributions from decision makers about sustainable development; (b) to publically announce the winning of the prestigious EURO/ROADEF Challenge; (c) to summarize the achievements after two years of the OSD chair's existence. The workshop attracted around 200 people. The programme included opening interventions by Mr. E. Boustouller (President of Microsoft France), Gen. X. Michel (Commander of Ecole Polytechnique), and Dr. Ph. Baptiste (Director of the INS2I CNRS Institute), a joint presentation by the two scientific co-directors (L. Liberti, LIX and Y. Hamadi, MSR), an interview to the winning EURO/ROADEF challenge team, a keynote presentation by R. Bernard (MS), and a round-table on the integration of sustainable development constraints in the enterprise decision process. See the corresponding entry in `http://chaire-osd.polytechnique.fr/accueil/evenements-generaux/`.

- **Sessions**.

1. D. Brockhoff and K. Deb, "Session: Evolutionary multiobjective optimization". 21st Intl. Conf. on Multiple Criteria Decision Making, June 13-17, 2011, Jyväskylä, Finland.
2. V. Jost and N. Touati-Moungla, "Session: Transport durable". 12th annual congress of the ROADEF, 2011, St. Étienne, France.
3. V. Jost and N. Touati-Moungla, "Stream: Optimization for Sustainable Development". 24th European Conference on Operational Research (EURO), July 2010, Lisbon, Portugal. Two sessions: "Green vehicle routing and scheduling", "Multi-objective optimization".
4. N. Touati-Moungla, "Session: VRP and applications". International Symposium on Combinatorial Optimization (ISCO 2010), March 2010, Hammamet, Tunisie.

Individual presentations are listed in `http://chaire-osd.polytechnique.fr/accueil/activites/conferences/`.

- **Sponsorship**.

1. CPAIOR 2010 `http://cpaior2010.ing.unibo.it/`
2. CP 2010 `http://cp2010.cs.st-andrews.ac.uk/`
3. TOGO 2010 `http://www.lix.polytechnique.fr/togo10/`
4. CSDM 2010 `http://www.csdm2010.csdm.fr/`

The chair also sponsors an invited seminar series. The details are at `http://chaire-osd.polytechnique.fr/accueil/activites/seminaires/`.

## 1.3 Roadmap

The general roadmap for the next three years is to keep on attaining the above objective, continue to excel within the international scientific community, and to start exploiting acquired knowledge for teaching and education purposes within the realm of sustainable development.

### 1.3.1 Scientific objectives

We shall devote considerable attention to the two most promising topics identified during the first two years of work: sustainable energy and sustainable transportation, as well as carrying out innovative methodological work in optimization techniques. We shall also continue to look for potentially interesting ideas linked to sustainable development, but this will no longer be the focus of our work.

More precisely, the following research topics will be explored in depth.

1. *Smart grids and energy production*. Smart grids are electrical grids that are supposed to be able to withstand practically every conceivable breakdown, integrate renewable energy sources, and adjust tariffs locally by means of a fine-grained interconnected control and counter network (see Ch. 3, 5, 4). Two industrial partners of coniderable standing (Alstom and EDF) have shown some interest in working with the OSD Chair on this subject. Specifically, Alstom is interested in solving the real-time energy market problem at the minute scale, as well as in integrating the different timescale problems routinely solved to maintain a complex electrical grid. EDF is interested in the scheduling of downtime in its network of nuclear plants. This line of research will be pursued by means of two postdoctoral fellowships dedicated to this topic, as well as by inviting field experts to give seminars.
2. *The rational use of energy in new generation buildings*. Typically, new buildings integrate sensors and controllers that work together to try and reduce thermal dissipation without impacting the quality of life. Such controls regulate temperature, humidity, hot water, and domestic appliances. Sensors must monitor the building state with respect to the inolved factors. Next generation buildings also include local energy generation facilities, such as solar cells, wind exploitation technology, or thermal collection from nearby plants. This generates several decision problems, from the reliable location of sensors to the rational use of local versus grid energy (in this sense this research is connected to smart grids). This line of research involves a full-time postdoctoral fellow (Ch. 12), whose current task is to improve an existing, open source architecture software with a wide distribution. We hope the visibility of this work to be remarkable, as well as the interest spawned in the community. This line of research will continue to be pursued by means of one postdoctoral fellowships dedicated to this topic, as well as by inviting field experts to give seminars.
3. *Environment-conscious transportation*. Typically, transportation and routing plans are optimized with respect to cost. Regard for the environment causes trade-offs in the cost-optimal plans, and requires different solution methodologies, which vary according to the kind of environmental objectives considered. In general, the decision maker is presented with an efficient set of routing plans involving several modes of transportations. This line of research involves a full-time postdoctoral fellow, a part-time Ph.D. student and a permanent researcher (see Ch. 4, 7, 8, 9, 10, 11). This line of research will continue to be pursued by means of one postdoctoral fellowships dedicated to this topic, as well as by inviting field experts to give seminars.
4. *Innovative methodologies in optimization technology*. Identifying and modelling problems is useless if one cannot solve them. A part of our budget will be dedicated to employing two postdoctoral fellows to investigate methods for robust multiobjective optimization, as well as TO inviting field experts to give seminars. Although for the moment we chose to focus on stochastic optimization methods (Ch. 13, 14, 15, 16), other promising techniques include mathematical programming based approaches to robustness, chance-constrained programming, and exact methods.

We shall also endeavour to continue looking for promising ideas that support the sustainable development paradigm in all forms (see e.g. Ch. 17, 18).

### 1.3.2 Didactical objectives

The OSD Chair will partially support the forthcoming *Master Parisien de Recherche Operationnelle* (MPRO) `http://uma.ensta.fr/mpro/`, insofar as the MPRO has an interest in teaching sustainable development and energy-related topics, in connection to Operations Research, to young graduates aiming to fulfill a research or professional career in industry or academia. The support will be in the form of payment of hourly teaching rates, institution of a "best master dissertation" prize, as well as in the form of facilitating, through the Chair's contact network, the match between finishing master students with corresponding industrial internships.

### 1.3.3 Dissemination objectives

The OSD Chair members and related co-workers will be encouraged to disseminate their scientific findings to relevant international conferences and workshops, where they will be invited to also organized special-interest sessions, as well as submit their completed work to high-caliber international scientific journals. A certain amount of sponsorship for conferences and workshops is also planned.

An important dissemination objective is that of forming tactical partnerships with industries that are interested in sustainable development, in order to find new problems and test innovative techniques on the field. Such tactical partnerships may eventually become strategic and long-term.

## 1.4 Contents of this report

The rest of this document is a collection of scientific reports on the several different research topics the OSD chair team worked on during the first two years of existence of the chair. The report is organized in four parts (sustainable energy, sustainable transportation, innovative methods for optimization and promising ideas), each of which consists of several chapters.

# Part I

# Sustainable Energy

# Chapter 2

# An ILP Approach to schedule outages of nuclear power plants

**Winning submission of the the EURO/ROADEF Challenge 2010**

VINCENT JOST, DAVID SAVOUREY

*Submitted to Journal of Scheduling*

*We address the problem of planning outages of nuclear power plants submitted by EDF as the Challenge EURO/ROADEF 2010. As we won the first prize of the contest in the senior category, our approach may deserve some interest: it is conceptually simple, easy to program and computationally relatively fast. The quality of our solutions, although a few percent behind the best known ones, still leaves hope for our approach to become competitive. Indeed, to fit with the deadline of the contest, we made several major simplifications without testing alternatives. Several assumptions and choices could be studied more carefully in further studies. We therefore try to present both our method and some ideas to improve it.*

## 2.1 Context, motivations

The subject of the Challenge ROADEF/EURO 2010 was proposed by EDF [1]. Nuclear power plants need to be stopped regularly for maintenance and refuel. The problem mainly consists, over a 5 years horizon, in deciding when to stop nuclear power plants and by how much to refuel the reactors, so as to both ensure the ability to robustly fulfil a stochastic demand, while minimizing the expected cost of production. All details can be found on the web site of the challenge [1]. From now on, we assume that the reader is familiar with the subject of the challenge, or at least has a global view on the 21 types of constraints.

Our original motivation to participate to the challenge was to understand the relevance of optimization techniques in energy management problems. Within the context of our research group "Optimization for Sustainable Development", proving our know-how on a real world problem was also a good way to promote our visibility and credibility.

Our original motivation was not to produce as good solutions as possible. We wanted to understand the problem as much as we wanted to solve it. Indeed, when it turned out, less than one month before the deadline, that our approach had a chance to become competitive, it was to late. We chose to work on robustness rather than optimality. This choice was also driven by the study of the recent history of

the challenge: Hidden X-instances might be slightly different, bigger and harder than known instances. All this explains why we have only some 3000 lines of used code, maybe 2 or 3 times less than the very competitive teams.

Unfortunately, the problem was too complex for us to propose a very pure and transparent algorithm. In order to find acceptable solutions within the deadlines, we had no more time to study insightfully the following question: "Given fixed dates of outage, how should we choose values of refueling so as to both ensure feasibility and near-optimality ?" On the other hand, we grasped intuitively more features about the problem than we had time and abilities to translate them into algorithms as discussed in Section 2.7.

This paper is organized as follows : In Section 2.2, a global presentation of our approach is given. The Section 2.3 focus on the scheduling sub-problem and the way we modelled it. In Section 2.4, algorithms used to solve the refueling sub-problem are presented. Numerical results are given into Section 2.5. Section 2.6 consists in an analysis of the instances: we present some features of the instances that help designing efficient ways to solve the real world cases of the problem (that is, at least for EDF in France). Finally, a discussion about our approach and the way it could be improved is given in Section 2.7.

## 2.2 Overview of the method

Conceptually, our approach consists in a hierarchical decomposition of the problem into three phases : we fix some variables before working on fixing subsequent other variables.

1. *ILP* : Fix the dates of outages (variables $ha_{ik}$),
2. *Dichotomy* : Fix the values of refueling (variables $r_{ik}$),
3. *Greedy* : Fix the production levels (variables $p_{its}$ and $p_{jts}$).

The first phase is mainly done using an ILP, the second phase using a dichotomised search on the values of refueling and the third, running greedily the dynamic of the nuclear plants over the time horizon. Hence, within this paper, "ILP" refers to the first phase while "Dichotomy" refers to the second phase. However, since we often need to check exactly the feasibility of the decisions taken, we use the "Greedy" subroutine in various places, and not only in the third phase.

### 2.2.1 Specifications, utilization and principle of the ILP

The ILP encodes exactly the constraints linking the dates of outages of various nuclear plants (CT14-21), as well as the bounds on these dates (CT13). The ILP also encodes constraints on minimum distances between pairs of outages of each given plant (see section 2.3.3). Theses (constraints on) distances are not given in the subject. We designed them as a way to aggregate approximately the constraints on the dynamic of production, interruptions and refueling in each nuclear plant (CT3-13) to help our ILP finding solutions that lead to the global feasibility of the problem.

These distance constraints being only necessary and not sufficient to satisfy (CT3-13), we run the Greedy routine at the end of the ILP. If the Greedy routine says that (CT3-13) can be satisfied, we go to the second phase. If not, we increase the minimum date ($TO_{ik}$) on an outage on which we detected infeasibility (technically this is when we observe a fuel level before outage $(i,k)$ greater than the allowed $AMAX_{ik}$).

The ILP obviously needs an objective function to find appropriate outages dates that will lead to a global solution with good objective value. As discussed in section 2.3.3, a binary variable $p_{ih}$ serves as a rough estimate of whether plant $i$ will be able to produce at week $h$ or not (depending on the dates

of outages of plant $i$). Having pre-computed some estimate $val_{ih}$ on the economical value of producing on $i$ at $h$ the ILP maximizes $\sum_i \sum_h val_{ih} p_{ih}$. Although the variables $p_{ih}$ are the only ones in the objective functions, we only care about the variables $x_{ikh}$ so as to fix the dates of outages $ha_{ik}$ after the solution of the ILP.

### 2.2.2 Specifications, utilization and principle of the Dichotomy

Once the dates of outage are fixed, we are looking for refueling values. We say that a set of refueling values $\{r_{ik}\}$ is feasible if we are able to find a global feasible solution using these values. Our goal is to maximize the ability to produce with nuclear plant while keeping a global feasible solution. As it was too hard for us to compute directly feasible refueling with high values, we use a kind of target-checker search on possible values. This method relies heavily on the monotonic feasibility of refueling values: If $\{r_{ik}\}$ is an infeasible set and $r'_{ik} \geq r_{ik}$ for all $(i,k)$, then $\{r'_{ik}\}$ is also infeasible. The method relies also on the fact that generally we prefer higher refueling values than smaller. We therefore start with $r_{ik} = RMIN_{ik}$. Then, given a feasible vector "$low$" of refueling values, we try to guess by how much we can increase them, thus obtaining a vector "$high$". If "$high$" is feasible, "$low \leftarrow high$". If "$high$" is infeasible, "$high \leftarrow (high + low)/2$". These rules are iteratively applied until "$low$" and "$high$" are almost equal (for all $(i,k)$). At the end we have a feasible set of refueling values, and we didn't find opportunities to improve any value in this set.

### 2.2.3 Specifications, utilization and principle of the Greedy

Our approach needs ways to make sure, during the first two phases, that we will be able to find feasible solutions based on the values we chose for dates and values of refueling. To do this, we need to chose production levels, either for each nuclear plan separately of for nuclear plants together. Therefore, two different greedy procedures are used.

The first one does not take into account the scenarios, that is to say there is no demand, and no modulation is required. This procedure is presented in details with Algorithm 1. It is used in order to check that chosen dates of outages for each plant are not violating the constraints involving only that plant. It is also used to compute how much modulation we can afford on a cycle of a plant, before conflicting with constraints involving only that plant.

The second procedure takes into account the demand, it chooses production levels for each scenario, by modulating if necessary. To do this, a rolling horizon procedure is used. For each scenario, we follow the dynamic of fuel for each nuclear power plant, producing at $PMAX_i^{t,s}$ in modulation phase and following the profile constraint otherwise. Whenever the production exceeds the demand in a scenario, we modulate on chosen power plants until the production equals the demand. We modulate first on the plant for which we have the largest estimate of allowed modulation. If there exists a scenario and a time step on which more modulation would lead to violating some constraints but production is still above demand, the procedure answers "no". This method is described more precisely with Algorithm 2.4.

## 2.3 Scheduling outages

### 2.3.1 Variables of the ILP

We use the following binary decision variables (which allow exact reformulations of the variables $ha_{ik}$).

$$x_{ikh} = \begin{cases} 1 \text{ if the k-th outage of plant } i \text{ starts at week h,} \\ 0 \text{ otherwise;} \end{cases}$$

We also use the following auxiliary binary decision variables, which allow to express linearly a heuristic cost function that helps producing values for the $x$ variables which yield a good final solution.

$$p_{ih} = \begin{cases} 1 \text{ if plant } i \text{ is likely to be able to produce at week } h, \\ 0 \text{ otherwise;} \end{cases}$$

Contrary to variables $x$, the values of variables $p$ are not intended to be used outside of the ILP. They just help defining (and hence finding) the quality of outages dates.

### 2.3.2 Reformulation of constraints CT13-21

In the subject, each constraint in the set [CT14-21] is defined relatively to a subset, denoted $A_m$, of nuclear plants and also relatively to a number of weeks, denoted $Se_m \in \mathbb{Z}$. $Se_m$ describe a minimum distance (if $\geq 0$) or a maximum overlapping (if $\geq 0$).

**Constraint** $CT13$

Each outage $(i,k)$ has at most one starting week $h$, which must be in within the bounds $[TO_{ik}, TA_{ik}]$:

$$\forall (i,k) \in I \times K, \sum_{h \in [TO_{ik}, TA_{ik}]} x_{ikh} \leq 1$$

$$\forall (i,k) \in I \times K, \forall h \notin [TO_{ik}, TA_{ik}], x_{ikh} = 0$$

If $k \geq 1$ and if the $k$-th outage of plant $i$ starts at week $h$, then the $k-1$-th outage of plant $i$ must occur, with a starting week not later than $h - DA_{i,k-1}$[1].

$$\forall h \in H, \forall (i,k), \sum_{h'=0}^{h-DA_{i,k-1}} x_{i,k-1,h'} \geq x_{ikh}$$

**Constraint** $CT14$

This constraints considers outages as intervals $(ha_{ik}, ha_{ik}+DA_{ik})$, and requires separation of these intervals. We reformulate it as a packing constraint:

$$\forall m, \forall h, \sum_{i \in A_m} \sum_{\substack{k \text{ s.t.} \\ DA_{ik}+Se_m-1 \geq 0}} \sum_{h'=h-DA_{ik}-Se_m+1}^{h} x_{ikh'} \leq 1$$

However, the above constraints are not sufficient, because outage for which $DA_{ik} \leq -Se_m$ are not considered. If it exists, each such outage must not interfere with any of the packing described above.

[1] The following constraints is written in the last paragraph, page 27 of the subject

$$\forall m, \forall i \in A_m, \forall k \ \ s.t. \ \ DA_{ik} \leq -Se_m, \forall h,$$
$$x_{ikh} + \sum_{i' \in A_m} \sum_{\substack{k' \ \ s.t. \\ DA_{i'k'} + Se_m - 1 \geq 0}} \sum_{h'=h-DA_{i'k'}-Se_m+1}^{h+DA_{ik}+Se_m-1} x_{i'k'h'} \leq 1$$

We invite the reader to take time digesting the last constraint. Indeed we completely overlooked it in the first place. The ILP had to find solutions that were claimed infeasible by the checker of EDF before considering that something was missing to encode CT14. This is typically the kind of bug that could have shown up only on X instances...

**Constraint** *CT*15

It is the same constraint that *CT*14, but restricted to outages which intersect the interval of week $[ID_m, IF_m]$.

For convenience, we introduce the following notations :

$$w_{ikm} = [ID_m - DA_{ik} + 1, IF_m],$$

$$y_{iki'k'm}(h) = [h - DA_{i'k'} - Se_m + 1, h + DA_{ik} + Se_m - 1].$$

$$\forall m, \forall h \in H, \sum_{(i,k) \in A_m} \sum_{\substack{h' \in [h-DA_{ik}-Se_m+1, h] \\ h' \in w_{ikm}}} x_{ikh'} \leq 1$$

Then, the second part of *CT*15 can be expressed as :

$$\forall m, \forall i \in A_m, \forall k \ \ s.t. \ \ DA_{ik} \leq -Se_m, \forall h,$$
$$x_{ikh} + \sum_{i' \in A_m} \sum_{\substack{k' \ \ s.t. \\ DA_{i'k'} + Se_m - 1 \geq 0}} \sum_{\substack{h' \in y_{iki'k'm}(h) \\ h' \in w_{i'k'm}}} x_{i'k'h'} \leq 1$$

**Constraint** *CT*16

For each week $h$, the number of decoupling dates in interval $[h, h + Se_m - 1]$ must be smaller or equal to 1.

$$\forall m, \forall h \in H, \sum_{(i,k) \in A_m} \sum_{h'=h}^{h+Se_m-1} x_{ikh'} \leq 1$$

**Constraint** *CT*17

For each week $h$, we consider interval $[h, h + Se_m - 1]$ and impose that at most one coupling date belongs to this interval. It is equivalent to say that at most one decoupling date belongs to $[h - DA_{ik}, h - DA_{ik} + Se_m - 1]$. Thus, we have :

$$\forall m, \forall h \in [0, H + \max_{ik} DA_{ik} - 1], \sum_{(i,k)\in Am} \sum_{h'=h-DA_{ik}}^{h-DA_{ik}+Se_m-1} x_{ikh'} \leq 1$$

**Constraint** *CT*18

Again, we consider interval $[h, h + Se_m - 1]$. There must be at most one coupling date or decoupling with this interval. Outage $(i, k)$ couples or decouples in $[h, h + Se_m + 1]$ iff $ha_{ik} \in [h - DA_{ik}, h + Se_m - 1]$. Thus, *CT*18 can be rewritten as :

$$\forall m, \forall h \in [0, H + \max_{ik} DA_{ik} - 1], \sum_{(i,k)\in A_m} \sum_{h'=h-DA_{ik}}^{h+Se_m-1} x_{ikh'} \leq 1$$

**Constraint** *CT*19

There must be at most $Q_m$ decoupling dates within every intervals $[h' + L_{ikm}, h' + L_{ikm} + TU_{ikm}]$ :

$$\forall m, \forall h \in H, \sum_{(i,k)\in A_m} \sum_{h'=h-L_{ikm}-TU_{ikm}+1}^{h-L_{ikm}} x_{ikh'} \leq Q_m$$

**Constraint** *CT*20

For a given week $h$, the number of outages overlapping this week must be smaller than $N_m(h)$.

$$\forall m, \sum_{(i,k)\in A_m} \sum_{h'=h-DA_{ik}+1}^{h} x_{ikh'} \leq N_m(h)$$

**Constraint** *CT*21

Constraints *CT*21 were almost not reformulated, we expressed it nearly as in the subject :

$$\forall m, \forall h \in IT_m, \forall t \in h, \quad \sum_{i\in C_m} \sum_{k\in K} \sum_{h'=h-DA_{ik}+1}^{h} PMAX_i^t \cdot x_{ikh'} \leq IMAX_m$$

### 2.3.3 Minimum distances between the outages within each plant

The variables $x$ do not allow to express the constraints concerning the production of nuclear plants. Indeed, constraints [CT3-11] imply that, even if refueling at the minimum allowed $RMIN_{ik}$ and producing at $PMAX_i^t$, the plant needs several weeks for its fuel level to fall below $AMAX_{i,k+1}$.

Let $i \in I$ and $C \subseteq [0..K[$, we say that a set of week of outage $\{ha_{ik}\}_{k\in C}$ is "feasible" is there exists values $\{ha_{ik}\}_{k\in[0..K[\setminus C}$ such that there exists fuel, production and refueling levels such that plant $i$ satisfies the constraints (involving only itself) [CT3-13].

It turns out that a set of outages dates $\{ha_{ik}\}_{k\in[0..K[}$ might be infeasible, while fixing all but one of these values might still be feasible. One might then wish to write constraints on $\{ha_{ik}|k \in C \subseteq [0..K[\}$ "as they are needed" (for example, when the current solution violates them). But this would lead to an iterative approach calling the ILP solver as a subroutine.

Concrete values in the data set suggest that feasible outage dates are highly (and mainly) constrained by minimum distances between (consecutive) pairs. In other words, under constraints involving pairs of outages of a plant, a set of value $\{ha_{ik}\}_{k\in[0..K[}$ is often feasible or requires shifting outages to the future by only one week to become feasible.

Our trade-off is thus to rely only on constraints involving pairs. To do so, we pre-compute exactly the values (with $k < k'$):

$$Earl(i, k, k_2, h) := \text{ first week } h_2 \text{ for which } x_{ikh} \text{ and } x_{ik_2h_2} \text{ are feasible together}$$

These values (*Earl*) are then all encoded in our ILP, with the following constraints:

$$\forall h \in H, \forall (i,k), \forall k' > k,\ x_{ikh} + \sum_{h'=0}^{Earl_{i,k,k_2,h}} x_{i,k_2,h_2} \leq 1$$

The values $Earl(i, k, k', h)$ are deduced on each plant $i$ independently, using a dynamic program to first compute the following quantities.

Let $MinFl(i, k, h)$ be the minimum fuel level that we can achieve at week $h$, assuming that outage $k$ has been completed (that is, with $h \geq ha_{i,k} + DA_{i,k}$). Let $MinFuel(i, k, k_2, h, h_2)$ be the minimum reachable fuel level at week $h_2$ such that $k_2$ is the last completed outage, and such that outage $k$ has started at week $h$.

Assuming these values are computed, one deduces easily earliest feasible week of an outage and minimum distances between outages of one plant. This amounts to ask whether *MinFl* and *MinFuel* are lower than *AMAX*. Notice first that computing values *MinFuel* knowing the values *MinFl* is very similar to computing *MinFl* starting form week 0 with fuel level $XI_i$.

The rest of section 2.3.3 describes the dynamic program. Constraint [CT10] implies that the dynamic of refueling is non-linear (although affine): if $t-1$ is the first time step of outage $(i, k)$, then the fuel level at time $t$ is given by $fuel(i, t)\Psi_{i,k}(fuel(i, t-1)) + r_{ik}$ with

$$\Psi_{i,k}(x) := \frac{Q_{i,k} - 1}{Q_{i,k}}.(x - BO_{i,k-1}) + BO_{i,k}$$

(where $Q_{i,k}$ is equal to 3 or 4 in practice).

Constraint [CT6] says that, if the fuel level is less than a threshold $BO_{i,k}$, then, until the plant enters an outage, the production is constrained to a proportion of its capacity $PMAX_i^t$. This proportion decreases with the fuel level, and is given by a (piecewise-linear decreasing function $PB_{i,k}$). We simplify, and indeed, over-constrain [CT10] (which normally allows a tiny margin) into:

$$p(i, t) = PMAX_i^t * PB_{i,k}(fuel(i, t))$$

.

To achieve $MinFl(i, k, h)$, we assume that each refueling is done at $RMIN_{ik}$ and that consumption is always equal to $PMAX_i^t$ in modulation phase. The second hypothesis is false in general because using modulation to adjust sharply the fuel level before an outage might help decreasing it at the end of

the profile phase. Anyway, although not exact, these hypothesis are almost exact given the order of magnitude of the numbers in the data set.

One concludes that, in order to compute $MinFl(i,k,h)$ we shall use the following equations describing fuel levels and power.

$$p(i,t) = \begin{cases} 0 & \text{if plant } i \text{ is in outage at } t \\ PMAX_i^t & \text{if } fuel(i,t) \geq BO_{ik} \\ PMAX_i^t.PB_{i,k}(fuel(i,h)) & \text{if } fuel(i,t) < BO_{ik} \end{cases} \tag{2.1}$$

$$fuel(i,t) = \begin{cases} XI_i & \text{if t=0} \\ \Psi(fuel(i,t-1)) + RMIN_{ik} & \text{if outage } k \text{ starts at } t-1 \\ fuel(i,t-1) - p(i,t-1).D & \text{otherwise} \end{cases} \tag{2.2}$$

Of course, to achieve $MinFl(i,k,h)$, we still need to choose the dates of outages. $MinFl(i,k,h)$ is set to $+\infty$ at the initialization of the dynamic program. It may remain infinite, indicating that is is impossible to complete outage $k$ by week $h$. We start with initial conditions $MinFl(i,0,0) := XI_i$.

We compute the following values (as they are needed and without storing them): let $Conso(i,k,h-1,x)$ be the quantity of fuel consumed during all time steps of week $h-1$ under equations (2.1) and (2.2), assuming that plant $i$ is in cycle $k$, and with fuel level $x$ at the beginning of that week.

$MinFl(i,k,h)$ satisfies a Bellman relation. It is the minimum between two values:

$$\begin{cases} MinFl(i,k,h-1) - Conso(i,k,h-1,MinFl(i,k,h-1)) \\ \Psi_{i,k}(MinFl(i,k-1,h-DA_{ik})) + RMIN_{ik} \end{cases} \tag{2.3}$$

The first case is always acceptable, but the second can be chosen only if both $MinFl(i,k-1,h-DA_{ik}) \leq Amax_{ik}$ and $TO_{ik} \leq h - DA_{ik} \leq TA_{ik}$.

Concerning $MinFuel(i,k,k_2,h,h_2)$, we don't go into computational details here. Let us just mention that $MinFuel(i,k,k_2,h,h_2)$ satisfies a relation quasi-identical to (2.3) as it is the minimum between:

$$\begin{cases} MinFuel(i,k,k_2,h,h_2-1) - Conso(i,k_2,h_2-1,MinFuel(i,k,k_2,h,h_2-1)) \\ \Psi_{i,k}(MinFuel(i,k,k_2-1,h,h_2-DA_{ik_2})) + RMIN_{ik_2} \end{cases} \tag{2.4}$$

### 2.3.4 Computing an approximate objective function

One first issue that we had to solve, before evaluating costs, is to evaluate the nuclear availability at each step, or at least at each week. We chose to evaluate the availability of each plant for each week. Actually, we assume here that a plant either produce at $PMAX$ or is offline (hence neglecting modulation and profile phase effects !).

Let the following binary variable be

$$p_{ih} = \begin{cases} 1 \text{ if we forcast a production (at } PMAX) \text{ on plant } i \text{ at week h,} \\ 0 \text{ if we forcast no production on } i \text{ at } h; \end{cases}$$

Variables $p$ must be constrained to 0 while on outage:

$$p_{ih} + \sum_{k} \sum_{h'=h+1-DA_{ik'}}^{h} x_{ikh'} \leq 1 \qquad \forall i, h$$

$p$ should be set to 0 in case of fuel shortage. Concerning fuel shortage, we assume that each cycle starts at level $RMAX_{ik'}$ and that the fuel is consumed at $PMAX_{i,t}$. It is therefore immediate to precompute the number $N_{ikh'}$ of weeks during which we can produce at $PMAX$ before fuel stock falls below 0. The following constraints is added to the ILP.

$$p_{i,h} \leq \sum_{k} \sum_{h'=h-N_{ikh'}}^{h-DA_{ik}} x_{ikh'} \qquad \forall i, h \tag{2.5}$$

Now that we have an estimate of the availability of plant $i$ at week $h$, we can express an objective function. One of our first idea was

$$\text{objective} = \max \sum_{h} \sum_{i} Dem_h.p_{ih} \tag{2.6}$$

where $Dem_h$ is the average demand at week $h$.

Many hypothesis made in this section seem unjustified oversimplifications of the problem. Notice however the various advantages of the approach. Thanks to the objective function, the ILP is decomposable on subsets of plants that are not interacting through [CT14-21]. Assuming fuel levels of $RMAX$ after outage is an under approximation of the fuel level we can ensure, but often, the maximum fuel allow $SMAX$ is not much higher than $RMAX$. Moreover, when $RMAX - RMIN >> 0$, the objective function tends to help satisfying constraints 2.3.3 discussed in section 2.3.3, because scheduling the next outage much before the end of the $N_{ikh}$ allowed weeks results in a loss in the objective function.

## 2.4 Production and refuelings levels

In all this section, we consider that $ha_{ik}$ are already set, so they could be considered as implicit inputs of the algorithms discussed here.

### 2.4.1 Fuel Levels without modulation

Algorithm 1 computes bounds on fuels levels for a given plan of outages and refuels, without taking into account the demand. For every nuclear power plant $i$, the algorithm iterates increasingly every time slot. At each step, the production is set as follows : if power plant $i$ is in a modulable production period, then the production is set to $PMAX_i^t$. If power plant $i$ is in a profiled production period, production is set by the profile constraint. If power plant $i$ is in outage, production is null. The refuelings are done according to the input $r_{ik}$. During loop over time steps, several bounds on fuel levels are computed and stored :

- $amaxGap_{ik}$ : difference between $AMAX_{ik}$ and fuel level $x[i]$ just before the refueling $r_{ik}$ ;
- $smaxGap_{ik}$ : difference between $SMAX_{ik}$ and fuel level $x[i]$ just after the refueling $r_{ik}$ ;
- $famax_{ik}$ : minimum between $amaxGap_{ik}$ and $smaxGap_{ik}$ ;
- $nomodfuel_{ik}$ : fuel level $x[i]$ at the end of cycle $(i, k)$.

**Algorithm 1**: $bounds(r_{ik})$

**Data**: $r_{ik}, i \in I, k \in K$
**Result**: $famax_{ik}, amaxGap_{ik}, smaxGap_{ik}, nomodFuel_{ik}, i \in I, k \in K$, a set of power plant indexes *toReschedule*

```
begin
    for i ∈ I do
        initialize state[i], cycle[i], x[i] ;
    for i ∈ I do
        for t ∈ T do
            if state[i] = MOD then
                prod[i] ← PMAX[t] ;
            if state[i] = PROF then
                prod[i] ← PROF[i, t] ;
            if state[i] = OUT then
                if t = ha_ik then
                    compute amaxGap_ik ;
                    if x[i] violates AMAX[i][cycle[i]] then
                        toReschedule+ = i ;
                    update x[i] with r_ik ;
                    if x[i] violates SMAX[i][cycle[i]] then
                        toReschedule+ = i ;
                    update smaxGap_ik ;
                    update famax_ik ;
                prod[i] ← 0 ;
            x[i] ← x[i] − prod[i] · D ;
            if t + 1 = ha_{i,k+1} then
                nomodfuel_ik ← x[i] ;
            update state[i], cycle[i] for next time step ;
end
```

### 2.4.2 Exact production level affectation

**Algorithm 2**: $affectScenar(s)$

**Data**: $r_{ik}, i \in I, k \in K$, scenario index $s$
**Result**: boolean value and $p_{its}$ and $p_{jts}$ for $i \in I, j \in J, t \in T, s \in S$
**begin**
  **for** $i \in I$ **do**
    initialize $state[i], cycle[i], x[i], fprime[i]$ ;
  **for** $t \in T$ **do**
    $dem \leftarrow DEM[t][s]$ ;
    **for** $j \in J$ **do**
      $prod[j][t][s] \leftarrow pmin[j][t][s]$ ;
      $dem \leftarrow dem - prod[j][t][s]$ ;
    compute $puissMax$ (available nuclear production at time $t$) ;
    **if** $puissMax > dem$ **then**
      $toModul \leftarrow puissMax - dem$ ;
      **for** $i \in I$ **do**
        $capaModul[i] \leftarrow \min(PMAX_i^{t,s}, fprime[i])$ ;
      **while** $toModul > 0$ **do**
        choose $i^*$ such that $fprime[i^*] \geq fprime[i] \forall i$ and $state[i^*] = MOD$ and $capaModul[i^*] > 0$ ;
        **if** *there is no candidate for* $i^*$ **then**
          return *false* ;
        $modul[i^*] \leftarrow \min(toModul, capaModul[i^*])$ ;
        update $fprime[i^*], capaModul[i^*]$ ;
        $toModul \leftarrow toModul - modul[i^*]$ ;
    **for** $in \in I$ **do**
      **if** $state[i] = MOD$ **then**
        $prod[i] \leftarrow PMAX_i^{t,s} - modul[i]$;
      **if** $state[i] = PROF$ **then**
        $prod[i] \leftarrow PROF[i, t]$ ;
      **if** $state[i] = OUT$ **then**
        $prod[i] \leftarrow 0$ ;
      update $x[i]$, $cycle[i]$, $state[i]$, $fprime[i]$ for next time step ;
    fill demand for each time step $t$ with type 1 power plants sorted by increasing cost $C_j^{t,s}$ ;
    return *true* ;
**end**

Algorithm 2 affects power production levels for each power plant for a given a scenario $s$. Outage dates and refuelings are known. It is a rolling horizon procedure. As for Algorithm 1, several informations are maintained for each nuclear power plant : the current index of cycle $cycle[i]$, the current state $state[i]$ and the current fuel level $x[i]$. In addition, an heuristic lower bound $fprime[i]$ on the remaining capacity of modulation is also computed. It takes into account the current maximum modulation $MMAX_{i,k}$ and the current and following gaps to $amaxGap_{ik}$ and $smaxGap_{ik}$. This indicator is (re)computed at each beginning of cycle, and decreased according the modulation already done in the current cycle. At a given time step $t$, these steps are followed : affect to type 1 power plants their minimum production levels and update the demand in consequence ; compute the maximal available nuclear production capacity $puissMax$ with nuclear power plants in profile production state and in modulation production state. If $puissMax$ is bigger than the demand, modulation will be necessary at this time step. Quantity $toModul$ is initialised to $puissMax - dem$. For each nuclear power plant $i$, a quantity $capaModul[i]$ is then computed, as the minimum between current maximum production $PMAX_i^{t,s}$ and remaining capacity

of modulation $fprime[i]/D$. Then, a loop is executed until *toModul* becomes nul. At each iteration, the nuclear power plant with the biggest value of $fprime$ is chosen, among plants in modulation state which a non null instataneous capacity of modulation (ie $capaModul[i] > 0$). The index pf this chosen power plant is denoted $i^*$. Once $i^*$ is computed, the quantity to modulate on this power plant is set to $\min(toModul, capaModul[i^*])$. Then, $modul[i^*]$, $fprime[i^*]$, $capaModul[i^*]$ are updated. Once this modulation phase is over, all nuclear production levels are affected for nuclear power plants. Then, all informations maintained over nuclear power plants are updated if necessary. Once this has been done for every time step, there is a new rolling time procedure to fill if needed the production to stick on the demand. At each time, type 1 power plants are sorted increasingly according to their costs $C_j^{t,s}$. Then, we use fill the production on the type 1 power plants, until the demand is reached.

### 2.4.3 Avoiding lacks of fuel

**Algorithm 3**: $fillIdleProd(r_{ik})$

**Data**: $r_{ik}, i \in I, k \in K$
**Result**: new values for $r_{ik}, i \in I, k \in K$
**begin**
  **for** $k \in K$ **do**
    **for** $i \in I$ **do**
      **if** *outage* $(i,k)$ *exists and* $nomodfuel_{ik}$ *is null* **then**
        increase $r_{ik}$ so that fuel level becomes null at the very end of cycle $(i,k)$ ;
        adjust $r_{ik}$ to satisfy constraints on $SMAX_{ik}$, $RMIN_{ik}$ and $RMAX_{ik}$ ;
    Recompute bounds on fuel levels with Algortihm 1
**end**

It could happens that a power plant can not product at the end of its production period, due to a lack of fuel. Algorithm 3 aims at avoiding these situations. We are given refuelings $r_{ik}, i \in I$ and fuel bounds $nomodfuel_{ik}, i \in I, k \in K$. Then, this method iterates on cycles. For each cycle $k$, we consider successively each nuclear power plant $i$, and we check if the fuel level is null at the end of cycle $(i,k)$, with the estimator $nomodfuel_{ik}$. If this quantity is null, we increase $r_{ik}$ so that the fuel level becomes null at the end of cycle $(i,k)$. Once this has been done for every nuclear power plant, we recomputes bounds on fuel level using Algorithm 1. Then, the next cycle is considered, and so on.

### 2.4.4 Increasing refuels

We would like to maximize the amount of refuelings over all nuclear power plants, without breaking the feasibility of the solution. Because it may be hard to ensure this feature, we use an optimistic algorithm that increase refuelings, and we will deal with the feasibility later. Increasing refuels is done in Algorithm 4. We consider iteratively every cycles. For each cycle, all nuclear power plant are treated independently. Given a power plant $i$ and a cycle $k$, a quantity $add_{ik}$ is computed and added to $r_{ik}$. To compute $add_{ik}$, we need to fulfill the following requirements :

- do not violate $SMAX_{ik'}$, for $k' \geq k$ ;
- do not violate $AMAX_{ik'}$, for $k' > k$ ;
- do not violate $RMAX_{ik'}$, for $k' \geq k$ ;
- $nomodfuel_{ik} \leq BO_{ik}$ ;
- maximize $add_{ik}$.

**Algorithm 4**: $incRefuels(r_{ik})$

**Data**: $r_{ik}, i \in I, k \in K$
**Result**: new values for $r_{ik}, i \in I, k \in K$
**begin**
  **for** $k \in K$ **do**
    **for** $i \in I$ **do**
      **if** *outage* $(i,k)$ *exists* **then**
        Compute maximal value for $add_{ik}$ (quantity of fuel to add to $r_{ik}$) so that : no fuel bounds constraints are violated within the current and the following cycles ; fuel at the end of cycle $(i,k)$ is no more than $BO_{ik}$ ;
        $r_{ik} \leftarrow r_{ik} + add_{ik}$ ;
    Recompute bounds on fuel levels with Algortihm 1
**end**

As we have an heuristic evaluation of fuel levels at the end of every cycles with *nomodfuel* values, we are able to estimate the maximal amount of fuel that we can add to $r_{ik}$ without breaking bound constraints over *AMAX*, *SMAX* and *RMAX*. To target being at $BO_{ik}$ at the end of cycle $(i,k)$, we just suppose the power plant $i$ products at *PMAX* each day of cycle $k$ and get back a quantity of fuel needed at the beginning of cycle $(i,k)$.

After each power plant has been considered for cycle $k$, we recompute indicators on distance to fuel bounds using Algorithm 1.

### 2.4.5 Optimal refuels : Up&Down refuels

**Algorithm 5**: $affectPowerOk(r_{ik})$

**Data**: $r_{ik}, i \in I, k \in K$
**Result**: a boolean indicating if power affecation is feasible
**begin**
  $ok \leftarrow true$ ;
  **for** $s \in S$ **do**
    $ok \leftarrow ok \& affectScenar(s)$ ;
  return $ok$ ;
**end**

Algorithm 5 is very simple. It just tries to affect production levels in evry scenario, and return true iff a feasible solution has been found.

Algorithm 6 is aimed to find biggest values of refueling such that a feasible solution for the general problem has been founds with these values. It maintains two sets of refueling values : *low* and *high*. The *low* set is always feasible (meaning that we are able to build a feasible solution from this set). The *high* set is an optimistic increasing based on the *low* set. The algorithm checks iteratively whether or not *high* set is feasible. In the positive case, *low* takes values from *high* and a new *high* set is computed from this new *low* set, using Algorithm 4. In the negative case, *low* does not change and *high* is decreased. The algorithm runs until *low* and *high* are different enough, or a time limit constraint has been reached. At the end of this algorithm, the *low* set is always feasible.

**Algorithm 6**: $upAndDown(r_{ik})$

```
Data: r_ik, i ∈ I, k ∈ K
Result: new values for r_ik, i ∈ I, k ∈ K
begin
    low ← r_ik ;
    high ← incRefuels(r_ik) ;
    while low ≠ high do
        if affectPowerOk(high) then
            low ← high ;
            high ← IncRef(high) ;
        else
            high ← middle(low, high) ;
end
```

## 2.5 Numerical Results

Configuration PC : 4 processors Dual Core AMD Opteron(tm) Processor 275 (2.2GHz, 1Mo cache), 16Go de RAM, ILOG CPLEX 12.1. One of the eight cores was used. NB: CPU time is computed before printing the solution in a file.

| Instance | Objective Value | CPU time (hh:mm:ss) |
|---|---|---|
| 0 | 8.73566e+12 | 00:00:00 |
| 1 | 1.69937e+11 | 00:00:08 |
| 2 | 1.46383e+11 | 00:00:16 |
| 3 | 1.5501e+11 | 00:00:15 |
| 4 | 1.12643e+11 | 00:00:29 |
| 5 | 1.26876e+11 | 00:00:47 |
| 6 | 8.68747e+10 | 00:01:56 |
| 7 | 8.36859e+10 | 00:02:20 |
| 8 | 8.7091e+10 | 00:04:31 |
| 9 | 8.68907e+10 | 00:14:49 |
| 10 | 8.06461e+10 | 00:04:28 |

## 2.6 Analysis of the data

The subject of the challenge in full generality is highly intractable. However there are structural properties that can be (robustly) found in all the 16 instances proposed by EDF. These properties are important to state precisely and formally, because they (might) help finding efficient resolution techniques. Let us the most structuring properties we were able to find.

**Preprocessing PMIN.** Studying the subject only and not the instances, observe that minimum power of type 1 plants can be assumed to be 0. Indeed, as a preprocessing, for all $(j, t, s)$ one can decrease $PMIN_{jts}$, $PMAX_{jts}$ and $DEM_{jts}$ by $PMIN_{jts}$. To let the cost invariant by this preprocessing, we also need to increase the cost function by $PMIN_{jts}.D/S$.

**Monotonic feasibility of refuels.** The main structuring characteristic of data is that non-nuclear availability is large enough so that it always allows to match demand, whatever the nuclear power available. On the other hand, too nuclear availability might require too much modulation and lead to global infeasibility. These two properties imply that, given fixed outage dates, if $\{r_{ik}\}$ is infeasible and $RMIN_{ik} \leq r_{ik} \leq r'_{ik}$ for all $(i,k)$ then $\{r'_{ik}\}$ is also infeasible.

**Approximation and evaluation of costs.** Nuclear production costs are roughly between 15 and 21, while type 1 plants costs typically vary between 10 and 10k. Most of the time, the most expansive nuclear plant is cheaper than the cheapest non-nuclear one, but this not always true. The costs of type 1 plants are indicated in Figures 2.1, 2.2 and 2.3. Type 1 completion costs are convex piecewise linear for each scenario and time step. Hence, associated marginal costs are non-increasing step functions (of nuclear power available). One should be careful on his hypothesis when aggregating on scenarii and time steps. However, if we rely on the Figure 2.2, that is, if nuclear availability is assumed uniform in a given week, we observe that marginal cost is constant (for small value on the x-axis), then piecewise linear, then strictly convex. This implies that the cost can be sharply approximated on some interval, by an affine or a quadratic function. Further Log(-log) plots should provide more insight on the nature of the curves 2.1 and 2.2 for $x \geq 25000$. For future investigations we recommend to include (an evaluation) the cost of nuclear power within such plots. More importantly, one should uses approximate cost functions that zoom on the range of likely (or possible) nuclear capability of each week.

**Laminarity and decomposability of nuclear plant interactions.** Insightfully, one can draw the hypergraph of constraints CT14-21. To do so, associate a vertex with each nuclear plant, and an hyperedge $(A_m)$ with each constraint. On all instances B, we observe that this hypergraph is roughly the same (after relabelling the plants). The reason is that that sets $A_m$ rely on the geographical positions of french nuclear plants. the hypergraph has a laminar structure (for any 2 pair of edges $Am_1$ and $Am_2$, either $Am_1 \cap Am_2 = \emptyset$ or $Am_1 \subseteq Am_2$ or $Am_2 \subseteq Am_1$). In particular, often constraints CT14 involve only 2 plants with large $Se_m$, hence strongly linking these 2 plants. The connected components of the hypergraph are of size 2,4 or 6 (with a total of more than 50 plants), yielding around 9 components. This implies that demand constraints CT1 are the only one linking these 9 subgroups of nuclear plants in the model, so that just dualizing CT1 allows to split the problem in much smaller parts. We used this non-connectivity of the hypergraph only in rescheduling phase of the ILP: if the solution of the ILP doesn't allow to find feasible production plans for a subset $I^*$ of plants, then we rerun the ILP only with plants connected to at least one plant in $I^*$.

## 2.7 Ideas for future work

When participating in such a challenge, it is hard to explore all mathematical and algorithmic ideas that come to mind. When studying the subject, and even the instance, it is very hard to guess whether an approach will be fast, robust and or competitive. Because of the time constraint, and because we spent a huge proportion of our time debugging and being infeasible, we had to make the choice of extreme simplicity in order to have a chance to produce something within the contest. But the context has changed. We hope that many subproblems will be studied thoroughly, so has to provide insight and subroutines aiming at solving them. We want to share our ideas concerning such issues in this section.

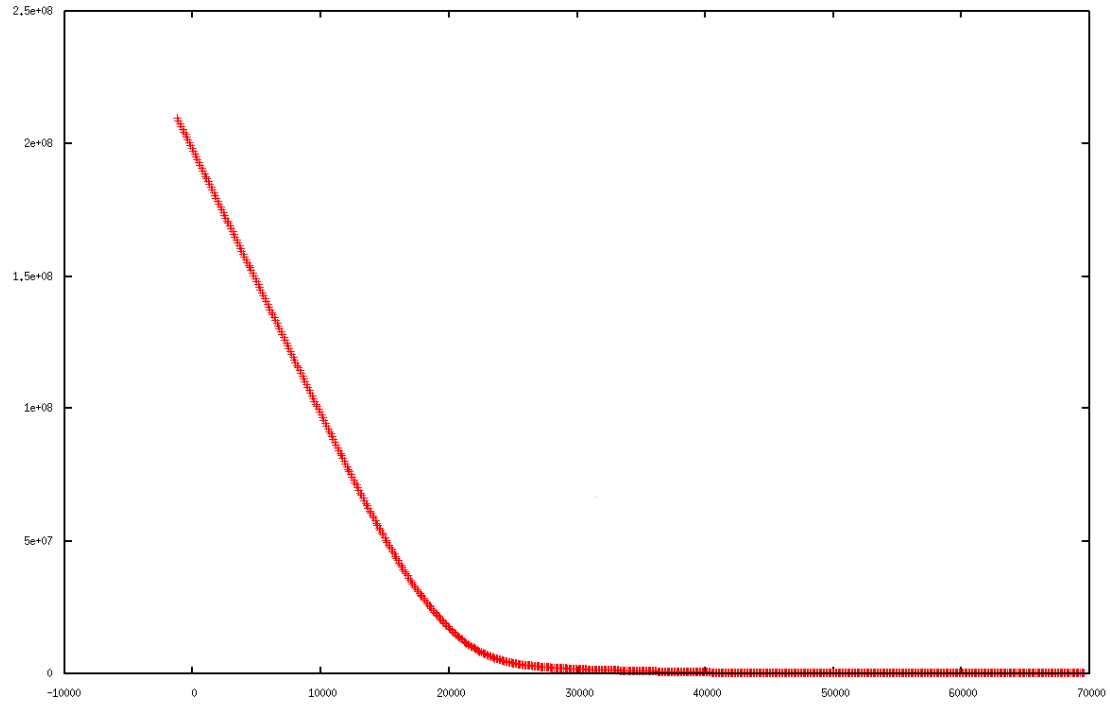



Figure 2.1: Cost of completion by non-nuclear plants as a function of total nuclear availability, averaged on all steps and scenarii of week 63, instance B6.

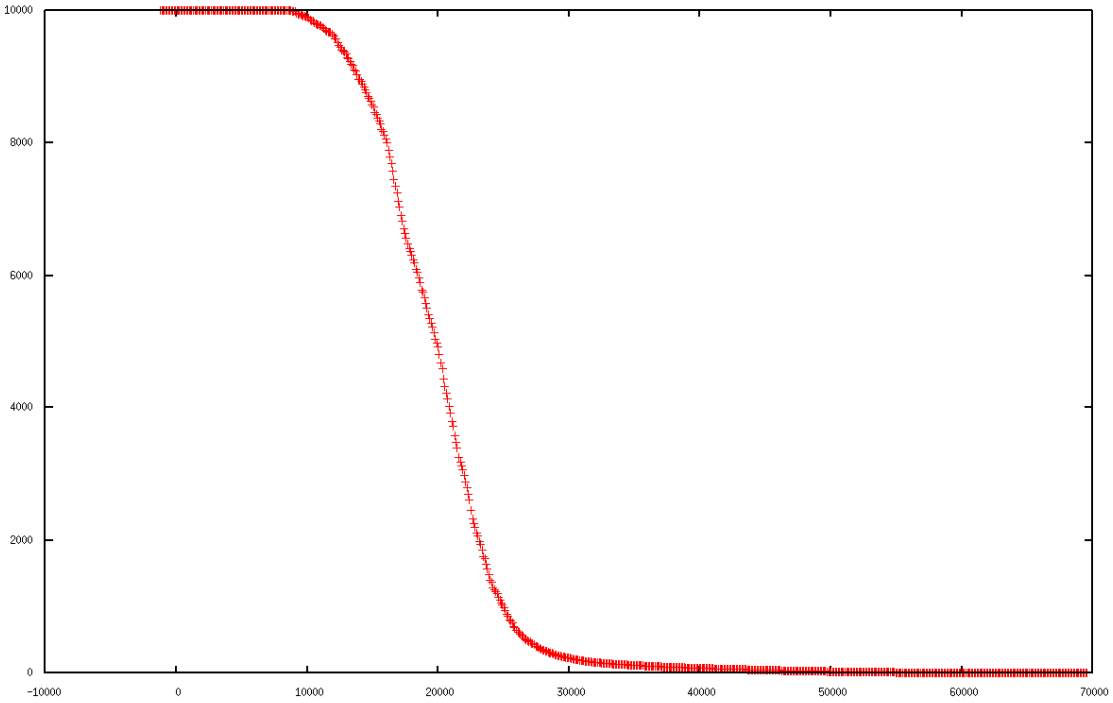



Figure 2.2: Marginal cost of completion by non-nuclear plants as a function of total nuclear availability, averaged on all steps and scenarii of week 63, instance B6.

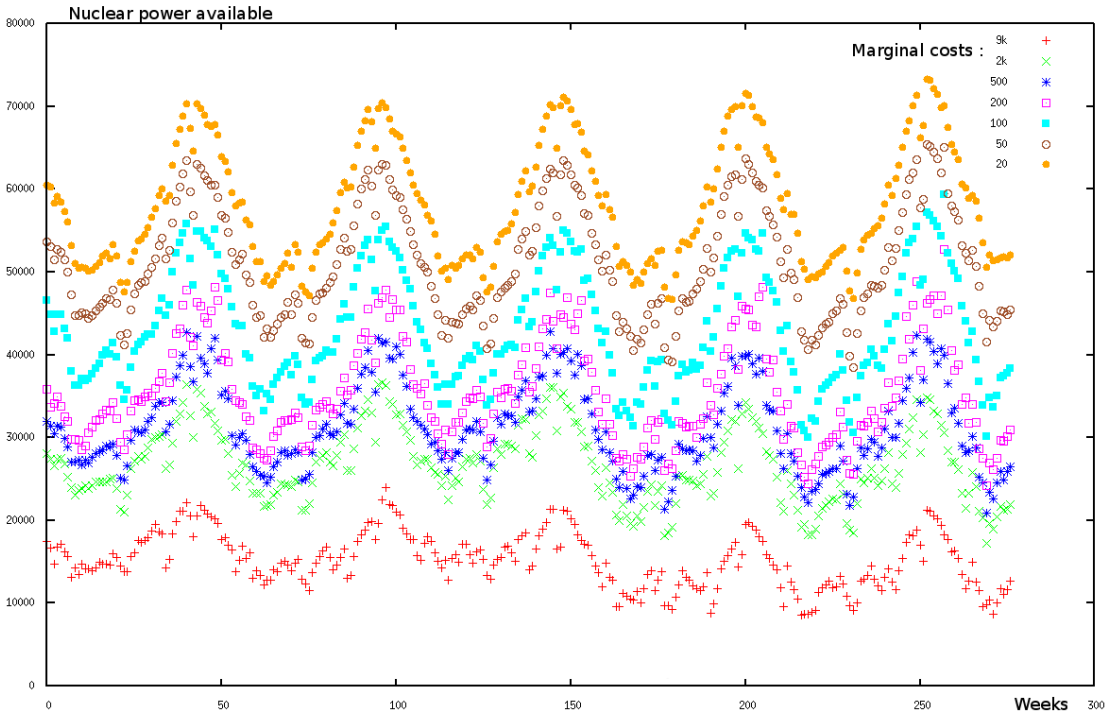



Figure 2.3: Marginal cost of completion by non-nuclear plants as a function of the week (x-axis) and nuclear availability, averaged on all steps and scenarii, instance B6.

The main characteristic of our approach is its hierarchical decomposition nature. This kind of strategy might be considered inefficient and not noble. But we gave some arguments showing that subsequent choices of variables might be evaluated approximately in compact implicit ways. Moreover, since other resolution techniques are successfully addressed by other competing teams, we prefer to stick to our approach in this discussion.

Assuming that the hierarchical structure is preserved, there are several issues that need to be addressed to judge the ability of our decomposition to provide solutions with highly competitive objective value.

### 2.7.1 Improving the ILP

**Implicit refueling and fuel stocks at the end of cycles.** Although simplistic and preliminary, our work challenges the idea that refueling values should be intimately studied with dates of outages. The hypothesis that fuel stock after an outage is limited to only RMAX should be improvable using (one) more dates of outage than only the last one (as done in (2.5)).

**Modulation and profile phases.** Using variables $p_{ikh}$ (production within each cycle) instead of just $p_{ih}$, there might be some ways to express the possibility of modulation. Moreover, assuming (observing ?) that the profile phase is convex, there might be ways to impose it using only linear inequalities, that is, only with upper bounds on the production. Notice that modulation and profile destroy the assumption that production is binary and constant within each week.

**Evaluating objective costs within the ILP.** The problem with the variables $x_{ikh}$ is that they don't allow to reformulate easily (an approximation of) the objective function of the subject. To evaluate the cost of a schedule, we should compute the amount of nuclear power dedicated to each pair $(t, s)$. This last information seems too detailed, because it yields too many variables. One may wish to compute only an approximation of the completion cost by non-nuclear plants for each week (aggregated on steps and scenarii). Such values are plotted in section 2.6. We tried such an approach by approximating, for each week, the function of Figure 2.1 piecewise-linearly. Indeed, we tried both upper and under approximations, using exponential thresholds (like 20,50,100,200,500,2k,9k) on the derivative (*i.e.* marginal costs), as indicated in Figure 2.3. Because the function of Figure 2.1 is convex, minimizing a piecewise linear approximation can be expressed with linear inequalities only. The methods works, indeed, it works so well that the optimal value found by CPLEX was only very few percents away from the exact value of the final solution. However the time taken by CPLEX on instance B9 to find an optimal solution was around 40 minutes. Moreover and more surprisingly, the quality of the solution was not really improved (it was a little better or worse depending on the instances). So we closed this way of research during the challenge. The reason why this nice idea didn't bring improvements in the solution seems to be that something else is too roughly taken into account, and should be improved before thinking about cost evaluation within the ILP.

**Polyhedral study and preprocessing constraints.** We tried to provide to CPLEX strong constraints, but further studies on the linear relaxation of our formulation are required to judge its quality. Distance constraints and CT21 are a priori highly redundant and naively written, bunch of constraints CT21 should be advantageously reformulated in the format of CT20. Assuming we use non-correlating objective functions (such as 2.6), the ILP could be submitted to CPLEX for each connected components one at a time (or in parallel).

**Using constraint based and local search solvers.** After all, why solving the ILP at optimality while we know it is only an approximation of the global problem ? One might wish to apply other resolution paradigms originating from the SAT community like pseudo-boolean programming [132]. Local search [53] seems also a promising way in order to find good solutions faster.

### 2.7.2 Improving the refueling and the power assignment

**Modulation for value.** Our approach doesn't look into the future demand when dealing with a given scenario at a given time step. Obviously, we lose objective value with this simplification.

**Faster greedy procedures.** Because computational time wasn't a key point that needed improvement, we make intensive use of nuclear power affectation in order to check feasibility. We clearly recompute again and again similar affectations when adjusting the refueling values. In some instances, the time to adjust refueling is therefore substantial. In order to accelerate recognizing set of refueling that are infeasible, it would be useful to identify scenarii and step windows that are the most critical. Recomputing only (or first) on these critical points would be natural.

**Smarter view on good refueling.** One thing that we overlooked is that although $r_{ik}$ cannot be increased, fuel level might be far bellow $BO_{ik}$ at the end of cycle $(i, k)$, yet it might be possible to increase $r_{ik-1}$. More generally, our view on refueling would benefit visualizing many outputs and comparing with better solvers to evaluate how much we lose at this point. This idea can be generalized easily, as we explain now.

### 2.7.3 Creating collaboration between solvers of the challenge.

Following our decomposition, one might wish to evaluate each competing solver of the challenge not on the global aspect of the subject, but only in solving subproblems. The following sketch of protocol should provide an easy framework. Given solvers $S_1$ and $S_2$, and given an instance $I$ of the problem, let $S_1$ solve $I$ and obtain solution $I_1$. Gather part of the values in $I_1$ (like $\{ha_{ik}\}$ and/or $\{r_{ik}\}$). Fix these values in $I$ to obtain a more constrained instance $I'$. Run $S_1$ and $S_2$ on $I'$ and compare speeds and values. Optionally, fix some more parts of the solutions obtained and so on.

# Chapter 3

# Modelling and optimization techniques for smart electrical grids

Paula Leite, Leo Liberti, and Nelson Maculan

*Manuscript in preparation*

*There is a lot of talk about Smart (electrical) Grids, and several informal definitions as to what impediments a smart grid should be able to cope with; the literature fails to some extent to provide a formal definition concerning what a smart grid should be able to actually do. Moreover, from the point of view of optimization techniques, no satisfactory survey of the state of the art was written yet. This report attempts to provide such a survey, with the explicit aim to direct subsequent research on this topic towards promising subjects.*

## 3.1 Introduction

The term "smart grid" refers to the design of electrical networks of the next generation, which should address all the troubles raised by the current state of technology. In particular, smart grids should avoid large-scale blackouts generated by single points of failure; should integrate different sources of renewable energy; should be both predictive and reactive with respect to sudden changes in demand; should be scalable, localized rather than centralized, and organized hierarchically; and should allow customers to reduce their costs by introducing flexibility and transparency of billing and accounting procedures. Furthermore, the whole grid should work automatically, even though unsupervised, rely on an advanced informatic infrastructure, and take smart decisions on critical issues.

The first Conference on Innovative Smart Grid Technologies, sponsored by the IEEE Power & Energy Society (PES), hosted by the National Institute of Standards and Technology (NIST) and technically co-sponsored by the IEEE Communications Society (ComSoc), the IEEE Computer Society, the IEEE Power Electronic Society (PELS), the IEEE Signal Processing Society (SPS) and IEEE-USA , was held January 19-21, 2010 in Gaithersburg, Maryland, USA.

Above all, "smart grid" is nowadays a fashionable term which means different things according to the context of study. This survey is motivated by an initial literature review on the state of the art of optimization techniques within the field of smart grids. The result of our initial review was a failure: too many sources of documentation of extremely diverse nature, most of which would not correspond to the standard definition of a "survey". What does exist in the literature is a set papers dealing with

optimization methods addressed to specific aspects of electrical networks; some of these aspects also turn up in smart grids. Most existing papers dealing with smart grids are preliminary, in the sense that they discuss their definition; some of these papers present deeper analyses involving also some optimization techniques. In this report we present a survey of our findings.

There is no standard definition for the term "smart grid". According to the testimony of M.W. Howard, senior VP of the Electric Power Research Institute (EPRI) before the House subcommittee for energy and air quality (May 3, 2007), a smart grid is an "advanced communication and data acquisition system to provide real-time analysis by a distributed computing system that will enable predictive rather than reactive responses to blink-of-the-eye disruptions". According to the testimony of T. Casey before the House committee on energy independence and global warming (Feb 25, 2009), a smart grid is in many ways like an "Internet for Electricity, a network of devices that are monitored and managed with real-time communications and computer intelligence. A Smart Grid discovers grid performance and conditions using intelligent sensors capable of detecting problems and/or opportunities for improvement that are widely distributed on the electric grid. It makes adjustments locally or communicates back to the central control center where it is combined with information from other grid devices and utility systems". According to R. Brown [89], the electricity distribution system of the past is radial and dumb, whereas in the future it will be meshed and intelligent.

The idea of Demand Response (DR) between transmission distribution and consumption parts in a grid supported by advanced metering infrastructure (AMI) is the basis of a smart grid, see [81]. A smart grid needs to include two-way digital technology to control appliance at costumer's houses to save energy, reduce cost and increase reliability and transparency. It overlaps the electricity distribution grid and net metering systems, see [172].

Smart grids are considered to be "green technology" and are more environmentally friendly although often prove to cyber attacks. For the US Dept. of Energy (2006), DR is defined as "charges in electric usage by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized", see [172].

Demand Response (DR) is a change electricity use by end-use customers within a window time in response to some type of incentive or price signal. Measurement and evaluation are very important for DR operations and improvement. If a DR structure operates under a regulated tariff, independent third-party evaluation will commonly be required, see [164]. An effort to develop measurement and verification standards for Demand Response (DR) is important to define DR terms and event timing, and to identify standards methods of evaluating DR performance, see [102].

Smart grids, given their potential for rapid response, are used to demand dispatch resource for plug-in vehicles. Over the next years the use of Internet will be fundamental for the large number of consumer devices: from televisions to household appliances to cars. The Internet will be used to real-time control of demand to help balance generation and load. Optimization models will be useful for organization and designer of grids, and scheduling plug-in vehicles, see [87].

In the rest of this report we shall illustrating what electrical grids actually do, and how optimization methods help their design. For each of the sections below we give a survey of the papers in the literature, focussing on those which we deem the most important.

## 3.2 Basic properties of electrical networks

1. **Mathematical study of very high voltage power networks I: The optimal DC power flow problem [77]**

   The optimal power flow problem involves minimizing the power loss over the lines by setting the voltage and power delivered at the nodes of an alternating current (AC) network. The voltage and

power distribution must obey Kirchhoff's and Ohm's laws and it is subject to certain bounds. This is the first paper of a series dedicated to mathematical study of the optimal power flow problem. A very high voltage approximation is used in the methodology, which is an asymptotic analysis where the small parameter is the inverse of the reference voltage of the network.

Even though the application deals with AC, this paper considers the analogous direct current problem in order to simplify the equations involved. It must be highlighted that these equations retain some of the properties of the original problem.

For this paper the perturbation theory for nonlinear programming is the main mathematical tool. The authors review the equation of the direct current power network and define the problem of minimizing the loss of power over the network. They introduce the high voltage approximation and show that, after a proper scaling, the limit problem is well defined. They also review the mathematical tools from the perturbation theory for nonlinear programming that are needed. They combine these two results in order to state a third. The limit problem is also analyzed and it is shown that it has nonqualified constraints, although there exist multipliers associated with the solution.

The analytic expansion for the optimal value and the solution are shown by the authors and they give physical interpretations of the expansion of the solution. The most important aspect of the result is that it suggests a possible extension to the AC power flow problem.

2. **Mathematical study of very high voltage power networks II: The AC power flow problem [78]**

   This is the second paper of a series dedicated to mathematical study of very high voltage power networks. A very high voltage approximation is used in the methodology, which is an asymptotic analysis where the square root of one of the small parameters is the inverse of the reference voltage of the network.

   The objective of this work is to apply the recently developed perturbation theory in optimization in order to compute the expansion of the optimal power flow problem (OPF).

   It is shown that the active and reactive parts of the limit problem are decoupled and therefore justify the so-called direct current approximation. The authors give an analytic expansion for the solution of the problem and prove the convergence of two decoupled numerical algorithms. For some specific cases they show that the solutions of the equations may be interpreted as critical points of some potentials.

   This paper focus on the power flow problem (PFP) in alternating current networks, because there is already much to say about the expansion of the solution of the equation of PFP without any optimization process.

   The authors show the very high voltage approximation (VHVA) and the convergence of two decoupled numerical algorithms, the classical one and the one of Carpentier in the VHVA setting. They also show that the limit problem has a unique solution, which is regular and has a block-diagonal Jacobian and they justify and improve the DC approximation as a by-product of the computation of the first term of the expansion. This work covers the case in which networks have negligible resistances and also show that the equations of PFP are those of the critical point of certain potentials. The active and reactive equations are shown to derive also from a potential. In the VHVA setting, the solutions of these critical point equations appear to be isolated solutions of the associated optimization problems.

   This paper confirms the existence of an unique solution and, when a set of small parameters is close to 0, it shows its asymptotic expansion. The scaled first-order expansion of the solution has a certain decoupling property: phases are determined by inputs of active power. This aspect justifies the direct current (DC) approximation which relates phases and inputs of active power.

   One can find in the literature several methods based on linearization of the PFP equations based on the observation that, in real world high voltage networks, the difference of phases is small and the value of voltages is approximately constant. This work complete and justify these heuristics methods by a strong mathematical framework.

The high voltage approximation (the combination of the classical approximation with the hypothesis that the nominal voltage is large) allows the DC approximation to be derived and a formula for the first-order variations of voltages that is also a DC-type problem to be proved.

3. **Mathematical study of very high voltage power networks III: The optimal AC power flow problem [79]**

   In this paper, it is shown the way to apply the perturbation theory for nonlinear programming problems to the study of the optimal power flow problem.

   In this paper, the small parameter is the inverse of the square root of the reference voltage of the network and it is called the very high voltage approximation. The authors show how it is possible to obtain, under natural hypotheses, the second order losses expansion and first order solutions expansion. The problem of minimizing losses of active power over a very high voltage power networks is such that the computation of the active and reactive parts are decoupled.

   The high order expansion of the value function, solution and Lagrange multiplier is also obtained and it is shown that the classical direct current approximation may be justified and improved using the framework of very high voltage approximation.

   The paper covers the optimal power flow problem and the limit problem. It also establishes the uniform boundedness of solutions to the family of perturbed problems. The first the first order expansion of solutions and the high order expansions are computed.

   The results show that the limiting problem has a unique and explicit solution. The problem of minimizing losses of active power over a very high voltage power networks may be decomposed into two independent subproblems (active and reactive parts). Each of these subproblems has an interpretation in a DC setting.

   The high voltage approximation combines the classical approximation with the hypothesis that the nominal voltage is large. Combining this approximation with the theory of nonlinear programming with perturbations, it is possible to obtain an expansion of the value function and the problem solution. Also the quadratic subproblem splits into two independent subproblems related to the active variables. This asymptotic analysis justifies the direct current approximation.

4. **Minimizing Effective Resistance of a Graph [161]**

   In a weighted graph, the effective resistance between two nodes is equivalent to the electrical resistance seen between the nodes of a resistor network with branch conductances given by the edge weights. Besides the analysis of electrical networks, many other applications and fields use the effective resistance such as Markov chains and continuous-time averaging networks.

   This work focus on studying the problem of allocating edge weights on a given graph in order to minimize the total effective resistance. It is shown that this is a convex optimization problem and can be solved efficiently either numerically or, in some cases, analytically. It is demonstrated that the optimal allocation of the edge weights can reduce the total effective resistance of the graph by a factor that grows unboundedly with the size of the graph.

   The authors give several interpretations to the effective resistance minimization problem (ERMP). They show that the ERMP is a convex optimization problem, which can be formulated as a semidefinite program (SDP).

   This has several implications, both practical and theoretical. The ERMP can be solved efficiently and the convexity of this problem allows the authors to form necessary and sufficient optimality conditions and several associated dual problems (with zero duality gap). A lower bound on Rtot is obtained, given any feasible allocation of conductances. In an electrical network, Rtot is related to the average power dissipation of the circuit with a random current excitation. The authors use the duality gap in a simple interior-point algorithm for solving the ERMP. They describe several families of graphs for which the solution to this problem can be found analytically, by exploiting symmetry or other structure. This work shows that of all graphs on n nodes, the optimal value of the ERMP is smallest for the complete graph and largest for the path and some numerical examples are shown.

The authors list some variations on the ERMP that are convex optimization problems and can be handled using similar methods: minimizing effective resistance between a specific pair of nodes, minimizing the sum of effective resistances to a specific node and minimizing maximum effective resistance.

It is shown in [327] that resistance functions are concave. In [46], an analysis of the load flow equations provide bounds on the number of stable load flows for a given network topology and set of power injections. The 1990s state of the art in optimal power flow and dispatching is surveyed in [196]. Variational representations of electrical networks by means of factor graphs are discussed in [352].

## 3.3 Economic issues

1. **When to Reap and When to Sow: Lowering Peak Usage With Realistic Batteries [49]**

   Large clients, in some energy markets, are charged for both total energy usage and peak energy usage, which is based on the maximum single energy request over the billing period. The problem of minimizing peak charges was recently introduced as an online problem. In this problem, a battery (assumed to be perfectly efficient) is used to store energy for later use. The authors extend the problem to the more realistic setting of lossy batteries, which lose to conversion inefficiency a constant fraction of any amount charged (e.g. 33%). The authors provide efficient and optimal online algorithms as well as possibly competitive online algorithms. They evaluate the heuristic algorithms on real and synthetic data.

   This work uses three datasets: a regular business day's demand from an actual Starbucks store, a simulated weekday demand of a residential user, and a randomly generated demand sequence.

2. **On Computational Issues of Market-Based Optimal Power Flow [355]**

   The authors study the computational challenges brought up by the deregulation of the electricity market and present new formulations and algorithms for the robust optimal power flow problem (OPF) such as the trust-region based augmented Lagrangian method (TRALM), step-controlled primal-dual interior point method (SCIPM) and constrained cost variable (CCV). They test and compare these formulations with other existing ones applying large-scale power system models.

   The TRALM is more theoretically rigid and the SCIPM is better for real-time applications in terms of computational performance. CCV is proposed by the authors with the objective of improving the scalability of market-based OPF computation. It embeds market-induced piecewise cost functions into inequality constraints as opposed to the objective function.

   The results showed that the methods discussed in this paper are reliable and better than some existing ones in solving large-scale market-based nonsmooth OPFs, however if they are working with non-convex cost, in order to reach global optimal solutions, they must be combined with global optimization techniques.

3. **On the Supply Function Equilibrium and its Applications in Electricity Markets [319]**

   The authors of this paper work with the Supply Function Equilibrium (SFE) as a model of competition on electricity markets. The authors show that through relaxations of supply functions more efficient algorithms can be derived. They work with two examples; the demonstration that continuous equilibrium could be impossible and that it is possible to converge to a linear equilibrium through learning in linear supply system.

## 3.4 Power network vulnerability analysis

1. **The $N - k$ Problem in Power Grids: New Models, Formulations and Computation [65]**

This work focus on finding small vulnerabilities in a power grid, in other words, whether there exists a set of arcs whose removal causes the system to fail. Computational and theoretical results are presented involving a mixed-integer model and a continuous nonlinear model related to this question.

Two models are presented:

(a) a new linear mixed-integer programming formulation. It models the problem as a competition between an "attacker" and a grid controller. The fictional attacker wants to disable the network, and a controller tries to prevent a collapse. This prevention is done by the selection of which generators to operate and the adjustment of generator outputs and demand levels.

(b) a new continuous nonlinear programming formulation. It aims at compactly capturing the interaction between the network structure and its underlying physics.

Although both formulations give better results than a pure enumerational approach, the second formulation appears particularly effective and scalable. The solution to this formulation tend to concentrate the attack on a relatively small number of lines, while at the same time investing small portions of the attack budget on other lines. The two models are consistent as the severity of the attack decrease as the scale increases for both of them.

2. **Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization [125]**

This paper gives an overview of a complex systems approach to large blackouts of electric power transmission systems caused by cascading failure. The authors study series of blackouts, their statistics and dynamics, with global models and not just particular blackouts. The data from worldwide blackouts show that a power law governs the frequency of large blackouts. This fact makes the risk of large blackouts consequential and it is consistent with the power system being a complex system designed and operated near a critical point. Power system overall loading or stress relative to operating limits is a key factor affecting the risk of cascading failure. The models for power system blackout and cascading failure show critical points with power law behavior as load is increased. The authors suggest that power system operating margins evolve slowly to near a critical point and this idea is confirmed by a power system model. The engineering responses to blackouts that improve the system, the economic pressure to maximize the use of the grid and a steady increase in electric loading cause the slow evolution of the power system.

The ultimate motivation to the analysis and understanding of blackouts is the primordial importance of the electrical infrastructure to society. Even though large blackouts dont occur very often, the observed statistics show that their risk cannot be discarded. This can be explained by the fact that the increase of the blackout cost is much greater than the decrease of the occurrence probability of a blackout in a power law manner as the size of the blackout increases.

The authors cover blackout mechanisms; discuss the evidence for a power law in blackout probability and the consequences for blackout risk; summarizes abstract and power system models of cascading and discuss the observed critical points in these models. They also expand the discussion to describe and model a power system slowly evolving with respect to engineering and economic forces and discuss some initial consequences for blackout mitigation.

3. **Using Graph Models to Analyze the Vulnerability of Electric Power Networks [190]**

In this work the electric power delivery networks are modeled as graphs. The authors study two power transmission grids (Nordic and the western states in the U.S.). They calculate values of topological (structural) characteristics of the networks and compare their error and attack tolerance (structural vulnerability). They also perform a structural vulnerability analysis of a simple and fictitious electric power network, using different strategies to decrease the vulnerability of the system.

The article focuses on electric power delivery systems. These systems can often be represented as networks. The structure of networks is mathematically described in terms of graphs. For an electric power grid, the vertices can be power plants, stations and power users, and the edges

power lines. The authors want to study how the performance of networks is affected by the removal of vertices and edges, to compare the structure of different networks, and to analyze how the change of structure affects the vulnerability of networks.

The authors introduce concepts and results from graph theory, study two electric power transmission grids (Nordic and the western states), calculate values of topological characteristics of the networks and compare their error and attack tolerance with two theoretical reference model networks. They also present a discussion of the pros and cons of the use of graph models when analyzing the vulnerability of electric power networks.

The work shows that the Nordic grid is more scattered than the grid of the western states. Numerical simulations of the structural vulnerability demonstrate that the two electric power networks exhibit similar disintegration patterns, both when it comes to random failures and deliberate attacks. All the studied networks disintegrate considerably faster when the vertices are removed deliberately than randomly. This article shows that in a structural vulnerability analysis of a fictitious network with a simple structure a wider spectrum of threats and hazards can be represented as the removal of vertices and edges.

4. **Cyber Security and Power System Communication — Essential Parts of a Smart Grid Infrastructure [140]**

   This paper highlights the role of the combination of cyber security and power system communication systems in the infrastructure of a smart grid. It gives an explanation of the facts of the PSC systems with partly vulnerable structure.

   Many power utilities have installed SCADA/EMS and industrial control systems, which were opened up from the design phase, but they had very limited security incorporated in the system solutions. The openness of the system was quite appealing, but now there is the problem with information and IT security. This fact is very serious and must be taken into account for system daily operation and control by each utility.

   PSC and cyber security issues are crucial parts of the information infrastructure, such as a smart grid system. The integrated SCADA/EMS systems and administrative office IT environments must now be separated. Cyber security issues become increasingly important, when dealing with smart grids.

## 3.5 Optimization of power restoration after a failure

1. **Using mixed-integer programming to solve power grid blackout problems [64]**

   The authors work with the prevention of large-scale cascading blackouts in power transmission networks through optimization problems. They consider many thousands of scenarios of externally caused damage and present computation with networks with up to 600 nodes and 827 edges.

   The paper highlights how large-scale failures of national power systems, such as the ones occurred in Brazil in 1999 and in the US and Canada in 2003, affected people over wide geographical areas and caused a great economic impact. The problem of minimizing the chances of catastrophic blackouts in large networks is very complex and involves engineering, economic and political issues. The authors present two distinct models and algorithmic tools to deal with the combinatorial difficulties related to this problem.

   The first model is concerned with the decision of which lines must have their capacity increased in order to guarantee that in each scenario all flows are within capacities after removing some edges from the network. The authors first describe a natural Mixed Integer Programming formulation of the problem which can be large and impractical depending on the number of scenarios. Because of that they subsequently present a projected formulation and an algorithm for solving the problem via this projected formulation using Branch & CUT. The second model considers the dynamics of a cascade, and assumes that no action is taken during the course of a cascade.

Both models are essentially static and aim at reinforcing a network so that in the event of some edge outages the network is better able to avoid a blackout. In a real time problem one desires to stop a cascade from becoming large-scaled.

2. **Analytical tools for power system restoration - conceptual design [370]**

   This work presents a conceptual framework for executing monitoring and assessment functions during system restoration. The authors identify analytical tools in system security monitoring that can undergo a few modifications in order to be applied in system restoration monitoring. A survey is done to cover the analytical tools that can be applied in restoration assessment and the authors propose a knowledge-based expert system architecture to accomplish this task. The importance of this research is highlighted by the increasing threats of blackouts and the fact that systems are operating closer to their limits nowadays.

3. **An Efficient Algorithm for Distribution Network Restoration [290]**

   The authors have previously proposed in [291] the use of a combination of expert system and mathematical programming approach for power system restoration. In this more recent work they propose a fast and efficient algorithm to solve the distribution system restoration based on a mathematical programming approach. They show that with the division of the problem in two steps and with the addition of restoration strategies the computation time can be reduced 30% if compared to [291]. These two steps are: the maximization of available power to the area affected, and the minimization of the amount of unsupplied energy. They are solved by a mixed integer programming approach.

## 3.6 Robust power network design

1. **From Hierarchical to Open Access Electric Power Systems [199]**

   This paper shows the modeling, monitoring and control of electric power systems through large-scale dynamic systems' point of view. The author summarizes the current hierarchical operations, assesses underlying assumptions, presents the challenge of operating electric power systems over very broad ranges of system conditions as an open sensing, estimation, and control problem. This work also shows a vision of an information-based multilayered Dynamic Energy Control Protocols (DECPs) framework for facilitating evolution into open access just-in-time (JIT) and just-in-place (JIP) electricity services of the future.

   Toady's system topologies can be described as large power plants, often located far from load centers; utilities supplying their customers without depending much on the neighboring utilities; and utilities interconnecting for reliability reasons, to help each other during major equipment outages. This causes the electric power grid to have several voltage levels, converted from one to the other several transformers, leading to an extra high voltage (EHV) meshed transmission backbone network, and local lower voltage networks closer to the consumers. The overall objective of traditional electric power system is simple and consists of minimizing total cost subject to reliability constraints. On the other hand it can be very complex to be implemented if considering it as a single problem of decision making for very large-scale dynamic systems.

   In today's electric power industry there is a growing concern and focus toward the reliability of the grid through robust design. The author has formalized more complex models in order to review the current models. The also presented the class of models for hierarchical automatic generation control (AGC) in the United States and automatic voltage control (AVC) in Europe.

   While the hierarchical and open access operating models are qualitatively different in their complexity, the author has approached both of them by recognizing the underlying structures in these systems. He affirms that independently from which type of the operating models a lot depends on the dynamics, monitoring, and control of the interaction variables between different levels of the system.

The open-access future paradigm is probabilistic in its basic nature and allows a truly dynamic interaction between performance objectives, sensing, monitoring, control, and coordination. Its success depends on aligning uncertainties with the right industry levels and on the adequate management of interactions under uncertainties. The author suggests that it is essential for Researchers should work closely with its industry partners to begin transforming both the backbone and local distribution electric power networks into future smart systems.

The main goal of this work can be summarized as introduction of a model-based decision logic at the layers of the distribution systems, and the simulation of the performance of future automated distribution grids with options to choose between distributed generation, price responsive demand, and dynamic islanding of customers with special needs, a system capable of responding to attacks on both hardware and software.

2. **Economic Impacts of Advanced Weather Forecasting on Energy System Operations [371]**

   The energy consumption and the performance of industrial and residential facilities are strongly influenced by weather conditions mostly because energy is used more often for conditioning large and exposed spaces and for generating industrial-scale utilities.

   This work analyzes the impacts at the decision-making hierarchy of the power grid of adopting advanced weather forecasting systems and this is done through case studies. It is shown that numerical weather prediction (NWP) models can provide high-precision forecasts and uncertainty information that can greatly enhance the performance of planning, scheduling, energy management, and feedback control systems. The forecasting capabilities of the Weather Research and Forecast (WRF) model are also analysed in several application domains.

   The authors cover an integrative study of weather forecasting, uncertainty quantification, and optimization-based operations. They first analyze the impact of increasing the forecast horizon on energy management operations for a multi storage hybrid system and on the temperature control of a building system then they seek to understand under which conditions weather forecasts are beneficial and compare the forecasting capabilities of an empirical modeling approach with those of the Weather Research and Forecast (WRF) model.

   The work suggests that costs and power losses can be reduced by incorporating accurate forecasts and that WRF forecasts significantly outperform those obtained with empirical models, specially for medium and long-term horizons.

3. **Voltage-Drop-Constrained Optimization of Power Distribution Network Based on Reliable Maximum Current Estimates [145]**

   This paper aims at achieving tree-shaped power distribution networks optimum design, considering the voltage drop effect. The width of the power lines is adjusted in a way such that the network occupies the minimum possible area under specific voltage drop constraints. The authors base the optimization on precising the maximum current estimates through statistical means (extreme value theory). In the experiments, the authors designed power grids for different topologies and voltage drop tolerances in a typical benchmark circuit.

   The drop of the effective voltage level supplied may degrade the performance of the circuit or even cause faulty logic signals and circuit malfunction. The power lines have been built very wide in order to reduce their resistance and avoid significant voltage level drop.

   The design of power networks that occupy smaller areas as possible, but satisfying constraints on IR-drop requires the knowledge of the maximum possible currents flowing at any time through the lines of the network.

   This paper covers the problem of power network design as a constrained optimization problem, gives the basic statistical results for the estimation of maximum currents using extreme value theory and describes the ensuing procedure for the optimum selection of power line widths.

   The power grid optimization procedure consists of two main parts:

   - obtaining the estimates of maximum currents along each line of interest, using the statistical concepts

- the application of a proper optimization algorithm.

4. **Cooperative Sensor Networks for Voltage Quality Monitoring in Smart Grids [66]**

   The authors propose the use of self-organizing sensors networks in order to obtain a fully decentralized voltage quality monitoring architecture. In these networks, each node can access the performances of the monitored site and also the global performances of the monitored grid section, and by accomplishing this each node may automatically detect local voltage quality anomalies. Furthermore, the system operator can assess the system voltage quality index for each section of the grid without requiring a central fusion center acquiring and processing all the node acquisitions which makes the architecture highly scalable, self-organizing and distributed.

## 3.7 Local and domestic management

1. **Smart Demand-Side Energy Management Based on Cellular Technology [2]**

   The power grids built more than a century ago are not prepared for modern requirements. One problem that must be highlighted is the lack of information on the consumption of electricity provided to users and network operators. To upgrade the existing grids to smarter ones is very expensive and to developing countries it is particularly difficult to benefit from these solutions. The authors focus on a system accessible to these countries which intends to reduce energy consumption and wastage. This system uses cell phones to display information and allow consumers to control appliances in their homes.

   The authors show an example from Malta, a micro-state with an area of only 316- square-kilometers where IBM and its partners are investing in the monitoring of electricity use in real-time and setting variable tariffs that reward customers who decrease the consumption of power. It is a $91 million project.

   The GSMA (GSM Association) has estimates that in 2007 there were approximately 270 million mobile phone users in Africa. These numbers make the cell phones an ideal instrument to provide feedback about electricity consumption to consumers in this part of the globe. A GSM Modem at the user's house will send data from a smart meter about electricity consumption to a server. This information will then be displayed on the consumers' mobile phone. The server can also receive control signals (to switch an appliance ON or OFF ) from the user's phone. If a control signal is received, the GSM modem will then issue the command to the appliance via the smart meter.

## 3.8 Electrical vehicles

1. **Optimal Transition to Plug-in Hybrid Electric Vehicles in Ontario-Canada Considering the Electricity Grid Limitations [171]**

   The transport sector is growing at a fast rate and demands more energy every day. It also increases air pollution and the emission of greenhouse gases. In Canada the transport sector represents almost 35% of the total energy demand and is the second highest source of greenhouse gas emissions. Therefore, alternative fuels are so important specially to this country.

   Hybrid Electric Vehicle (HEV) technology can be an alternative. It may reduce the consumption of gasoline and the greenhouse gas emissions. However it depends on a single hydrocarbon fuel source and has range limitations. The Plug-in Hybrid Electric Vehicle (PHEV) is similar to HEV but has a larger onboard battery and a plug-in charger and does not depend on a single fuel source. It can run up to 30 Km on battery power alone. It is important to highlight that this is a technology still under development. There are a few concerts regarding energy storage costs, range and durability.

A multi-interval DC optimal power flow (OPF) model with loss factor considerations was developed by the authors, based on the 10 zone model of Ontario's electricity system, during base-load time intervals, in order to find the maximum volume of extra load due to PHEVs that can be added to Ontario's electricity network between the years of 2009 and 2025. With a piecewise linearization of power losses, the authors have a Mixed Integer Linear Programming (MILP) problem. The objective of this model is to minimize the total cost.

The result of the study showed that by 2025 500.000 PHEVs can be introduced into Ontario's transport sector without affecting the reliability of the grid and without the need of a new infrastructure to cover the electricity requirements.

## 3.9 Information systems for smart grids

1. **Information services for smart grids [218]**

   The Interconnected and integrated electrical power systems are very complex and may face some challenges such as an older structure, scarce availability of generation near load centers, distributed resources, dynamic reactive compensation, congestion management, reliability coordination, supply and cost of natural resources for generation, etc. The objective of this work is to overcome the grid management challenges through new information services.

   This paper highlights that information semantics is a new approach for handling complex systems and can be applied to discover knowledge and support the decisions of grid operators by focusing on making machines more interactive. These software tools are based on semantics and ontologies that use web-addressable sensors. Information semantics supports systems, data, documents, agents and spans ontologies, knowledge representations, semantic web, natural language processing, and knowledge management.

   The operational environment of today's power system is basically composed by many distributed tools and components. The interfaces of the components must work in a "plug and play" concept similar to the hardware used in the substation automation. There has to be data sharing amongst applications, and integration of functions of various tools for the modeling. One task that should start now is the integration of the functions of the individual services so that they can perform more complicated functions. The concepts suggested in this work such as the utilization of information semantics and integrating data will be required to sustain social and environmental obligations for the electric utility industry and will be important to guarantee reliability and efficiency of the new smart grid.

## 3.10 Scheduling and running the grid

1. **Multi objective optimalisation of smart grid structure [213]**

   This article focus on the future smart grid structure optimization. Investment can be made in solutions such as: smart gauges, line breakers, redundant secondary network or the total old network could be replaced. The author highlights that the lack of energy sources and the aging of the existing power system calls for a smart grid structure.

   Power system should supply energy with the guarantee of high reliability for the customer, low user end price of delivery and sustainability for the global supply.

   The MultiObjective optimization described in this work is concerned with the minimization of the delivery price and also the maximization of the reliability and sustainability. The authors maximize the security and the sustainability by limited investment sources.

   There a several ways to raise the network reliability, such as, by adding some control (smart) gauges, by building in some remotely controlled line breakers, by adding some secondary gauges, by making a redundant network part and by replacing the old line

The results of this work showed that putting as many intelligent smart gauges as possible it brings advantages compared to the replacement of the old network. It also showed that reliability of the network can be raised by remote controlled line breakers and the self healing auto reconfiguration. There is no need to change the existing voltage level. A low level network duplication makes possible the local REN generation (PV) and works as an local micro grid or Uninterruptible Power Supply, reducing considerably the amount of nondelivered- energy.

2. **Optimal operation of dispersed generation under uncertainty using mathematical programming [174]**

   This paper presents a mathematical model including different kinds of dispersed generation units with respect to their technical characteristics and the optimization technique used. The authors apply stochastic programming extensions to mixed-integer linear programs. An advanced scenario decomposition algorithm is used to cope with the resulting stochastic integer programs that are too big for standard solvers.

3. **Smart Grid Design for Efficient and Flexible Power Networks Operation and Control [281]**

   There is a growing need of a more adaptive and secure power system for the US. Consumers also demand more power quality and reliability of supply and delivery. Because of that there is an increase interest in Smart Grids. This paper shows Smart Grid intelligent functions and also a special case for the development of Dynamic Stochastic Optimal Power Flow (DSOPF) technology as a tool for the Smart Grid design.

   The Smart Grid aims at providing grid observability, creating controllability of assets, enhance power system performance and security, and reducing costs of operations, maintenance, and system planning. The authors of this paper develop the Smart Grid functions that improve interactions of agents (e.g. telecommunication, control, and optimization to achieve adaptability, self-healing, efficiency and reliability of power systems)

   There are some challenges to be faced ahead such as the lack of predictive control signals to operate devices and lack of energy storage devices which affect deployment of smart devices. Funding is also necessary to the development of new technologies in this area for the Smart Grid.

   The use of Adaptive Dynamic Programming (ADP) as a method to handle complex power system problems with prediction under uncertainty has been enhanced by the recent advances in computer science and the Artificial Intelligence and Applied Mathematics communities.

   ADP combines the experience of Optimal Control and Dynamic Programming (DP) in reaching optimal utility functions and also judgments performance function that evaluates the reliability and performance of the first stage of ADP optimization. These methods have been used in aerospace, power systems, and in other management portfolio.

   The author has been developing DSOPF. This computational algorithm has controllability and interoperability, reliability, adaptability and sustainability, anticipatory behavior and affirmation of Security.

   This paper proposes the development of a Research and Learning Center which will provide researchers with the study of standards and the comparative study on modeling, predictive, and control tools for electric power industry. It proposes future research, for example, to the development of advanced techniques for measuring peak load reductions and energy-efficiency savings from smart metering, demand response, distributed generation, and electricity storage systems; to the investigation of the feasibility of a transition to time-of-use and real-time electricity pricing and to the development of algorithms for use in electric transmission system software applications.

4. **Proactive Energy Management for next-generation building systems [372]**

   This paper shows a day-ahead prediction of disturbances affecting costs and efficiency which can lead to an accurate forecast of the daily electricity demand profile. This is done through the on-line solution of mixed-integer nonlinear programming problems. The disturbances involve weather conditions, fuel prices, heat gains, and utility demands. One advantage of this proactive

framework is that it can capture occupant behaviors and market interactions by using agent-based models.

## 3.11 Open problems in smart grids

Minimization of the time for sensing and responding to signals in an electrical grid [146, 295, 90, 202, 12, 14, 11, 13, 9]

**[202] Grid of the Future - A. Ipakchi and F. Albuyeh**

Different and antagonist factors modulate the most profound changes the power energy industry has ever faced: environmental targets, demand response (DR), support to plug-in electric vehicles, distributed generation and storage. The existing electricity distribution grid faces a challenging transition: from counting on quasi steady and individually reliable generation sources, dealing with localized decisions and manual operations, to the automatic and dynamic integration of a variety of intermittent wind, solar and other renewable generation sources, the "smart grid" scenario.

Novel businesses and regulatory drivers are being set, having environmental issues as the main modulators of such new regulations: renewable portfolio standards (RPS)[1] deal with greenhouse gases limitation, DR and energy conservation measures, defining obligations towards increasing the production of energy whilst providing a minimum percentage of the electricity production from renewable energy sources. About 32 USA states, accounting for approximately half of the energy sales in the whole country, had established RPS targets by August 2008, with California setting 20% of the energy production coming from renewable sources by 2010, and 33% by 2020.

As reliability of energy supply is a strategic factor, the evolution of regulatory mechanisms is mandatory and the implementation of advanced metering infrastructures (AMI) targeting DR is an ongoing process. Wind and solar generation are the main actors in the new electricity generation portfolio, with wind being the fastest growing segment (and the most intermittent form of electricity generation). In a more complex scenario, even batteries from plug-in electric vehicles could act as energy providers in the role of distributed storage, a very important subsidiary service in the smart grid. While wind turbines present a power curve in which typical operation includes a fast ramp, i.e., minutes to be up (in contrast to seconds to shut down), solar cells have the potential of very localized operation, while still not in conditions of economically competing with wind farms.

The main issues posed by these renewable energy sources can be summarized as:

- **Transmission**; placement a non-trivial problem since good sites for wind farms, mainly, and solar plants, may be in areas having low capacity, or non existing, transmission lines;
- **Distribution**; new energy flow patterns, including the introduction of bidirectional flows, will need new and smart protection and control mechanisms;
- **Interconnection standards**; unification and power scaling will impact interconnection standards;
- **Operation**; dealing with intermittent energy generation is a challenge;
- **Forecasting and scheduling**; accurate climate forecasting and smart energy scheduling and dispatch is needed.

Due not only to environmental, but also to economical reasons, plug-in electric vehicles are definitely penetrating the market. Based on the existing specifications of such vehicles, the impact of having electric cars charging during evening and night will eventually more than double the average residential

[1] *Renewable Electricity Standard* (RES) at the United States federal level, and *Renewables Obligation* in the UK.

load during such period. Moreover, as existing distribution systems were designed decades ago, the concurrent charging of vehicles in a localized situation, e.g., a parking lot, may overload and/or unbalance the distribution system.

The deployment of AMI will soon allow much more dynamic tariffs to take place, together with market -based prices, so that end-use devices, e.g, smart chargers, would be able to decide about a consumption style by having a view of the grid status and associated dynamic prices. On the other hand, the system operator will be able to monitor and control demand though price signals. The NIST/Gridwise Architecture Council is a mentionable initiative towards integrating demand-side resources with distribution and transmission operations.

Given its geographically distributed nature, the smart grid impacts all operational and enterprise information systems. The combined application of distributed intelligence, congestion management strategies and market-based dynamic pricing is a almost certain feature of the future IT systems involved in the smart grid. In fact, one can draw a parallel between (i) the transition from existing electric generation, transmission and distribution apparatus the to the new smart grid scenario, and; (ii) the transition from the existing, mostly stand-alone, operational and enterprise information systems to the integrated (operational and enterprise aspects), distributed, secure and agile computational system expected in the deployment of a smart grid.

A cloud computing approach seems a natural target since it could not only minimize the need for the development of new customized applications with the same functionality of the ones already in use, but also allow for an integrated view of all information and operation systems via a web-based middleware. Such software design style would accommodate, directly or indirectly, the incremental transition in course as well as any new operational requirements needed by smart grid.

# Chapter 4

# Combinatorial optimization for electric vehicles management

Nora Touati-Moungla and Vincent Jost

*To appear in Renewable Energy and Power Quality Journal*

*Growing concerns about environmental quality of cities are calling for sustainable road transportation technologies. Electric Vehicles (EV), for public and private transport, can contribute significantly to the lowering of the current pollution levels. However, the EV use is currently facing several weaknesses among which are: limited driving range, high cost and overall limited efficiency. This report aims at specifying some key contributions of combinatorial optimization for an efficient electric vehicles management.*

## 4.1 Introduction

Distribution and transportation systems have been intensively studied in the operations research literature [347]. 73% of all oil consumed is Europe is used in transport and road transport accounts for 25% of CO2 emissions of the overall transport activity. From both an environmental and energy points of view, the introduction of EV should be a first priority for the reduction of primary energy consumption.

Although higher concerns are the opportunities EV provided in terms of efficiency and flexibility in the use of energy, the EV use however is currently facing several weaknesses among which are: (1) The low energy density of batteries compared to the fuel of combustion engined vehicles, (2) EV often have long recharge times compared to the relatively fast process of refueling a tank and (3) The scarcity of public charging stations.

Electric Vehicles Management (EVM) is a relatively recent problem, its purpose is to expedite the establishment of a costumer convenient, cost-effective, EV infrastructure. Inspite the relevance of the problem, a few small research communities in this field work on some aspects of this problem. In this work, we discuss some important issues of this problem and show how CO tools can be used for solving some challenging subproblems.

Routing of EV is a major aspect of EVM, it consists of designing routes for maximizing the autonomy of vehicles, efficient EV routing plays a major role for encouraging EV use. We discuss in this report this problem, we present a mathematical formulation of the *energy shortest path problem* and the *energy*

*routing problem* and we expose some relationships between these problems and other well-known routing problems.

Limited driving distance between battery charges is a fundamental obstacle to broad consumer adoption of EV. In order to eliminate this fundamental disadvantage and increase consumer acceptance and usage of EV, a sufficient number of charge stations is required. The objective is to establish a charging network that is conveniently placed in familiar places to meet consumers needs. The localization of EV charge stations is known in CO field as the *facility location problem*, we define here this problem and we expose some models proposed in the literature.

The Self-Service Electric Vehicles (SS_EV) is a key concept for developing urban clean mobility. The "free" use of EV would cause either overflow or shortage of vehicles at some stations at some times of the day. This system require a redistribution of EV over the stations, this problem is generally modeled as a special *pick-up and delivery problem*, we discuss in this report some of its characteristics and resolution methods.

The remainder of this report is as follows. In section 4.2 we discuss routing of EV and its relationships with some routing problems in the literature. In section 4.3 we address the charge stations localization. We expose redistribution of EV in the context of a SS_EV system in section 4.4 and we conclude this report in section 4.5.

## 4.2 Routing

The rising and highly-variable cost of fuel, increase the importance of efficient vehicle routing. Key advantages of EV are their ability to recover braking energy to be restored to the battery (regenerative braking) and their zero energy consumption in congested environment. Efficient EV routing is critical to the operational profitability and customer satisfaction of EV use, especially in light of highly concurrence with fuel of combustion engined vehicles.

### 4.2.1 Energy shortest path problem

The Energy Shortest Path Problem (EnSPP) consists of finding an optimal origin-destination route for EV with rechargeable batteries taking into account energy recuperation during deceleration phases. This problem can be modeled with a directed graph $G = (N, A)$, where $N = M \cup \{s, t\}$, $M$ is the set of nodes representing the junctions, $s$ and $t$ the source and destination nodes respectively and $A$ represents the set of arcs. With each arc $(i, j)$ is associated a positive (resp. negative) value $e_{ij}$ indicating consumption (resp. gain) of energy on arc $(i, j)$. The battery of each vehicle has a maximum capacity $C$ and cannot be discharged below zero. The EnSPP consists of finding optimal origin-destination routes for EV by maximizing the vehicles battery charge at the destination node.

Very scarce works interest to this problem in the literature. In [22], the authors formalize the EnSPP problem as a generalization of the shortest path problem with hard constraints (which impose that the battery cannot be discharged below zero) and soft constraints (which impose that the battery cannot store more energy than its maximum capacity). A generic shortest path algorithm was proposed to solve the problem. The authors developed a prototypic software system for energy efficient routing where data is obtained by combining geospatial data (of OpenStreetMap) and elevation data (of the NASA Shuttle Radar Topographic Mission).

The formulation of the EnSPP is given as follows:

$$\min \quad u_t \tag{4.1}$$

$$\text{s.c.} \quad \sum_{i \in M} x_{si} = \sum_{i \in M} x_{it} = 1 \tag{4.2}$$

$$\sum_{j \in N} x_{ij} - \sum_{j \in N} x_{ji} = 0, \forall i \in M \tag{4.3}$$

$$x_{ij}(u_j - u_i - e_{ij}) \geq 0, \forall (i,j) \in A \tag{4.4}$$

$$0 \leq u_i \leq C, \qquad \forall i \in N \tag{4.5}$$

$$x_{ij} \in \{0,1\}, \qquad \forall (i,j) \in A \tag{4.6}$$

$$u_i \in \mathbb{N}^+, \qquad \forall i \in N \tag{4.7}$$

where variables $x_{ij}$ indicate whether arc $(i, j)$ is used or not and $u_i$ is the battery remaining storage capacity at node $i$. Constraints (4.2)-(4.3) define the path structure for the vehicle. Energy constraints (4.4) express the fact that if arc $(i, j)$ is used, the battery remaining storage capacity at node $j$ can be greater than $u_i + e_{ij}$ (in the case where $u_i + e_{ij} < 0$).

The EnSPP is a generalization of the shortest path problem with time windows [121] where $e_{ij}$ represents the travel time on arc $(i, j)$ and the time window at each node is equivalent to $[0, C]$. Hard and soft constraints are equivalent to time windows constraints; Hard constraints express the fact that a vehicle must arrive at the costumer before the end of time window and soft constraints express the fact that if a vehicle arrives before the beginning of the time window, it wait at no cost. In a more general case [22], the EnSPP is a generalization of the Shortest Path Problem with Resource Constraints (SPPRC) [203] where resource consumption represents energy consumption, the amount of available resource is equal to $C$, and the residual resource at node $i$ is equivalent to $u_i$.

### 4.2.2 Energy vehicle routing problem

The Vehicle Routing Problem (VRP) [347] is an important problem in the fields of transportation, distribution and logistics. It can be modeled with the graph $G$ described at section 4.2.1, where nodes represent costumers. A set $K$ of identical vehicles are available, each one has a maximum capacity $Q$. The VRP consists of finding a set of minimum cost origin-destination routes, such that each costumer $i \in M$ is visited by exactly one vehicle to satisfy a specific demand $d_i$. The total customers demand satisfied by the same vehicle must not exceed the vehicle capacity.

Distribution activities cause major problems with regard to noise and air pollution, so, there is a need for EV in urban distribution activities. We introduce here the VRP where vehicles are electric ones, called Energy Vehicle Routing Problem (EVRP), its formulation is given as follows:

$$\min \quad F(x,u) \tag{4.8}$$

$$\text{s.c.} \quad \sum_{k\in K}\sum_{j\in M} x_{ij}^k = 1, \forall i \in N \tag{4.9}$$

$$\sum_{(i,j)\in A} d_i x_{ij}^k \leq Q, \forall k \in K \tag{4.10}$$

$$\sum_{i\in M} x_{si}^k = \sum_{i\in M} x_{it}^k = 1, \forall k \in K \tag{4.11}$$

$$\sum_{j\in N} x_{ij}^k - \sum_{j\in N} x_{ji}^k = 0, \forall i \in M, \forall k \in K \tag{4.12}$$

$$x_{ij}^k(u_j^k - u_i^k - e_{ij}) \geq 0, \forall (i,j) \in A, \forall k \in K \tag{4.13}$$

$$0 \leq u_i^k \leq C, \forall i \in N, \forall k \in K \tag{4.14}$$

$$x_{ij}^k \in \{0,1\}, \forall (i,j) \in A, \forall k \in K \tag{4.15}$$

$$u_i^k \in \mathbb{N}^+, \forall i \in N, \forall k \in K \tag{4.16}$$

where $x_{ij}^k$ indicate whether arc $(i,j)$ is used or not by vehicle $k$ and $u_i^k$ is the battery remaining storage capacity of vehicle $k$ at node $i$. The objective function (4.8) can be considered as the sum of the remaining battery storage capacity of all vehicles ($\sum_{k\in K} u_t^k$) or the maximum remaining battery storage capacity of all vehicles ($\max_{k\in K} u_k^k$) at the destination node. Constraints (4.9) ensure that each costumer is visited exactly once. Constraints (4.10) ensure that demand of each route is within the capacity limit of the vehicle serving the route. Constraints (4.11)-(4.12) are path constraints. Constraints (4.13)-(4.14) ensure compatible remaining storage capacity at each node for each vehicle.

As showed in section 4.2.1, the EVRP can be considered as a generalization of the vehicle routing problem with time windows [214] where the objective is to minimize the total travel time of all vehicles or the minimization of the makespan (minimization of the maximum total travel time). Additionally, this problem is a special case of a much studied problem in the literature, the Pick-up and Delivery Problem (PDP) [301], where pick-up are not hard constraints: partial pick-ups are allowed and pick-up is not performed when the vehicle is full. Acceleration (resp. deceleration) phase can be considered as a delivery (resp. pick-up) operation.

Lets $\hat{G} = (\hat{N}, \hat{A})$ be a directed graph where $\hat{N}$ is the set of nodes and $\hat{A}$ is the set of arcs. $\hat{G}$ is constructed using the graph $G$ as follows: let be $P(i)$ the set of predecessor nodes of node $i \in N$, $|P(i)| = l^i$. Each node $i \in N$ is duplicated into $l^i$ nodes $\hat{N}_i = \{i^1, ..., i^{l^i}\}$, so, $\hat{N} = \bigcup_{i\in N} \hat{N}_i$. For each new node, we assign an incoming arc from one predecessor node:

$$\hat{A} = \bigcup_{i\in N}\{\hat{A}(j^1, i^1) \bigcup \hat{A}(j^2, i^2) ... \bigcup \hat{A}(j^{l^i}, i^{l^i}) \mid j^1, ... j^{l^i} \in P(i)\}.$$

where $A(j^1, i^1) = \bigcup_{k\in\hat{N}_{j^1}}\{(k, i^1)\}$. For each node $i \in N$, we associate to the duplicated nodes $i^1, ..., i^{l^i}$ the weights $q^1 = e_{j^1 i}, ..., q^{l^i} = e_{j^{l^i} i}$. Node $i \in \hat{N}$ represents a pick-up (resp. delivery) node if $q_i < 0$ (resp. $q_i > 0$).

Optimizing the vehicle braking routes, through maximization of final remaining energy, can result in an increase of travel time. To make energy vehicle routing effective, the travel time has to be taken into account in the optimization model. As our knowledge, no work considers the travel time in energy routing problem. Such multiobjective problem can be solved using adaptation of multiobjective routing methods [209, 339].

In practice, the road links have different combinations of energy Consumption/Recuperation (C/R) levels and delays, associated with road furniture such as traffic lights and roundabouts, and road topography and geometry such as inclines. This causes energy C/R variations (resulting from acceleration

and deceleration) over links with the same road category and distance. Therefore, instead of constant values, more realistic considerations of energy C/R have to be established. On one hand, energy C/R can be a function of speed variations over links. A number of works in the literature interest to model/estimate travel speed [188], these works can be exploited to estimate energy C/R. On another hand, energy C/R has to be considered as stochastic instead of constant values for assessing the risk of running out of energy before arriving at the destination. No work in the literature tackle these issues yet.

## 4.3 Facility location

One of major barriers to EV success have arisen in limited number of refueling stations, due to the limited range of EV, the establishment of an infrastructure to facilitate EV refueling is a pressing concern. Due to the large capital costs involved in infrastructure investment, economic factors are very important in determining the number and location of stations. Therefore, studies must work to provide a theoretical basis for station deployment, such as with a facility location model, to economically and efficiently serve EV trips.

Location problems in general are spatial resource allocation problems dealing with one or more service facilities serving a spatially distributed set of demands. The objective is to locate facilities to optimize a spatially dependent objective like the minimization of average travel time or distance between demands and facilities. The most studied practical problem in this context concerns hydrogen station location. In [273], general criteria are proposed for identifying effective locations for early hydrogen stations, (1) close to areas with high traffic volume, (2) in places to provide fuel during long distance trips, (3) at high profile locations to increase public awareness, and (4) in places that are accessible to individuals who are buying their first fuel-cell vehicle. These criteria are also necessary in EVM to ensure consumer confidence in the reliability of the refueling network.

In order to develop an overall broad perspective on the facility location literature so that we can appropriately model EV specific problem we present below three basic models categories that are related to refueling-station location problem [358]:

*Models based on the maximum covering location problem*: In [189] was introduced the "flow capturing location model" to represent the goal of locating facilities to serve passing flows. Any flow that uses a path that passes through the location of the facility was considered to be captured. The problem is formalized as a maximum covering location problem, it consists of locating $p$ facilities so as to "capture" as many of passing flows as possible. This model was extended in [226] to the "flow refueling location model" for considering a flow refueled only if an adequate number of stations are spaced appropriately along the path.

*Models based on the set covering problem*: In [356], the author propose a facility location model for economically site slow recharging stations at scenic spots in order to conveniently serve all demand from single origin-destination journeys (via multi-stop refueling), with drivers using electric scooters at the destination area. This model was extended in [357] for accommodating multiple origin-destination trips. The purpose of the proposed models is to optimally site the refueling stations to cover overall passing flows on the paths of interest.

*Models based the maximum covering/shortest path problem*: This model [107] consists of the minimization of the path length between a given origin-destination pair of nodes and the maximization of the total demand covered by the facilities located at the nodes in the path. This model is extended in [48] to accommodate multiple origins and destinations.

## 4.4 EV redistribution

A SS_EV system represents an efficient alternative to use the private (petroleum fuel) vehicles in terms of resource sharing, cost and flexibility. This concept is currently very popular in many countries and should bring two advantages: a net reduction in the number of vehicles (and therefore parking spaces), and a reduction in air and noise pollution with the use of EV. In this system, a costumer can pickup one vehicle in a station, use it for a while and return it to another (or the same) station. Indeed, either an overflow or a shortage of vehicles can happen at one or more stations at some times of the day. This system must guarantee the availability of the vehicles at the pick-up points.

Redistribution problem consists of redistributing the vehicles among the stations in order to maximize their availability to the costumers. Redistribution is provided by a fleet of limited capacity tow-trucks located at various depots on the network. This problem can be naturally modeled as a pickup and delivery problem. In [130], this problem is viewed as a generalization of the pickup and delivery problem, formulated as a mixed integer problem and solved using a number of resolution methods including constraint programming, Lagrangian relaxation and a modified $A^*$ heuristic. The problem seems to be difficult to solve although for small instances, this is due to the splitup of pickups and deliveries, the small capacities of tow-trucks and non simple paths (a tow-trucks can return to a station already serviced).

In [96], the authors propose a balancing technique which consists of switching from an unfavorable state to a favorable one. A favorable state is defined as the distribution of vehicles in the stations that guarantees that the system can reach a given large horizon with the highest probability, assuming that no balancing action is conducted, and an unfavorable state is defined as the distribution of vehicles in the stations that guarantees that at least one of the station will run out of vehicles or will be overload before a given small horizon with a probability greater than or equal a given threshold. In [99], the authors consider the issue of recharge problem where an optimal level of charge that makes the vehicle available for costumers is defined. Because of the battery limited range, two major aspects have to be taken into account in the redistribution system, the physical and energy availability of vehicles at stations.

## 4.5 Conclusion

The succeeding of a transition from a conventional gasoline based transportation system towards a sustainable way of transportation, depends on a quite number of critical factors. The substitution of conventional vehicles through electric and/or hybrid vehicles involves economical, environmental and social aspects. The purpose of EVM is to encourage the transition to EV use and to expedite the establishment of a convenient, cost-effective, EV infrastructure that such a transition necessitates. Whereas the development of EV through battery autonomy grow extensively, very little research is dedicated to EV routing, recharge stations localization and vehicles redistribution. We discussed in this report these major aspects which represent challenging optimization problems.

Routing consists of designing routes for EV for maximizing the autonomy of vehicles, efficient EV routing plays a major role for encouraging EV use. This problems is much close to a number of well-studied routing models in the literature.

Compared to fossil fuels, batteries have a low energy content per weight ratio. This limits the radius of action of EV, making them mostly suitable for urban traffic and special transport applications, such as shuttle services. Siting sufficient refueling stations along intercity highways is central issue to ensure the completion of the long-distance travel demands in order to foster the use of EV.

There is an increase on utilization of video surveillance, global positioning systems and communication equipment installed on vehicles. A topic of interest is the development of a global mobility information systems through which EV charging station customers can locate EV charging stations prior

to starting on a trip, and can maximize their mileage and minimize their risk of running out of electricity.

Problems discussed in this report are subproblems of EVM that can be treated using combinatorial optimization tools. The global critical problem in the EV promotion is a chicken-and-egg infrastructure dilemma [359]: consumers will be reluctant to purchase vehicles until a sufficient number of refueling stations has been installed, while vehicle manufactures will not produce vehicles that consumers will not buy, and fuel providers will not invest in a new energy infrastructure until there is sufficient demand for it. Therefore, it is likely that governments will need to play a significant role in promoting any change to alternative fuels, although public support alone will not ensure the success of this transformation.

Summarizing the paper discussions, it can be stated that EVM systems including the three topics discussed in this report are a crucial path towards EV mobility. The key functionalities assigned to such a system are gain of driving range, the flexibility of recharging vehicles and vehicles availability in a self-service EV system.

# Chapter 5

# Optimal running and planning of a biomass-based energy production process

Maurizio Bruglieri, Leo Liberti

*Published in Energy Policy,* **36***:2430-2438, 2008*

*We propose mathematical programming models for solving problems arising from planning and running an energy production process based on burning biomasses. The models take into account different aspects of the problem: determination of the biomasses to produce and/or buy, transportation decisions to convey the materials to the respective plants, and plant site locations. Whereas the "running model" is linear, we propose two "planning models", both of which are Mixed-Integer Nonlinear Programming problems. We show that a spatial Branch-and-Bound type algorithm applied to them is guaranteed to converge to an exact optimum in a finite number of steps.*

## 5.1 Introduction

Producing energy derived from fossil carbon-based fuels is proving costly to both the environment (in terms of pollution) and society (in terms of monetary investment). As the prices of crude oil increase, governments and other institutions are researching the most cost-efficient ways to produce energy from alternative sources [201]. One of the most popular contendents is energy produced by biomasses of several kinds [306]. In [270] the competitiveness of biomass-based fuel for electrical energy opposed to carbon-based fuel is examined using a mathematical programming model. Among the advantages of this type of energy production, there is the potential for employing wasted materials of biological origin, like used alimentary fats and oils, agricultural waste and so on. A factory producing energy with such materials would benefit from both the sales of the energy and the gains obtained by servicing waste [18]. In [149] a mathematical program is proposed to localize both energy conversion plants and biomass catchment basins in provincial areas. Other mathematical models for specific biomass discrete facility location problems are developed in [155] and [109]. A model that combines detailed energy conversion plant optimization with energy/heat transportation cost is given in [334].

This report describes an optimization problem arising from the deployment of such an energy production process in central Italy. This involves several processing plants of different types (for example,

a liquid biomass plant, a squeeze plant and a fermentation-distillation plant). Some of these plants (e.g. liquid biomass plant) produce energy; others (e.g. the fermentation-distillation plant) produce intermediate products which will then be routed to other plants for further processing. There are several possible input products (e.g. agricultural products, biological waste), obtained from different sources (e.g. direct farming or acquisition on the markets) at different unit costs. Apart from the energetic output, there may be other output products which are sold in different markets (e.g. bioethanol obtained from the fermentation-distillation plant and sold in the bioethanol market). See Fig. 5.1 for a typical process flowsheet.

There are in fact three optimization problems relating to this description. The first (and simplest) is that of modelling the production process as a net gain maximization supposing the type of plants involved and the end product demands are known. The second is that of deciding the type of plants to involve in the process to maximize the net gain, subject to known end product demands. Although post-optimal sensitivity analysis may be used to gather hints on how to improve the process, an optimization model provides the ultimate process planning tool. The second model is in fact a simple variant of the first, in that we simply let some of the parameters of the first (linear) problem be decision variables in the second. The third problem is an evolution of the second, taking into account plant installation costs, some features of electricity production plants, and transportation issues [265, 266, 280]. We remark that we only carried out computational experiments on the first and second model, since the practical needs of the industry that commissioned the research were limited. The third model is supplied to show that this modelling approach can be extended to a more complicated and realistic setup.

Section 5.2 describes the model relating to the production process when the plant types are known ("running model"). Section 5.3 describes the model relating to the process planning ("planning model"), an exact mixed-integer linear reformulation thereof, and shows that an application of a standard spatial Branch-and-Bound algorithm (e.g. [320, 333]) yields a finitely convergent exact method. In Section 5.4 we discuss the application of the production process and planning models to a real-life case. Section 5.5 discusses the third model, and Section 5.6 concludes the paper.

## 5.2 Optimizing the production process

Modelling a flowsheet as that presented in Fig. 5.1 presents many difficulties. Notice that the products can be inputs, intermediate, outputs, or both (like alcohol, which is both an output product and an intermediate product). Likewise, processes can be intermediate or final or a combination (like the fermentation-distillation plant). Consider also that the decision maker may choose to buy an intermediate product from a different source to cover demand needs, thus making the product a combination of intermediate and input. Of course the input products may be acquired or produced at different locations and at different prices. Moreover, each flow arrow has an associated transportation cost. The time horizon for the optimization process is one year.

The central concept in our model is the process site. A *process site* is a geographical location with at most one processing plant and/or various storage spaces for different types of goods (commodities). A place where production of a given commodity occurs is represented by a process site with a storage space. Thus, for example, a geographical location with two fields producing maize and sunflowers is a process site with two storage spaces and no processing plant. The fermentation-distillation plant is a process site with no storage spaces and one processing plant. Each output in Fig. 5.1 is represented by a process site with just one storage space for each output good. In this interpretation the concepts of input, output and intermediate products, and those of intermediate and final process, lose importance: this is appropriate because, as we have emphasized earlier, these distinctions are not always well-defined. Instead, we focus the attention on the material balance and on the transformation process in each process site. Furthermore, we are able to deal with the occurrence that a given commodity may be obtained at different costs depending on whether it is bought or produced directly.

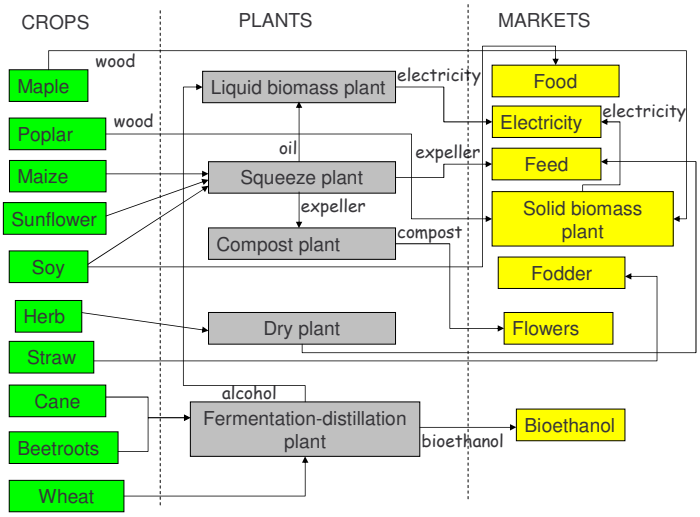



Figure 5.1: A typical process flowsheet.

We represent the process sites by a set $V$ of vertices of a graph $G = (V, A)$ where the set of arcs $A$ is given by the logistic connections among the locations. To each vertex $v \in V$ we associate a set of commodities $H^-(v)$ which may enter the process site, and a set of commodities $H^+(v)$ which may leave it. Thus, for example, the fermentation-distillation plant is a process site vertex where $H^-$(fermentation-distillation plant) = {cane, beetroots} and $H^+$(fermentation-distillation plant) = {alcohol}. Furthermore, we let $H = \bigcup_{v \in V}(H^-(v) \cup H^+(v))$ be the set of all commodities involved in the production process, and we partition $V = V_0 \cup V_1$ into $V_0$, the set of process sites with an associated processing plant, and $V_1 = V \backslash V_0$.

Fig. 5.2 is the graph derived from the example in Fig. 5.1.

The following parameters define the problem instance:

- $c_{vk}$: cost of supplying vertex $v$ with a unit of commodity $k$ (negative costs are associated with output nodes, as these represent selling prices; a negative cost may also be associated to the input node "waste", since waste disposal is a service commodity);
- $C_{vk}$: maximum quantity of commodity $k$ in vertex $v$;
- $\tau_{uvk}$: transportation cost for a unit of commodity $k$ on the arc $(u, v)$;
- $T_{uvk}$: transportation capacity for commodity $k$ on arc $(u, v)$;
- $\lambda_{vkh}$: cost of processing a unit of commodity $k$ into commodity $h$ in vertex $v$;
- $\pi_{vkh}$: yield of commodity $h$ expressed as unit percentage of commodity $k$ in vertex $v$;
- $d_{vk}$: demand of commodity $k$ in vertex $v$.

It is clear that certain parameters make sense only when associated to a particular subset of vertices, like e.g. the demands may only be applied to the vertices representing the outputs. In this case, the corresponding parameter should be set to 0 in all vertices for which it is not applicable.

The decision variables are:

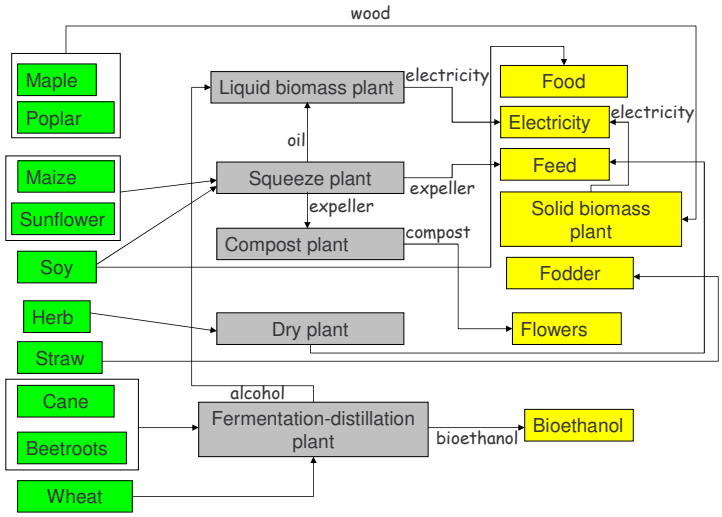



Figure 5.2: The graph derived from the example in Fig. 5.1.

- $x_{vk}$: quantity of commodity $k$ in vertex $v$;
- $y_{uvk}$: quantity of commodity $k$ on arc $(u, v)$;
- $z_{vkh}$: quantity of commodity $k$ processed into commodity $h$ in vertex $v$.

Since the output demands are known *a priori*, we would like to minimize the total operation costs subject to demand satisfaction. There are three types of costs:

- cost of supplying vertices with commodities:

$$\gamma_1 = \sum_{k \in H} \sum_{v \in V} c_{vk} x_{vk};$$

- transportation costs:

$$\gamma_2 = \sum_{k \in H} \sum_{(u,v) \in A} \tau_{uvk} y_{uvk};$$

- processing costs:

$$\gamma_3 = \sum_{v \in V} \sum_{k \in H^-(v)} \sum_{h \in H^+(v)} \lambda_{vkh} z_{vkh},$$

so the objective function is

$$\min \sum_{i=1}^{3} \gamma_i(x, y, z). \tag{5.1}$$

We need to make sure that some material conservation equations are enforced in each process site where a plant is installed:

$$\sum_{k \in H^-(v)} \pi_{vkh} z_{vkh} = x_{vh}, \ \forall v \in V_0, h \in H^+(v). \tag{5.2}$$

Notice that these constraints do not actually enforce a conservation of mass, for in most processing plants a percentage of the input quantities goes to waste; but it is nonetheless a conservation law subject to the yield properties of the particular transformation process of the plant.

Secondly, the quantity of processed commodity must not exceed the quantity of input commodity in each vertex:

$$\sum_{h \in H^+(v)} z_{vkh} \leq x_{vk}, \ \forall v \in V_0, k \in H^-(v). \tag{5.3}$$

Furthermore, we need the quantity of input commodity in each vertex to be consistent with the quantity of commodity in the vertex itself, and similarly for output commodities:

$$\sum_{u \in V:(u,v) \in A} y_{uvk} = x_{vk}, \ \forall v \in V, k \in H^-(v) \tag{5.4}$$

$$\sum_{u \in V:(v,u) \in A} y_{vuh} = x_{vh}, \ \forall v \in V, h \in H^+(v). \tag{5.5}$$

Finally, we have the bounds on the variables:

$$d_{vk} \leq x_{vk} \leq C_{vk}, \ \forall v \in V, k \in H \tag{5.6}$$

$$0 \leq y_{uvk} \leq T_{uvk}, \ \forall (u,v) \in A, k \in H \tag{5.7}$$

$$z_{vkh} \geq 0, \ \forall v \in V, k \in H^-(v), h \in H^+(v) \tag{5.8}$$

and some fixed variables for irrelevant vertices:

$$x_{vk} = 0, \ \forall v \in V_1, k \in H \backslash (H^-(v) \cup H^+(v)) \tag{5.9}$$

$$y_{uvk} = 0, \ \forall (u,v) \in A, k \in H \backslash H^-(v), \tag{5.10}$$

$$y_{uvk} = 0, \ \forall (u,v) \in A, k \in H \backslash H^+(u). \tag{5.11}$$

The main advantage to this model is that it can be easily extended to deal with more commodities and plants in a natural way, by adding appropriate vertices or changing the relevant $H^-(v), H^+(v)$ and related parameters.

This problem is a Linear Program, and can be solved by using one of several LP solvers (e.g. CPLEX [200]).

## 5.3 Planning the production process

In this section, we suppose no processing plants are yet present at the process sites. At each process site $v \in V_0$, we wish to install an appropriate processing plant chosen from a set $P(v)$ of possible plants (e.g. there may be different types of liquid biomass plants, each having different yield levels on the input commodities). We therefore wish to make decisions as regards the plant installation, feasible with the material balance constraints as in Section 5.2, which minimize the total operation costs.

We re-define the parameters $\lambda, \pi$ to make them dependent on a processing plant $p$ as follows:

- $\lambda_{vkhp}$: cost of using plant $p$ to transform a unit of commodity $k$ into commodity $h$ in vertex $v$ (this includes a per-unit estimate of the initial investment costs for building the plant);
- $\pi_{vkhp}$: yield of commodity $h$, using plant $p$, expressed as unit percentage of commodity $k$ in vertex $v$.

We consider the following additional binary decision variables:

$$w_{vp} = \begin{cases} 1 & \text{if plant } p \text{ is installed in vertex } v \\ 0 & \text{otherwise} \end{cases}$$

Moreover, the node capacities $C_{vk}$ and arc capacities $T_{uvk}$, which are considered as parameters in the previous model, are to be considered as decision variables instead, bounded above and below by relevant values.

The objective function (5.1) changes in the $\gamma_3$ term, which becomes:

$$\gamma_3' = \sum_{v \in V} \sum_{k \in H^-(v)} \sum_{h \in H^+(v)} \left( \sum_{p \in P(v)} \lambda_{vkhp} w_{vp} \right) z_{vkh}.$$

The material conservation constraints (5.2) become:

$$\sum_{k \in H^-(v)} \left( \sum_{p \in P(v)} \pi_{vkhp} w_{vp} \right) z_{vkh} = x_{vh}, \ \forall v \in V, h \in H^+(v). \tag{5.12}$$

The following constraints enforce consistency on the assignment variables (we remark that (5.13) allows a process site to host no plant at all):

$$\sum_{p \in P(v)} w_{vp} \leq 1, \ \forall v \in V_0, \tag{5.13}$$

$$\sum_{p \in P(v)} w_{vp} = 0, \ \forall v \in V_1. \tag{5.14}$$

Finally, constraints (5.3)-(5.8) are also part of the formulation.

### 5.3.1 Solution of the problem

The model described in Section 5.3 is a Mixed-Integer Nonlinear Programming problem (MINLP) with nonconvex terms in both the objective function and the constraints. Problems of this type are solved either by employing heuristic methods, like Multi Level Single Linkage (MLSL) [229, 255] or Variable Neighbourhood Search (VNS) [183, 254], or by using an $\varepsilon$-approximate method called spatial Branch-and-Bound (sBB) [333, 5, 341], which provides a proof of $\varepsilon$-optimality. sBB algorithms are Branch-and-Bound (BB) type algorithms (i.e. searches on trees where each node represents a restriction of the problem to a particular subdomain, the union of all the subdomains being the entire search space) where branching is possible on continuous variables appearing in nonlinear terms (branching on binary variables is carried out by fixing the variable at 0 in the left subnode and at 1 in the right subnode). Bounding is obtained by solving a suitable convex relaxation of the problem restricted at the current node's variable ranges. We assume that the branching scheme ensures that all binary variables are chosen for branching within a finite time limit (such a scheme is readily available by e.g. branching on each binary variable in turn). In general, when $\varepsilon = 0$, the sBB algorithm has no finite termination guarantee. Applied to this particular problem, however, the sBB yields a finitely terminating exact method, as shown in Thm. 5.3.2.

At each BB node we obtain a lower bound by solving a linear relaxation of the problem, built as follows.

1. Distribute products over sums, so that all the bilinear terms can be expressed as $w_{vp} z_{vkh}$ (also see [340]).

2. Replace each bilinar term $w_{vp}z_{vkh}$ by a new added variable $\zeta_{vkhp}$ called a *linearization variable*. More precisely, the bilinear terms in $\gamma_3'$ in the objective and in constraints (5.12) should be replaced by the corresponding linearization variable, yielding:

$$\gamma_3'' = \sum_{v\in V}\sum_{k\in H^-(v)}\sum_{h\in H^+(v)}\sum_{p\in P(v)} \lambda_{vkhp}\zeta_{vkhp}$$

and

$$\sum_{k\in H^-(v)}\sum_{p\in P(v)} \pi_{vkhp}\zeta_{vkhp} = x_{vh},\ \forall v \in V, h \in H^+(v). \tag{5.15}$$

Naturally, to keep the reformulation exact, we must add the bilinear constraints

$$(\zeta_{vkhp} = w_{vp}z_{vkh}),\ \forall v \in V, k, h \in H, p \in P(v) \tag{5.16}$$

to the formulation. These are called *defining constraints*. The usefulness of this step is that it isolates the nonconvex terms in (5.16).

3. Replace (5.16) by their convex envelopes:

$$\begin{aligned}
(\zeta_{vkhp} &\geq 0),\ \forall v \in V, k, h \in H, p \in P(v) && (5.17)\\
(\zeta_{vkhp} &\geq z_{vkh} + z^U_{vkh}(w_{vp} - 1)),\ \forall v \in V, k, h \in H, p \in P(v) && (5.18)\\
(\zeta_{vkhp} &\leq z^U_{vkh}w_{vp}),\ \forall v \in V, k, h \in H, p \in P(v) && (5.19)\\
(\zeta_{vkhp} &\leq z_{vkh}),\ \forall v \in V, k, h \in H, p \in P(v) && (5.20)
\end{aligned}$$

where $z^U_{vkh}$ is a tight upper bound to $z_{vkh}$ for each $v \in V, k, h \in H$. Constraints (5.17)-(5.20) are known as McCormick envelopes [271]. The relaxed problem is a mixed-integer linear relaxation of the original problem.

4. A linear (and hence convex) relaxation of the problem is readily obtained by relaxing the integrality constraints on the binary variables.

The linear relaxation thus derived is further tightened by adding Reformulation-Linearization Technique (RLT) cuts as in [4, 329, 328] by multiplying constraints by appropriate variables and then linearizing the resulting bilinear terms, as detailed below:

- constraints (5.3) by bound factors $w_{vp}$ and $(1 - w_{vp})$ for all $p \in P(v)$;
- constraints (5.13) and (5.14) by variables $z_{vkh}$ for all $k, h$, to obtain:

$$\begin{aligned}
\sum_{p\in P(v)} \zeta_{vkhp} &\leq z_{vkh},\ \forall v \in V_0, k, h \in H && (5.21)\\
\sum_{p\in P(v)} \zeta_{vkhp} &= 0,\ \forall v \in V_1, k, h \in H. && (5.22)
\end{aligned}$$

Constraints (5.22) are a particular subclass of RLT constraints ([329, 328]) called *reduction constraints* [250, 251, 249], with very interesting properties. In particular, although the bilinear defining constraints (5.16) are not in the formulation, it can be shown that in consequence of (5.21) and (5.22), a certain subset of them still hold at the relaxed solution. Constraints (5.21) are normal level-1 RLT constraints.

To sum up, the linear relaxation at each sBB node consists in minimizing $\gamma_1 + \gamma_2 + \gamma_3''$ subject to (5.15) and (5.3) with the RLT cuts derived from them, (5.4), (5.5), (5.6)-(5.8), (5.9)-(5.11), (5.13) and (5.14) with the reduction constraints (5.21) and (5.22) derived from them, and the McCormick envelopes (5.17)-(5.20). We shall call this linear relaxation $\bar{L}$.

**5.3.1 Proposition ([328], Prop. 8.11)**
*Let $W \subseteq \mathbb{R}^n$ and $Z \subseteq \mathbb{R}^m$ be the two nonempty polytopes (in variables $w$ and $z$ respectively) described in the planning model above, and consider the set $\Omega = \{(w, z, \zeta) \mid w \in W \wedge z \in Z \wedge$ (5.17)-(5.22)$\}$. Then, for any $(\bar{w}, \bar{z}, \bar{\zeta}) \in \Omega$, if either $\bar{w}$ is a vertex of $W$ or $\bar{z}$ is a vertex of $Z$, constraints (5.16) hold.*

**5.3.2 Theorem**
*As long as the branching scheme ensures that all binary variables are chosen for branching within a finite time limit, an sBB algorithm applied to the MINLP problem of Section 5.3 ((5.3)-(5.8),(5.12)-(5.14), excluding the linearization constraints of Sect. 5.3.1) converges to an exact solution or is shown to be infeasible in a finite amount of time.*

*Proof.* Because of the assumption in the branching scheme, after a finite amount of running time in the sBB algorithm, a node $N$ will be reached where all binary variables are fixed at either 0 or 1. Let $\bar{P}_N$ be the lower bounding LP at node $N$. $\bar{P}_N$ may be infeasible, unbounded or feasible with optimal solution $x^N$. If $\bar{P}_N$ is infeasible, the current Branch-and-Bound tree branch is pruned at the node $N$. If $\bar{P}_N$ is unbounded, the MINLP is also unbounded and the sBB terminates. Otherwise, if $\bar{P}_N$ is feasible, since the $w$ variables have values 0 or 1 by virtue of the branching process, $w$ will be at a vertex of its feasible polyhedron (which is a sub-polytope of $\{w_{vp} \in [0,1] \mid v \in V \wedge p \in P(v) \wedge \sum_{p \in P(v)} w_{vp} \leq 1$ if $v \in V_0$ and 0 othw.$\}$). By Prop. 5.3.1, this implies that $x^N$ is feasible w.r.t. the bilinear defining constraints (5.16). Thus, $x^N$ is also an upper bounding solution in the original problem and the node is fathomed, which implies that the current Branch-and-Bound tree branch is pruned at the node $N$. This also shows that the node is pruned even with $\varepsilon = 0$. Therefore, no branch of the Branch-and-Bound tree can be infinitely long, which proves finite convergence. To show exactness, we remark that since the algorithm was shown to converge even with $\varepsilon = 0$, the solution it provides is exact. Lastly, if no node yields a feasible solution $x^N$, the problem is infeasible. □

**5.3.3 Proposition**
*Let $L$ be the linear relaxation $\bar{L}$ subject to the integrality constraints on the $w$ variables. $L$ is an exact Mixed-Integer Linear Programming (MILP) reformulation of the planning model.*

*Proof.* Let $(w^*, z^*, \zeta^*)$ be an optimal solution of $L$. Since $w$ are either 0 or 1, again by Prop. 5.3.1 we have that $\forall v \in V, k, h \in H, p \in P(v)$ $(\zeta^*_{vkhp} = w^*_{vp} z^*_{vkh})$, which shows that the optimal solution of $L$ is feasible in the original planning model. □

By Prop. 5.3.3, we can also solve the planning model with a MILP solver (e.g. CPLEX [200]).

## 5.4 Computational experience

The main driving force for this report was a real-life instance of the production model occurring in the Marche region of Italy. The owners of an agricultural ground currently producing beetroots and wheat wanted to switch to a more diversified scheme which could provide enough biomass to fuel an energy production plant. This instance gave rise to an LP model with fewer than 100 variables and constraints, the solution performance details are totally irrelevant. The instance was solved by CPLEX 10.1 [200] to optimality. Most of the material used for this computational study can be found at `http://www.lix.polytechnique.fr/~liberti/bioenergy`.

As for the planning model, we considered two separate sets of instances. The first one is based on the real-life production instance mentioned above, modified to contain more production sites and potential plants at each site. The second one consists of randomly generated instances; see Table 5.1 for details.

We solved the instances in Table 5.1 to optimality using several solvers.

| Instance name | $\lvert V_0\rvert$ | max$\lvert P(v)\rvert$ | avg$\lvert P(v)\rvert$ | Vars | (where 0-1) | Constrs | Noncvx terms |
|---|---|---|---|---|---|---|---|
| planning1 | 2 | 4 | 3.50 | 41 | 7 | 31 | 20+13 |
| planning2 | 10 | 3 | 1.50 | 228 | 9 | 161 | 34+18 |
| planning3 | 10 | 3 | 1.50 | 228 | 9 | 161 | 34+18 |
| planning4 | 10 | 2 | 1.40 | 227 | 8 | 161 | 32+16 |
| planning5 | 4 | 3 | 2.00 | 165 | 7 | 106 | 24+14 |
| planning6 | 4 | 3 | 2.00 | 165 | 7 | 106 | 24+14 |
| planning7 | 4 | 2 | 1.75 | 164 | 6 | 106 | 22+12 |
| planning8 | 2 | 3 | 2.50 | 39 | 5 | 31 | 14+10 |
| planning9 | 9 | 3 | 1.33 | 198 | 5 | 139 | 8+8 |
| planning10 | 9 | 2 | 1.11 | 195 | 2 | 138 | 4+4 |
| rnd10 | 2 | 6 | 3.50 | 179 | 6 | 131 | 72+42 |
| rnd14 | 1 | 3 | 3.00 | 249 | 3 | 154 | 18+2 |
| rnd16 | 1 | 8 | 8.00 | 158 | 8 | 90 | 16+16 |
| rnd20 | 3 | 12 | 5.67 | 913 | 17 | 493 | 4+0 |
| rnd23 | 4 | 9 | 6.50 | 690 | 26 | 406 | 60+50 |
| rnd24 | 5 | 9 | 5.40 | 2379 | 26 | 1293 | 0+173 |
| rnd26 | 7 | 9 | 6.14 | 177 | 42 | 101 | 28+7 |
| rnd33 | 1 | 5 | 5.00 | 5361 | 5 | 2982 | 0+783 |
| rnd37 | 3 | 3 | 2.00 | 5491 | 5 | 2876 | 0+384 |
| rnd43 | 3 | 11 | 7.33 | 3289 | 22 | 1741 | 0+543 |
| rnd47 | 1 | 5 | 5.00 | 4686 | 4 | 2726 | 23+15 |
| rnd63 | 5 | 7 | 4.20 | 2681 | 21 | 1589 | 1957+923 |
| rnd96 | 20 | 9 | 5.25 | 7908 | 80 | 4011 | 46+10 |

Table 5.1: The planning model instances. The nonconvex terms are expressed in the form $o + c$ where $o$ is the number of nonconvex terms in the objective, and $c$ is the number of nonconvex terms in the constraints.

- BARON [320] (under the GAMS [86] interface), a global optimization sBB solver acting on the original MINLP formulation.

- CPLEX [200] (under the AMPL [154] interface), a BB solver acting on the MILP reformulation as per Corollary 5.3.3.

- MINLP_BB [247] (under the AMPL interface), a BB solver for MINLPs which only branches on integer variables, acting on the original MINLP formulation — MINLP_BB only guarantees optimality if the NLP relaxation of the problem is convex (which is not the case here), but it proves to be an effective heuristic for nonconvex MINLPs.

Running times were generally very low for all solved instances: well within 2s CPU time for most instances, with a few exceptions which clocked at under a minute (on a PIV 1.2GHz with 640MB RAM running Linux). Running times were not deemed to be significant comparative indicators, particularly in view of the fact that such problems need usually not be solved in real time. The results are given in Table 5.2.

There are at least two important conclusions that can be drawn from the results in Table 5.2. First, all the considered methods scale well with the problem size. We feel it is particularly important to remark that the global optimization solvers perform on this application as well as the CPLEX MILP solver. Secondly, the experimental results are in line with the conclusion of Theorem 5.3.2 (we note that all global optimization solvers were run with a convergence tolerance of $\varepsilon = 0$).

| Instance name | BARON | CPLEX | MINLP_BB |
|---|---|---|---|
| planning1 | -1993500.000006 | -1993500.000000 | -1993500.000000 |
| planning2 | -1487141.062815 | -1487141.062815 | -1487141.062815 |
| planning3 | -1293674.672558 | -1293674.672558 | -1293674.672558 |
| planning4 | -1298674.672558 | -1298674.672558 | -1298674.672558 |
| planning5 | -1592708.399909 | -1592708.399909 | -1592708.399909 |
| planning6 | -1282208.399909 | -1282208.399909 | -1282208.399909 |
| planning7 | -1287208.399909 | -1287208.399909 | -1287208.399909 |
| planning8 | -1688000.000000 | -1688000.000000 | -1688000.000000 |
| planning9 | -656478.709039 | -656478.709039 | -656478.709039 |
| planning10 | -482425.869355 | -482425.869355 | -482425.869355 |
| rnd10 | -10172670.9000 | -10172670.900000 | -10172670.900000 |
| rnd14 | -46793640.7549 | -46793640.754903 | -46793640.754903 |
| rnd16 | -2517826.01334 | -2517826.013343 | -2517826.013343 |
| rnd20 | -74172261.5403 | -74172261.540278 | -74172261.540278 |
| rnd23 | -225161225.148 | -225161225.147790 | -225161225.147790 |
| rnd24 | -24121864.8543 | -24130018.218155 | -23841483.937430* |
| rnd26 | -2609446.88325 | -2609446.883250 | -2609446.883250 |
| rnd33 | -720041744.091* | -922056072.533006 | -917574429.510109* |
| rnd37 | -348612891.222* | -352566723.200840 | -288592249.147212* |
| rnd43 | -78666258.8413 | -78666633.904820 | -78666633.904820 |
| rnd57 | -28534845.5788 | -28534845.578841 | -28534845.578841 |
| rnd63 | -138429853.758 | -143249498.228375 | -138391489.573639 |
| rnd96 | -14015008.6266 | -14015008.626632 | -14015008.626632 |

Table 5.2: Objective function values found by BARON, CPLEX, MINLP_BB. Non-optimal values are marked *.

## 5.5 A more realistic planning model

The models of Sect. 5.2 and Sect. 5.3 rely on several simplifications of real-life conditions. A truer picture would encompass other realistic features, as detailed below.

- Some of the plants considered in this report produce electricity. These have very specific properties and behaviours [265, 266], among which:
  1. in a true market situation (i.e. no subsidization), electricity prices vary during the course of a single day, as demand rises and subsides;
  2. some electricity production plants are often designed to produce electricity *and* heat (which is either stored or conveyed directly into buildings in the area) — such plants are called Combined Heat Power (CHP) [75];
  3. CHPs generate heat and electricity at the same hour and same location.
- Transportation costs do not depend linearly on the distances due to the different means of transportation used [280]. For very short transportation distances, tractors may be used, which have higher transportation cost than lorries, used for medium to long distances; for very long transportation distances, trains or ships are used. More generally, the geography of the production process region deeply influences the costs of the single process activities; [280] suggests a methodology that combines Geographical Information Systems (GIS) software with process analysis to estimate these costs.
- All production plants incur a one time installation cost. This varies according to many factors, the most important of which are production capacity and physical location. A given production

capacity can be obtained by building a single huge plant or many different small-sized plants. With respect to the latter choice, the former yields: (a) higher production efficiency but lower total efficiency (because usually it is difficult to convey auxiliary heat from one single place to many neighbouring urban areas); (b) lower specific building costs; (c) higher transportation costs due to distances [280].

It turns out that the planning model in Sect. 5.3 can be extended to accommodate most of the features above, excluding the electricity price variability over short time periods (point 1 in the list above). For this we would need an explicit time dependency in the model. Since the electricity price variability range is much shorter than the envisaged time horizon in the planning optimization (one day as opposed to one year), an hourly time discretization would multiply the number of variables by 8760 ($= 365 \times 24$), thus yielding a MINLP or MILP whose size is excessive with respect to the current solver state of the art. It is worth emphasizing, however, that since most of the available biomasses can be easily stored, the variability of electricity prices can be dealt with by adding biomass storage space near the electricity plants, which would either add more processing sites or simply reflect on the plant cost parameter. Storage space is not the only way to deal with the situation: the energyPRO model [265] proposes an electricity plant planning methodology that locally optimizes each plant over a yearly time horizon with hourly time-steps. The combined production of electricity and heat (point 2 in the list above) can be dealt by our model by simply introducing an output process site representing heat, and adapting the $\lambda$ and $\pi$ parameters relative to the CHP, various inputs and heat output to reflect the situation. As a consequence of point 3 in the list above, this modelling is not wholly satisfactory, as generation of heat is time-dependent because it is linked to the generation of electricity: but again this time dependency can be dealt with by using process sites representing heat storage capacity or simply adding to the plant cost parameter.

Nonlinear transportation costs are already fully dealt with by our model, for with each arc we associate a transportation cost which is not unitary but rather depends on the vertices adjacent to the arc. Since the arc length is not used anywhere in the model, each arc can be assigned its proper cost.

In order to cater for the last point, i.e. plant installation costs, we consider the parameters $I_{vp}$ = cost of building plant $p \in P(v)$ in vertex $v \in V_0$. The dependency of the installation cost with the vertex $v$ allows us to consider geography dependencies as outlined in [280]. We then add the following term $\lambda_4$ to the objective function:

$$\lambda_4 = \sum_{v \in V_0} \sum_{p \in P(v)} I_{vp} w_{vp}. \tag{5.23}$$

This does not change the convergence results given in Sect. 5.3.1.

We remark that although we treat plants as rather simple entities defined by their input, yields and outputs much like other nodes, this is a simplistic view. Some refined models that describe these these energy conversion units are given in [75, 76, 74]; in particular, each of these nodes may sometimes consist of a set of different plants and utilization points where transportation of energy/heat has an impact on the overall efficiency. More recently, a full supply-chain model was provided that also cater for multi-company interactions [335].

## 5.6 Conclusion

We describe a Linear Programming model for running a biomass-based energy production process and a Mixed-Integer Nonlinear Programming model for a simplified planning of the installation of processing plants used in the production process. Although the solution of the first model is readily obtained by any good quality LP solver, the second (nonlinear) model is nonconvex, exhibits multiple local minima and therefore needs to be solved using Global Optimization techniques. We show that a spatial Branch-and-Bound type algorithm converges exactly to the optimum and that the MINLP model can be reformulated

exactly to a MILP; this result is also apparent in the computational results, ranging over a set of realistic instances and a set of randomly generated ones. Finally, we extend the planning model to deal with more realistic features.

# Chapter 6

# Optimizing Architectural and Structural Aspects of Buildings Towards Higher Energy Efficiency

Álvaro Fialho, Youssef Hamadi, Marc Schonauer

*Manuscript in preparation*

*We aim at contributing to the issue of energy consumption by proposing tools to automatically define some aspects of the architectural and structural design of buildings. Our architecture starts with a building design, and automatically generates and searches a space of acceptable design variations. It outputs a variation which maximizes energy efficiency, and respect cost constraints. The optimization stage is done by the combination of an energy consumption simulation program, EnergyPlus [336], with a state-of-the-art multi-objective evolutionary algorithm. The latter explores the design search space, automatically generating new feasible design solutions, which are then evaluated by the energy simulation software.*

## 6.1 Introduction

The continuous rising of energy consumption is a current (and global) concern. On the one hand, there is the fact that energy is still mainly coming from non-renewable and limited sources. On the other hand, the more energy is consumed, the more carbon emissions are released in the atmosphere. According to a recent report from the *World Business Council for Sustainable Development* [360], the building sector is responsible for the most important energy consumption rate, estimated at around 40% of the total energy used worldwide. Surprisingly, the resulting carbon emissions are even higher than those of all the transportation sector combined. Reducing energy consumption, while not compromising the rising living standards of the ever-growing population, has become thus a matter of extreme importance towards a global sustainable future.

In this work, we aim at contributing to this issue by proposing tools to automatically define some aspects of the architectural and structural design of buildings, in order to passively maximize their energy efficiency, while maintaining the construction costs and occupants' thermal comfort at a reasonable level. It is important to note that, for the sake of simplicity, we are currently considering only the period during which the building is in "operation". This corresponds to around 80% of the total energy consumption

during the life cycle of a building [360], the other 20% being spent on its construction (including the manufacturing of the materials used) and demolition.

The optimization of the building designs is done by the combination of an energy consumption simulation program, EnergyPlus [336], with a state-of-the-art multi-objective evolutionary algorithm. The latter explores the design search space, automatically generating new feasible design solutions, which are then evaluated by the energy simulation software. This generate-and-test cycle is repeated until a satisfactory design is found, or another stopping condition is achieved.

This document is structured as follows. Section 6.2 surveys some related work. In Section 6.3, the EnergyPlus simulation program is briefly described, while Section 6.4 overviews the evolutionary algorithm employed. Finally, Section 6.5 presents the results of some preliminary experiments, while Section 6.6 concludes with some perspectives for further work.

## 6.2 Related Work

In [303], the authors optimize the shape of the building envelope in order to maximize solar efficiency at its façade, which directly affects the daylighting and the Heating, Ventilation and Air Conditioning (HVAC) energy consumption. Both variables are computed using the EnergyPlus software. As an extra, the authors suggest the use of CFD analysis to, for example, analyze the wind flow in the internal space of the building.

In [124], the objective is to minimize costs while maximizing energy efficiency. As choices for the optimization method, there are the type of windows and the type and quantity of the insulation material used in the walls. The thermal transmittance and the conductivity of each material are taken from the ASHRAE database [10], and the prices are artificially simulated according to the performance of the material.

In [91], many different applications are considered, with growing levels of complexity, which can be summarized as follows:

1. Starting with box-like offices facing each cardinal direction (squared one-floor buildings), the first objective is to maximize the energy efficiency by finding the best window dimension for each office and orientation.

2. In the second case, the geometry of the building and the space layout are fixed, the objective is to optimize its façade, mainly the size and placement of the windows, aiming at maximizing the energy efficiency.

3. In the following case, the objective is two-fold: minimize the cost of the materials used in the construction of the building, while also maximizing the energy performance. In a second step, not only the cost of the materials is considered, but also the energy saved during the construction phase and the energy spent to manufacture these materials.

4. In the last case, the objective is to automatically generate/evolve complete 3D architectural forms that are energy-efficient, while at the same time being in agreement with the architectural design intentions expressed by the architect (represented in terms of a well-defined set of rules).

Besides the building design, another component that significantly affects the energy consumption is the HVAC system. The optimization of the design of the HVAC system is the focus of the works presented in [367, 17].

## 6.3 EnergyPlus

EnergyPlus is a very complete energy analysis and thermal load simulation program, which is available[1] free of charge for all the main computer platforms. It can be seen as an ameliorated extension of the BLAST and DOE-2 energy simulation programs, which were developed after the 1970s energy crisis. Many enhancements have been continuously aggregated to it since its first release in 2001, mainly supported by the U.S. Department of Energy.

This software enables the evaluation of the energy consumption behavior of almost any kind of building, as defined by the user via an input file containing the design parameters, according to the meteorological trends (weather data) of the region where it is planned to be built. It is primarily a simulation engine, with both input and output being made via text files. Some GUIs and special plugins are provided by third-party developers in order to facilitate and extend its use. It is mainly used by "design engineers" and architects to, for instance, appropriately size the HVAC equipments, develop retrofitting projects, evaluate and/or optimize energy performance, etc.

In this work, we use it to evaluate the energy consumption of the design solutions automatically generated by a search/optimization algorithm, which will be described in the following.

## 6.4 Multi-Objective Optimization with Evolutionary Algorithms

As previously discussed, for the optimization of the buildings, the main objective in terms of sustainable development here is the reduction of energy use (while also possibly generating some energy, e.g., solar or eolic energy, but this is out of the scope for the time being). But the more energy-efficient is the building, the more expensive tends to be its construction. One might thus find a compromise between energy efficiency and construction costs.

Another well-known trade-off in this context is the exploration of natural daylighting versus thermal isolation. The bigger the windows, the more daylight will possibly come in; but windows are usually much less efficient in terms of thermal insulation than walls. Then, some savings in the use of electricity for lighting might be (even over-)compensated by a higher need of the HVAC mechanisms, which are usually the most energy-demanding equipments in a building. The non-use of these equipments, or their use at a smaller power rate, might significantly affect the total energy consumption; but at the same time it will also affect the thermal comfort of the people living/working in the given building, most probably in a negative way.

Based on these examples, it is clear that multiple objectives need to be taken into account in order to do a more realistic optimization of building designs. Evolutionary algorithms are very popular meta-heuristics for multi-objective optimization, mainly due to the fact that they are population-based, i.e., a set of solutions is continuously evolved in parallel. This enables the decision-maker to choose, in the end of the process, between many solutions that are optimal w.r.t. the considered objectives in different ways, not being comparable between each other, the so-called Pareto front.

A current trend to evaluate the quality or fitness of a solution in multi-objective optimization is the use of the hyper-volume measure [385]. Briefly, the fitness of a solution is equal to its contribution to the hyper-volume computed between the current Pareto front and a reference point in the search space. In this work, we are using the current state-of-the-art hyper-volume-based algorithm, referred to as *Hype* [45], which will now be briefly described in turn.

[1] EnergyPlus is available at http://apps1.eere.energy.gov/buildings/energyplus/

### 6.4.1 The Hype Algorithm

Besides the proposal of a new approximate way of calculating the hyper-volume, based on sampling, the Hype algorithm [45] differs from the other available variants of hyper-volume-based multi-objective evolutionary algorithms mostly in two aspects: its fitness assignment and its replacement/survival selection mechanism.

Concerning the former, as in the other existing algorithms, the fitness assigned to a solution is equal to its contribution to the current hyper-volume. The main difference is that each solution receives partial "credit" for the regions of the hyper-volume that are dominated by more than one solution, as exemplified in Figure 6.1.

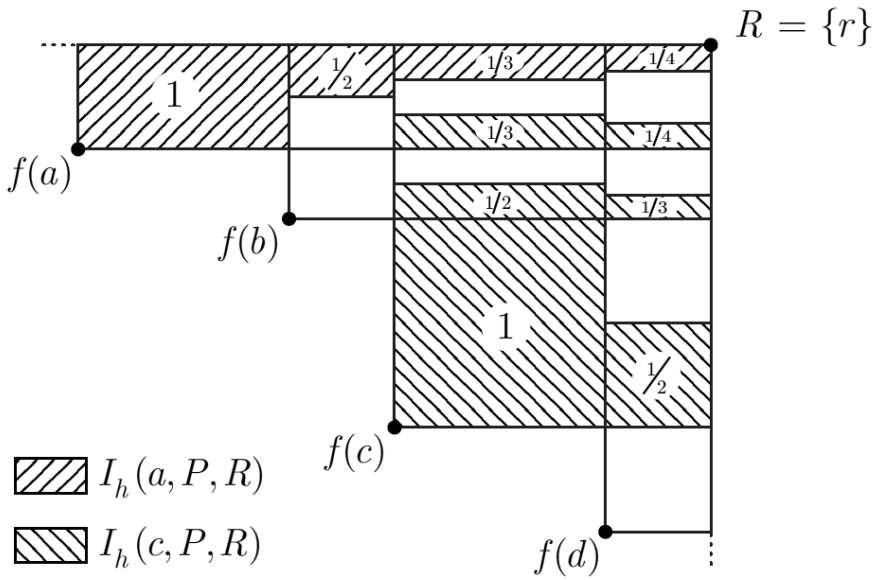



Figure 6.1: Illustration of the Hype credit assignment, extracted from [45].

And for the replacement or survival selection mechanism, Hype proposes a special way of doing it, as follows.

1. Starting from the merged parents+offspring population, it firstly divides it into non-dominance partitions (using the non-dominance sorting concept from the NSGA-II [117] algorithm). The first partition is the set of non-dominated solutions (the Pareto front) of the given population. By removing these solutions from the population, we have a new Pareto front, which is the second partition (or depth) of non-dominance, and so on. These partitions are then included one by one in the new population, whenever there is space available for the entire partition.

2. At some point, one partition will not entirely fit. It is then considered using the Hype fitness assignment method, as follows. Let $k$ be the number of individuals that needs to be removed from this partition so it can included in the new population. Briefly, at each iteration from 1 to $k$, the fitness (the contribution to the hyper-volume) of each individual of the given sub-population (the partition) is calculated, and the worst individual is removed.

This process is exemplified in Figure 6.2. All the solutions are represented by a red cross. The $\times$ marks the first partition of non-dominated solutions, which has 5 individuals, and thus can be entirely included in the new population (of size 8 in this case). The $\times$ marks the second partition (the new Pareto front when removing the first partition). It does not fit entirely in the new population, and only 3 out of

5 individuals are kept, marked with a black square, according to their contribution to the hyper-volume formed by this sub-population.

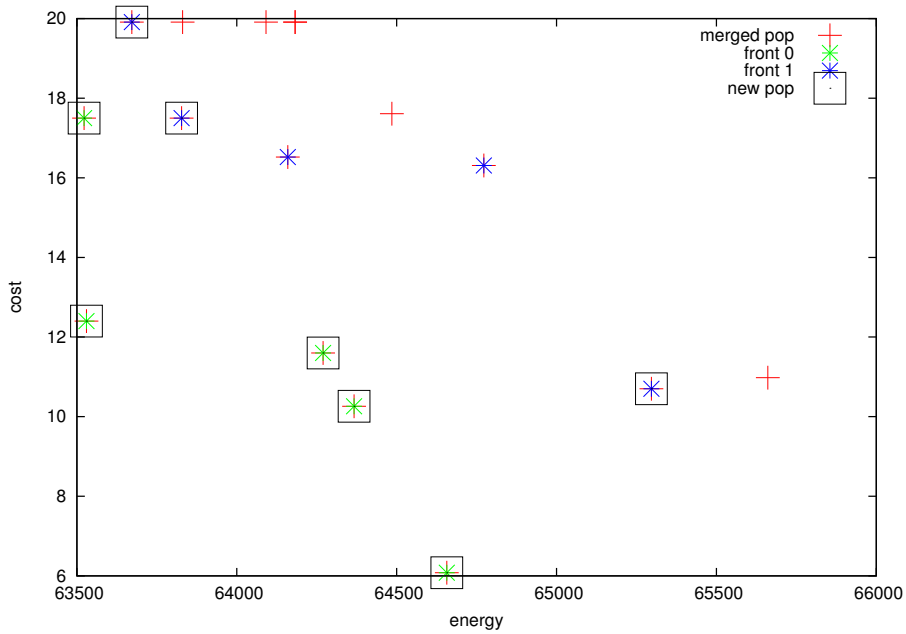



Figure 6.2: Example of the Hype replacement mechanism in action. The solutions marked with a black square are the ones chosen to remain for the next generation, while the others will be discarded.

## 6.5 Preliminary Experiments

This is a currently on-going work, still on its early stages. Some experiments were done in order to validate our experimental framework and to empirically verify the possible gains that might be achieved with it. These experiments will be described in the following.

### 6.5.1 The Building

The first building design tackled was taken from [138, Exercise 2C]. The description taken from the IDF file is the following:

> *Building: Single floor rectangular building 30.5 m (100 ft) x 15.2 m (50 ft). 5 zones - 4 exterior, 1 interior, zone height 2.4 m (8 ft). Exterior zone depth is 3.7 m (12 ft). There is a 0.6 m (2 ft) high return plenum. The overall building height is 3m (10 ft). There are windows on all 4 facades; the south and north facades have glass doors. The south facing glass is shaded by overhangs. The walls are woodshingle over plywood, R11 insulation, and gypboard. The roof is a gravel built up roof with mineral board insulation and plywood sheathing. The floor slab is 0.1 m (4 in) of heavy concrete. The windows and glass doors are double pane Low-e clear glass with argon gap. The window to wall ratio is approximately 0.3.*
>
> *The building is oriented with the long axis running east-west.*
>
> *Floor Area: 463.6 m2 (5000 ft2)*
>
> *Internal: Lighting is 16 W/m2 (1.5 W/ft2), office equip is 10.8 W/m2 (1.0 W/ft2). 1 occupant per 9.3 m2 (100 ft2) of floor area. The Infiltration is 0.25 air changes per hour.*

*HVAC: Single-zone unitary with DX cooling and gas heating serving NORTH PERIMETER zone. VAV with hot water reheat, return plenum, chiller, boiler, and tower serving the other four occupied zones.*

*Environment: Chicago, IL, USA, TMY2*

Different 3D visualizations of this building are shown in Figure 6.3.

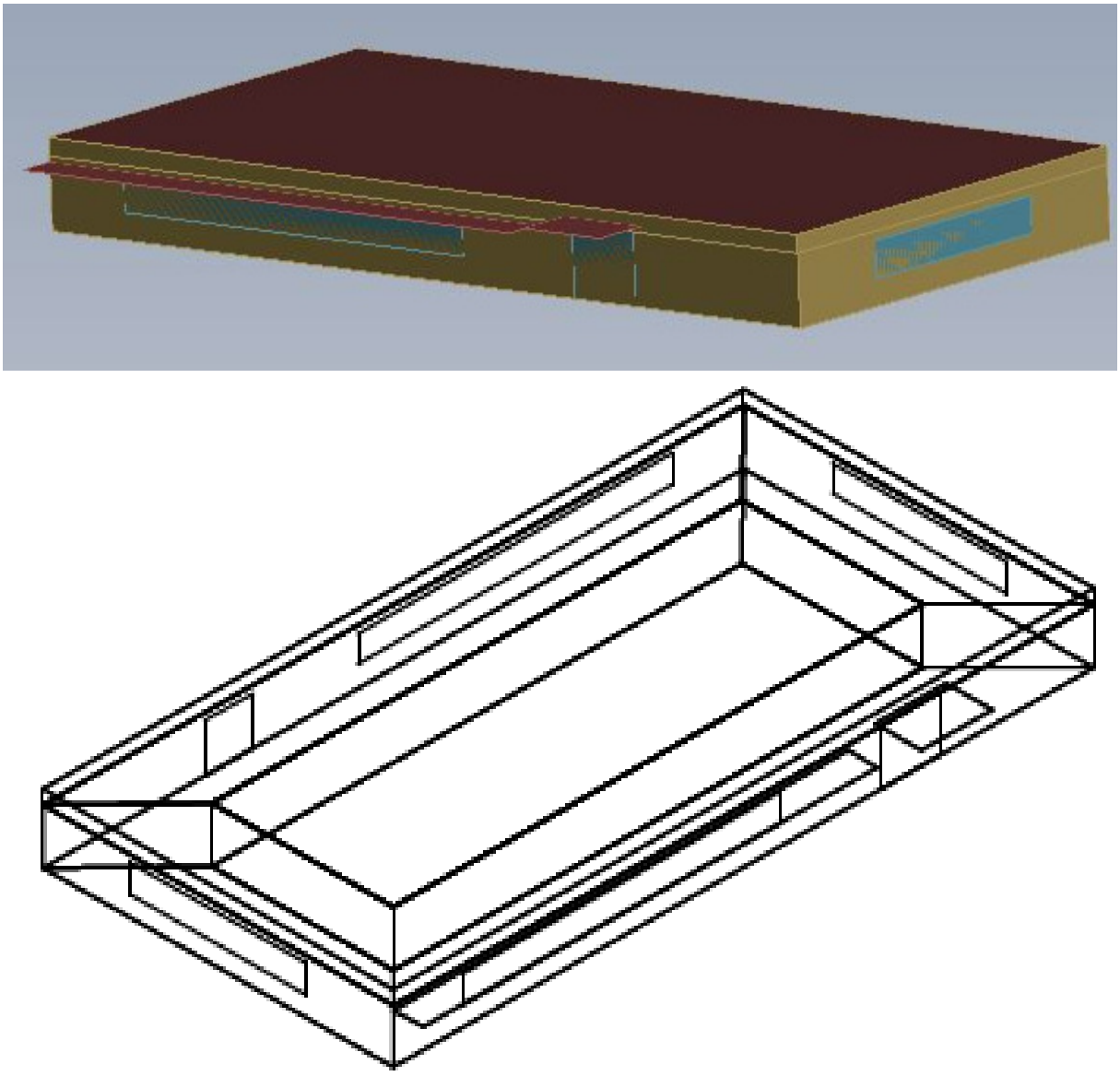


Figure 6.3: Different views of the building used in the first experiments.

### 6.5.2 The Problem

These preliminary experiments are considering two conflicting objectives: the minimization of both the energy consumption and the construction costs of the building. The former is taken from the "Total Site Energy" variable in the "Annual Building Performance Summary" output report, in kW/h. For the latter,

we consider the sum of the costs of the insulation materials used, in €/$m^2$, taken from a french retailer. Three problem variables are considered here, as follows.

The first one is the orientation angle of the building, expressed in degrees with respect to the real North axis. It is being currently considered as a real value between 0 and 360, although in practice some physical constraints might exist.

The other two represent the choice of material used for each insulation layer of the external walls. These walls are finally constituted by the aggregation of these insulation layers with two other layers, the external and the internal ones, which are not modified. Our current database of insulation materials contains 33 different kinds of material, with their corresponding thermal insulation performance and price. So, these are categorical variables, ranging between 1 and 33 (the "index" of each material), with partial ordering: materials #5 to #10 might be of the same kind but with different thickness (and consequently different performance), while material #11 might be a totally different one.

### 6.5.3 Evolutionary Algorithm

The implemented evolutionary algorithm uses the Hype [45] fitness assignment and replacement mechanisms, briefly described in Section 6.4.1. For the mating selection, it uses tournament with $t = 2$: for the generation of each new solution, two solutions are randomly selected from the main population, and the best between them (in terms of contribution to the hyper-volume, as defined by Hype) is selected for the mating.

Very simple variation operators are being currently used: a 1-point Crossover, and a Gaussian Mutation with mean zero and standard deviation 1 (mutiplied by 10 for the Orientation variable, because there is no meaning in doing very small variations on the building orientation). Crossover is applied at a fixed rate of 0.8, and Mutation at 0.1, an identical copy of the solution is done otherwise.

### 6.5.4 Analysis of Results

The results presented here are extracted from a single run, considering "cost versus energy consumption", using a (8+32)-EA: at each generating, 32 new solutions are created out of 8 solutions, and the population for the next generation is composed by the 8 best (according to the Hype procedure) of the (8+32) merged population. The stopping condition was fixed to 5000 fitness function evaluations.

Figure 6.4 shows the population (of size 8) after every 10 generations. The solutions marked with a black square represent the final Pareto front. As it can be seen, the algorithm seems to have converged to this solution set many generations before the end. In the near future, we could eventually look into how to on-line detect convergence, as in [354], in order to save computational time.

This set of final solutions is described in Table 6.1. The two rightmost columns show the energy consumption and the cost per square meter for each combination of angle and insulation materials.

For the sake of comparison, the original design of this building, using Orientation=0 (i.e., aligned with the real North axis) and the materials cited on its description in Section 6.5.1, achieves a total energy consumption of 65735.2 kWh per year, considering the meteorological data of the OHare International Airport in Chicago, USA (the same used in these experiments). In case there are no budget constraints, the most energy efficient solution (#1) will represent a reduction of around 5% in terms of energy consumption.

These results become more interesting when considering both objectives. We could read it in the following way: given a target energy consumption, for instance 63600 kW/h, the cheapest solution found (#4) will cost 5.34 €/$m^2$. In the opposite direction, given a budget of 8 €/$m^2$, the most energy-

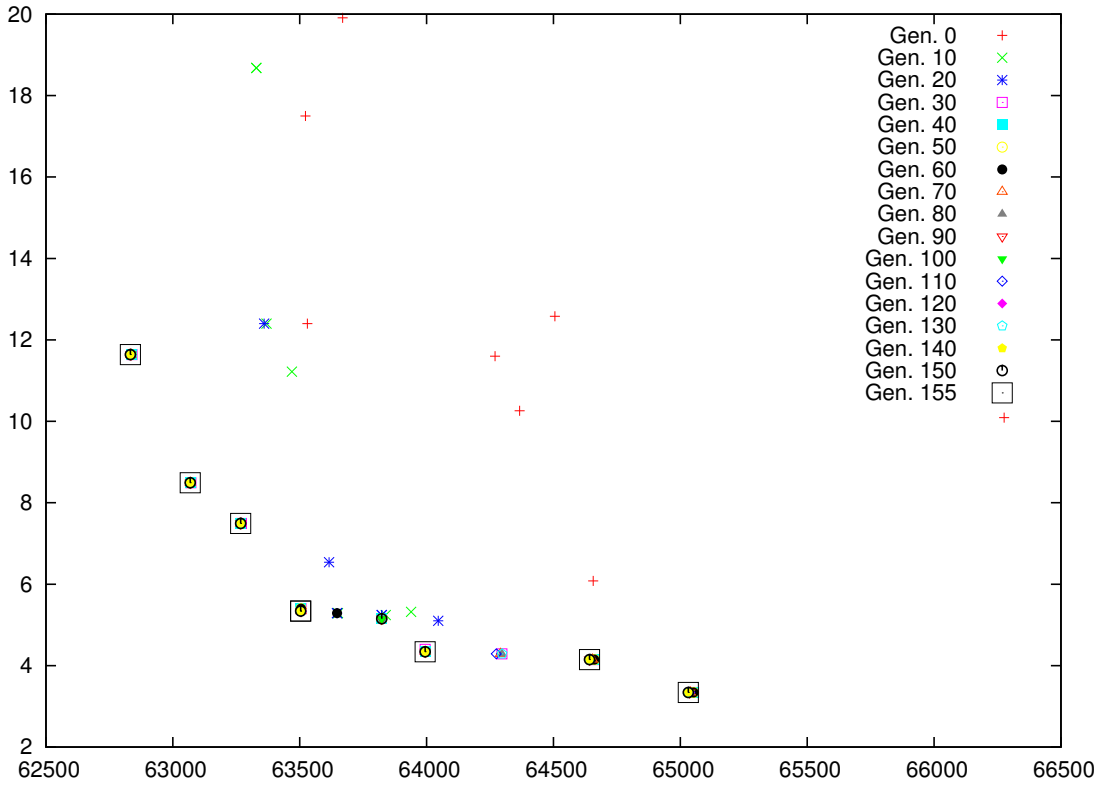



Figure 6.4: The populations after every 10 generations, up to the final Pareto..

efficient solution will be the number #3.

As it can also be seen, all these solutions use around the same value for the orientation angle of the building. This is intuitive, and now empirically verified: if we use the same combination of materials for all the external walls of the building, the optimization of the orientation angle becomes an independent problem. In order to further check it some extra experiments were done considering solution #1, varying the angle $\in [0 : 5 : 360]$. The results are shown in Figure 6.5. As it can be seen, in this case the optimum is really around 265 degrees, and up to 700kW/h (around 1% in this case) can be saved only by optimizing the orientation angle.

## 6.6 Conclusion and Further Perspectives

The experiments presented, although being very preliminary, showed to be very useful in order to validate the experimental framework, besides demonstrating that this project has a great potential in terms of contribution to sustainable development.

There are many other design aspects that shall be considered in the near future. For instance, we are currently working on including the size and placement of the windows as problem variables, considering an additional objective in this case, the absorptance of natural daylighting. Besides, for the time being, we are currently taking into account only objectives related to minimize the energy consumption in a passive way. Further work might also address the choice of HVAC systems that should be used, as well as their control parameters.

Additionally, for the moment we have considered only a single (and simple) building, while many

| S | Angle | Layer 1 | Layer 2 | kWh/year ⇓ | €/$m^2$ |
|---|---|---|---|---|---|
| 1 | 265.248 | coated glass wool KIC 200 4.5x1.2M | coated glass wool KIC 200 4.5x1.2M | 62833 | 11.64 |
| 2 | 265.037 | coated glass wool KIC 200 4.5x1.2M | coated glass wool KIC 100 8x1.2M | 63068.8 | 8.49 |
| 3 | 265.037 | comfort coated glass wool 45mm 15.6X0.6M | coated glass wool KIC 200 4.5x1.2M | 63267.4 | 7.49 |
| 4 | 264.917 | glass wool IBR nu 100 7X1.20M | glass wool IBR nu 100 7X1.20M | 63503.9 | 5.34 |
| 5 | 264.917 | glass wool IBR nu 100 7X1.20M | glass wool IBR nu 100 7X1.20M | 63503.9 | 5.34 |
| 6 | 264.811 | comfort coated glass wool 45mm 15.6X0.6M | coated glass wool KIC 100 8x1.2M | 63994.5 | 4.34 |
| 7 | 263.956 | comfort coated glass wool 45mm 15.6X0.6M | comfort coated glass wool 60mm 12X0.6M | 64642.1 | 4.15 |
| 8 | 263.956 | comfort coated glass wool 45mm 15.6X0.6M | comfort coated glass wool 45mm 15.6X0.6M | 65031.6 | 3.34 |

Table 6.1: Final set of solutions, ordered by the energy consumption.

other buildings should be tackled in order to further validate our approach. The plan is to create a benchmark set containing several building of very different nature, such as office buildings, warehouses, churches, hospitals, schools, etc, with different geometries (e.g., L-shape, U-shape), number of floors, HVAC systems, etc.

Finally, in order to make all this more realistic, we are currently starting a collaboration with an architecture company. They will assist us with all their expertise in what concerns architecture in the real-world, and possibly provide us the project of a real building to be optimized.

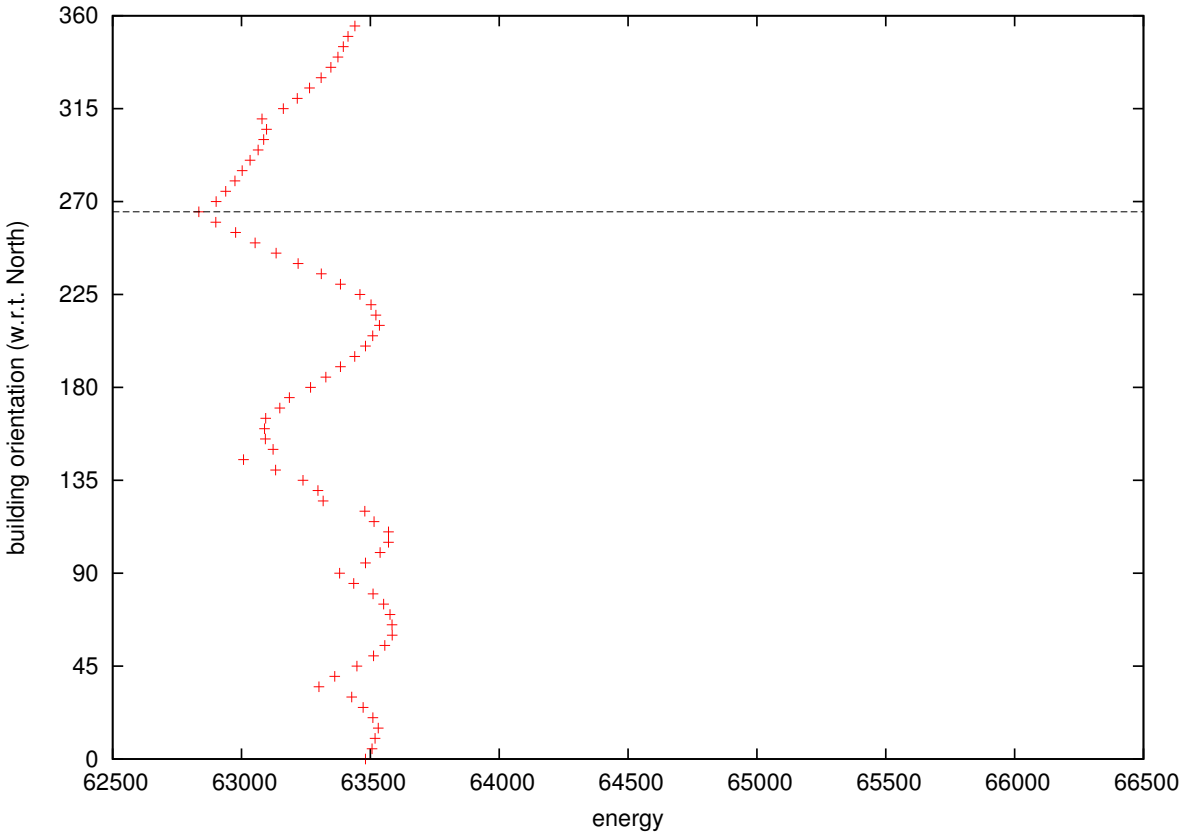



Figure 6.5: Energy consumption evaluation, in kW/h averaged over an year, for different orientation angles of the most energy-efficient building.

# Part II

# Sustainable Transportation

# Chapter 7

# On green routing and scheduling problem

Nora Touati-Moungla, Pietro Belotti, Vincent Jost

*In revision in International Transactions in Operational Research*

*The vehicle routing and scheduling problem has been studied with much interest within the last four decades. In this report, some of the existing literature dealing with routing and scheduling problems with environmental issues is reviewed, and a description is provided of the problems that have been investigated and how they are treated using combinatorial optimization tools.*

## 7.1 Motivation

During the last few years, Operations Research (OR) has extended its scope to include environmental applications [70, 110, 324]. Because in Europe 73% of the oil is used for transportation purposes, the need to design efficient plans for sustainable transportation is evident. Advances in the transportation planning process and in the efficiency of transportation systems are key components of the development of sustainable transportation.

The routing of vehicles represents an important component of many distribution and transportation systems and has been intensively studied in the OR literature [347]. In this report, particular consideration is given to routing and scheduling models that relate to environmental issues, we will denote this class of problems as Green Routing and Scheduling Problems (GRSP). In [324], the authors discuss different problems that relate to sustainable logistics, they focus on reverse logistics, waste management, and vehicle routing and scheduling problems. Some variants of routing and scheduling problems in connection with environmental considerations were described: the arc routing problem, which is considered as a major component in waste management, and the time-dependent vehicle routing problem which allows one to indirectly decrease gas emissions involved by transportation activity by avoiding congested routes.

We present in this report some general tools for transportation decision-making under assumptions related to economic, environmental and social considerations. An exhaustive review of sustainable transportation problems and their treatment by OR tools is out of scope here given the generality of this area of research. Most of the work in this domain is still very much in development and some applications have only just started. The aim of this survey is to provide a clear presentation on how combinatorial

optimization can contribute to sustainable transportation as well as a comprehensive survey covering all known green routing and scheduling problems and their variants. Therefore, we list some GRSPs that are studied in the literature [324] and identify some other problems which can be added to this class, we describe some models and solution methods that can be exploited for these problems, and we expose some multiobjective optimization methods which are essential for solving these particular problems.

### 7.1.1 Green routing and scheduling

Usually, routing and scheduling models are concerned with objectives of minimizing economic costs, but due to growing concerns about public health, global warming and economic safety, it is necessary to consider in the cost function the factor of environmental and social costs. The additional environmental and social constraints and objectives that must be taken into consideration often make the problem more difficult to both model and solve. We study in this report the best known routing and scheduling related problems arisen from sustainable transportation field:

- *Routing of Hazardous Materials (RHM):* The objective is the minimization of the risk on the population and the environment caused by the transportation of hazardous materials. The research area that investigates this problem is most advanced, while less research taking into account explicitly environmental impacts is dedicated to the other problems cited bellow.
- *Routing and Scheduling in Time-Dependent Environment (RS_TDE):* This class of problems contributes indirectly to reduce vehicles gas emission. The main objective is the minimization of a more realistic travel time by avoiding congested routes.
- *Waste Collection Vehicle Routing Problem (WCVRP):* This problem is a major component of waste management.
- *Multi-Modal Vehicle Routing problem (MMVRP):* This problem permits to manage many transportation modes and allows to perform priority-based routing for clean transportation (as rail transport).
- *Dial-a-Ride Problem (DARP):* This problem contributes indirectly to decrease the global taxis gas emission by promoting grouped transportation (grouped taxis) for decreasing the transportation fleet size and routes congestion, particularly in large cities.
- *Pick-up and Delivery Vehicle Routing Problem (PDVRP):* Permits the integration of the backward flow of waste in distribution systems for recycling for example.
- *Energy Routing Problems (ERP):* This problem is one of the least studied one in the context of sustainable transportation. It permits to promote the use of electric vehicles by maximizing the vehicle autonomy.
- *Air Traffic Management (ATM):* This problem can contribute to decrease the fuel consumption in planes.

These problems are very different from their structure, their contribution to transportation sustainability and their dedicated models and solution methods. Works on these problems are often unbalanced, this generally depends on the problem characteristics, for example RHM includes many aspects as risk definition and model, risk equity and uncertainty while the major aspect of RS_TDE is the travel time.

### 7.1.2 Limitations

A very large literature cover routing and scheduling problems presented in this survey, we do no give mathematical models of the presented problems since they are already well defined in the literature.

However, an exception concerns energy routing problems for which no model is given in the literature. For each problem, we describe its characteristics, how it contributes to sustainable transportation, some models and resolution methods related to problems taking into account environmental considerations only. For example, in pick-up and delivery vehicle routing problem, the pick-up can concern materials or goods, we only interest to works in the literature which consider the pick-up of waste for recycling for example.

expluding for reason of space.

### 7.1.3 Structure of the survey

This report is structured as follows. Section 7.2 describes the classical vehicle routing and scheduling problem. In Section 7.1.1, we describe the problems cited above, why they can be considered as green routing and scheduling problems, their particularity compared to the standard vehicle routing and scheduling problem, some related models and solution methods. We adress in Section 7.11 the caractreristics of green routing and scheduling problems and some classification schemes. Finally, we conclude in Section 7.12.

## 7.2 Vehicle routing and scheduling problem

The Vehicle Routing Problem (VRP) can be stated as follows: Consider a fleet of $K$ identical vehicles of fixed capacity $C$ available at a given depot to serve a set of costumers with fixed demand. We are given an oriented graph $G = (N, A)$, where $N$ is the set of nodes including the costumers and the depot, and $A$ the set of arcs connecting the nodes. Each arc $(i, j)$ is associated with a cost $c_{ij}$ and each costumer $i \in N$ has a demand $d_i$. The goal is to find a set of minimum cost vehicle routes that service every costumer such that:

- Each vehicle route originates and terminates at the depot.
- Each vehicle services one route and each costumer is visited by exactly one vehicle.
- The demand of each costumer is satisfied and the capacity of each vehicle is not exceeded.

The VRP is an important sub-problem in a wide range of distribution systems and a lot of effort has been devoted to research on different variants of this problem. Indeed, in practice, additional constraints or changes in the structure of the basic model are taken into account. We cite for example *the VRP with time windows* which involves time window constraints restricting the times at which a customer is available to receive a delivery, in *the VRP with Pick-up and Delivery*, each vehicle must visits the pick-up location before the corresponding delivery location, in *the VRP with Backhauls* customers can demand or return some commodities, in *the multiple depot VRP* the company may have several depots from which it can serve its customers and in *the open VRP* each vehicle is not required to return to the central depot after visiting the final customer (see [347] for a description of these VRPs). Many other variants of the VRP exist, however we focus in this work on vehicle routing and scheduling problems taking into account environmental considerations.

## 7.3 Routing problem for hazardous materials

### 7.3.1 Description

The transportation of hazardous materials (hazmat from here on) has received much interest in recent years, this results from the increase in public awareness of the dangers of hazmats and the enormous amount of hazmats being transported. The main target of this problem is to select routes from a given origin $s$ to a given destination $t$ such that the risk for the surrounding population and the environment is minimized, without producing excessive economic costs. The study of hazmat transportation problems can be classified in four main subjects: risk analysis [141], routing and scheduling [159], facility location [159], and treatment and disposal of waste [292], we focus in this section on routing and scheduling.

### 7.3.2 Environmental contribution

This problem is naturally a GRSP, since it contributes to minimizing the risks of release accidents on the population and the environment in hazmat transportation activities.

### 7.3.3 Related works

The difference between hazmat transportation and other transportation problems is mainly the risk. The risk makes this problem more complicated by its assessment, the related data collection and the solution of the induced formulations.

#### Problem characteristics

We address in the following three major particularities of RHM problem: risk assessment and risk equity:

**(a) Risk assessment:** Although the fact that the major target of RHM is the minimization of the risk, there is no universally accepted definition of risk (for a survey on risk assessment, see [143]). The risk on population caused by hazmat transportation depends on many factors, the most important of which are: the site of the accident, the meteorological conditions, the distribution of the population in the wider area under consideration and the transported hazmat type. It is pointed out in [159] that the evaluation of risk in hazmat transportation generally consists of the evaluation of the probability of an undesirable event, the exposure level of the population and the environment, and the degree of the consequences (e.g., deaths, injured people, damages). In practice, these probabilities are difficult to obtain due to the lack of data and generally, the analysis is reduced to consider the risk as the expected damage or the population exposure.

As the risk is a part of the objective function, it is quantified with a path evaluation function [143]. This function is not additive since the probability of a release accident on a link depends on the probability of a release accident on the traveled links of the path. This important property leads to non-linear integer formulations which can not optimized using a classical shortest path algorithm. Generally approximations are needed by considering additive functions (Usually considering independent release accident probabilities on links) for obtaining tractable models.

**(b) Risk equity:** When many vehicles have to be routed between the same origin-destination nodes, these vehicles are routed on the same path, hence the risk associated to regions surrounding this path could be high. In this case, one may wish distribute the risk in an equitable way over the population and the environment. The computation of routes with a fairly distributed risk consists in generating dissimilar origin-destination paths, i.e paths which relatively don't impact the same zones. Solution approaches

ca be classified in tow sets, *resolution-equity-based methods* and *model-equity-based methods*. In resolution-equity-based methods, equity constraints are taken into account in the resolution process. These methods are based on a dissimilarity index which permits to indicate when two paths are considered as dissimilar. We present on Table 7.1 some of these methods.

Table 7.1: Resolution-equity-based methods

| Method | Principle | dissimilarity index |
|---|---|---|
| Iterative Penalty Method [208] | Compute iteratively a shortest path and penalize its arcs by increasing their weights for discouraging the selection of the same arc set in the generated paths set in the next iteration. | |
| Gateway shortest-paths method [263] | Generate dissimilar paths by forcing at each time a new path to go through a different node (called the gateway node). | The absolute difference between areas under the paths (areas between paths and the abscissa axis). |
| Minimax method [227] | Select $k$ origin-destination shortest-paths and select among them iteratively a subset of Dissimilar Paths (DP) by means of an index that determines the inclusion or not of candidate paths in DP. | The length of common parts between the paths. |
| $p$-dispersion method [6] | Generate an initial set $U$ of paths and determining a maximal dissimilar subset $S$, i.e., the one with the maximum minimum dissimilarity among its paths. | The length of common parts or the common impact zones between the paths. |

Model-equity-based methods consists of taking into account equity constraints in the model formulation. In [166, 167], the authors propose an equity shortest path model that minimizes the total risk of travel, while the difference between the risks imposed on any two arbitrary zones does not exceed a given threshold $\epsilon$. In [92] was proposed a multi-commodity flow model for routing of hazmat, where each commodity is considered as one hazmat type. The objective function is formulated as the sum of the economical cost and the cost related to the consequences of an incident for each material. To deal with risk equity, the costs are defined as functions of the flow traversing the arcs, this imposes an increase of the arc's cost and risk when the number of vehicles transporting a given material increases on the arc.

**Routing and scheduling**

Hazmat routing is multiobjective in nature, since risk minimization accompanies the cost minimization in the objective function. In addition, other pertinent objectives can be considered as the travel time for minimizing the exposure of the driver to risk. Therefore, a set of alternative (Pareto optimal) solutions have to be computed (see section 7.11.2). Solution methods for hazmat routing can be classified in two categories:

**Local hazmat routing:** Consists of a one origin-destination hazmat routing and aims at selecting routes between a given origin-destination nodes for a given hazmat, transport mode and vehicle type. We present in Table 7.2 some works on local routing.

**Global hazmat routing:** A substantial work in the literature focuses on the selection of a single commodity routes between only one origin-destination pair. In practice, a better suited model is global routing, where different hazmats have to be shipped simultaneously among different origin-destination pairs. In [387], a multiobjective routing model that considers equity constraints is proposed. The model considers the following criteria: the general risk, the risk of special population, the travel time, the property damage, and the risk equity which is imposed using capacity constraints on the network links. The obtained model is equivalent to the capacitated assignment problem and a goal programming method is used to solve it. A pertinent model for global routing was proposed in [92] and described in section 7.3-(b), unfortunately the multicommodity flow model is mono-objective in nature. As our knowledge, the multiobjective aspect of this model is not yet studied for the considered problem.

Table 7.2: Local hazmat routing

| Author | Objective | Method |
|---|---|---|
| Shobrys (1981) [330] | Min. ton-miles traveled<br>Min. population exposure-tons | Weighting method |
| Robbins (1981) [316] | Min. the total length of shipment<br>Min. the size of the population brought into contact with the shipment | Weighting method |
| Current, ReVelle and Cohon (1988) [108] | Min. the population affected around the path<br>Min. the length of the path | Weighting method |
| Abkowitz and Cheng (1988) [3] | Min. ton-miles travelled<br>Min. population exposure-tons | Weighting method |
| Turnquist (1993) [350] | Min. of incident rates related to the release of hazardous material<br>Min. of the population exposed to the risk<br>Min. of the route length | Stochastic dominance |
| Karkazis and Boffey (1995) [215] | Min. expected damage effects | Branch-and-bound |
| I. Giannikos (1998) [163] | Min. Operating cost<br>Min. Total perceived risk<br>Min. Maximum individual perceived risk<br>Min. Maximum individual disutility | Goal Programming, Penalty Functions |
| Zografos and Androutsopoulos (2004) [386] | Min. Total travel time<br>Min. Total transportation risk | Objectives aggregation, Insertion heuristic |

### 7.3.4 Summary

Transportation of hazardous materials is a complex and seemingly intractable problem, principally because of the inherent trade-offs between social, environmental, ecological, and economic factors. Several important directions for future challenging research can be stem from this problem.

There is no common conceptual model for the RHM problem. Works in this field takes generally account of different considerations (economic, environmental and social) and significant simplifications are necessary to obtain tractable models. We present in the following some important characteristics of the RHM:

- *Risk categories:* explosion, toxicity, radioactivity, corrosiveness, infectious risk and burns risk.
- *Transportation modes:* most studies on RHM in the literature deal with road transportation mode [142]. Although rail transportation is safer (automatic control system, do not cross populated zones), and more capacitated, it received less attention. Works on marine, air and pipeline transportation of hazardous materials are scarce.
- *Affected agents:* the risk can affect population (the population leaving the area, employees working in the area, people present in facilities like schools and people crossing the area), territorial infrastructures (railways, electric lines) and natural elements (water bodies, green areas).
- *Meteorological conditions:* is an important factor which impacts considerably on the dispersion of material [215].
- *Temporal factor:* many factors can change as a function of time, as the population distribution on day and night [95] for example. New technological advances in communication systems and Global Positioning System (GPS) are challenging researchers to develop routing models and robust optimization procedures that are able to respond quickly to changes in the data.

Many difficulties follow from the transportation of hazardous materials, it can be classified into two major components, the definition of the problem and the solution of the formulations induced. An accurate model for hazmat transportation must be technically and computationally tractable.

The definition of the transportation of hazardous materials problem is based on risk assessment, which can be qualitative or quantitative. Qualitative risk assessment deals with the identification of

possible accident events and attempts to estimate the undesirable consequences and quantitative risk assessment deals with the numerical assessment of frequencies and consequences of incidents [142].

According to the characteristics presented above, it is possible to consider the RHM problem as a multiobjective problem on stochastic dynamic networks. However, the multiobjective routing of hazardous materials on stochastic dynamic networks did not receive much attention despite its relevance.

## 7.4 Routing and scheduling in a time-dependent dependent environment

The main difference between this class of problems and the classical VRP is the definition of travel time. When in classical VRP, the travel time is a function of the distance, in RS_TDE problems, the travel time is variable and depends on many factors among which are weather conditions, congestion and the time of the day. We describe in this section three VRPs dealing with more realistic considerations of the travel time, the Time-dependent Vehicle Routing Problem (TDVRP), the Dynamic Vehicle Routing Problem (DVRP) and the Real-Time Vehicle Routing Problem (RTVRP). These problems are important not just because the consideration of the travel time variations affects considerably the objective values, but also because the best solutions known for non time-dependent problem are in general infeasible when applied in time-dependent world.

### 7.4.1 Environmental contribution

Most traffic networks in large cities face high utilization levels and congestion. As a result, the road traffic conditions and its resulting variability can not be ignored in order to carry out a good quality route optimization. When taking into account congestion, RS_TDE problems can be considered as major components for dealing with urban freight transportation problems associated with negative environmental impacts such as air pollution and noise.

### 7.4.2 Time-dependent vehicle routing problem

When the VRP assumes that the costs or travel times are a scalar transformation of distance, the TDVRP is more adapted to real applications by taking into account variations of the travel time resulting from periodic cycles in the average traffic volumes. It is considered in this problem that the principal variation in travel time results from the time-of-day variation, the travel time between two points $i$ and $j$ is a function of the time of the day at the origin point $i$. A variety of models for the TDVRP are considered in the literature. We present briefly a classification of these models [197]:

1. *Basic Models (BM):* Time-dependency is integrated in the model using simple rules like multiplier factors associated with different periods of the day. Unfortunately, these assumptions are weak approximations of the real-world conditions where travel times are subject to more subtle variations over time.
2. *Models based on Discrete Travel time and Cost Functions (MDTCF):* In this kind of formulations, the horizon of interest is discretized into small time intervals. The travel time and cost functions for each link are assumed to be step functions of the starting time at the origin node. However, the assumption that travel times vary in discrete jumps is just an approximation of real-world conditions since travel times change continuously over time. Many of these models are dedicated to time-dependent shortest path problem [96] and time-dependent traveling salesman problem [268]. In [224], the authors consider the problem of path planning in networks including multiple time dependent costs on the links and use the dynamic programming algorithm principle.

3. *Models based on Continuous Travel time and Cost Functions (MCTCF):* Continuous travel time functions seem to be more appropriate to model real-world conditions. Unfortunately, the models obtained are difficult to solve without simplifying assumptions, so these models consider again an approximation of the real travel time variations.

4. *Queueing Models (QM):* Here, the traffic congestion component is based on queueing theory. This allows one to capture the stochastic behavior of travel times by generating an analytical expression for the expected travel times [364].

Works in this field show experimentally that the total travel times can be improved significantly by explicitly taking into account congestion during the optimization. Very few comparative framework on different models and solution methods are found. We show on Table 7.3 some works on the TDVRP.

Table 7.3: Some works on the time-dependent vehicle routing problem

| References | Models | Models characteristics | Solution methods |
|---|---|---|---|
| Brown, Ellis, Graves and Ronen (1987) [88] | BM | A solution where travel time fluctuations are ignored is first produced. Then, the loads for each truck are resequenced "manually" to take into account various factors such as traffic congestion during rush hours, road and weather conditions. | A collection of integer programming methods. |
| Malandraki and Daskin (1992) [268] | MDTCF | The problem is formulated as a mixed integer programming problem. | Nearest-neighbour (greedy) heuristic is proposed, as well as a branch-and-cut algorithm for solving small problems with 10-25 nodes. |
| Hill and Benton (1992) [188] | MDTCF | The model was based on time-dependent travel speed. | Experimentations based on a small example with a single vehicle and five locations are given. |
| Ichoua, Gendreau and Potvin (2003) [197] | MDTCF | The model was based on time-dependent travel speed. | A taboo search heuristic is proposed and experimentations are performed on Solomon's 100-costumers problems. |
| Woensel, Kerbache, Peremans and Vandaele (2007) [364] | QM | - | Both the static and the dynamic TDVRP were solved using ant colony optimization. |
| Hashimoto, Yagiura and Ibaraki (2008) [184] | MDTCF | Travelling time and cost functions values are time-dependent. | A local search algorithm. |
| Donati, Montemanni, Casagrande, Rizzoli and Gambardella (2008) [126] | MDTCF | The model was based on time-dependent travel speed. | A multi-Ant Colony System. |

provides a more complete list of references.

### 7.4.3 Dynamic vehicle routing problem

The DVRP is the dynamic counterpart of the VRP, where information relevant to the planning of the routes can change after the initial routes have been constructed. This class of problems have arisen thanks to recent advances in communication and information technologies that allow information to be obtained and processed in real time.

Traditionally, the DVRP is solved in one of two ways: when problem parameters become (customers, speed, travel time) known throughout the run, oblivious online algorithms are used. Alternatively, when all parameters are available but with uncertainties in their properties, stochastic optimization is used, which build the routing plan a priori, and then modifies it when changes in parameters properties occur. Most research in this area has focused on dynamic routing and scheduling that considers the variation in customer demands. However, there has been limited research on routing and scheduling with congestion and travel times variation. We present below some models for the DVRP proposed in the litterature:

1. *Models based on Simulations (MS):* A vehicle routing and scheduling plan is obtained by revisiting the vehicle routing and scheduling plan computed on the previous day by using real-time information

of present link travel times, whenever a vehicle arrives at a customer. This real-time information is provided by dynamic traffic simulation based on the current conditions of the day [338].

2. *Queueing Models (QM):* To capture travel times, these models introduce a traffic congestion component based on queueing theory. A major advantage of using these queueing models is that the real-life physical characteristics of the road network can be mapped immediately into the parameters of the queueing model. Moreover, the inherent stochasticity of travel times can explicitly be taken into account via the analytical queueing models [365].

3. *Stochastic Models (SM):* The travel times are subjected to stochastic variations [304].

Due to the complexity of this problem, heuristic methods are often used to obtain good solutions, tabu search algorithms were used to solve QM [365], genetic algorithms to solve MS [338] and local search heuristics to solve SM [304]. A discussion of the DVRP solution methods can be found in [308].

### 7.4.4 The real-time vehicle routing problem

A new generation of VRP are proposed in the literature for vehicle routing in more realistic settings. The RTVRP considers more accurate information about the travel times compared to the DVRP by considering real-time variations in travel times [296, 375]. Dynamic vehicle routing models imply that a vehicle en route must first reach its current destination and only after that can it be diverted from its route. However, an unpredicted congestion or other traffic impediment can be encountered on the way to its immediate destination. Using mobile technology, vehicle routing can be modeled in more realistic settings:

- Allows locating vehicles in real time.
- Enables the online communication between the drivers and the dispatching center.
- Capable to capture varying traffic conditions in real time and in the short run predict with high accuracy the travel time between a pair of nodes.

All these factors allow to send new instructions to drivers at any time, regardless of their location and status. These modeling approaches enable a better approximation of the real-world conditions.

### 7.4.5 Summary

It was established that the highest emissions of carbon dioxide occur in congested, slow moving traffic. RS_TDE models can contribute indirectly to decrease fuel consumption and gas emissions [324]. In reality, the links have different combinations of congestion levels, and delays associated with road furniture such as traffic lights and roundabouts, and road topography and geometry such as inclines. This causes speed variations (resulting from acceleration and deceleration) and therefore produces different times over links with the same road category and distance. In addition, these speed variations would lead to fuel consumption and therefore gas emissions variations. Therefore, there is a need to calculate vehicle routes which minimize gas emissions, not just to calculate routes minimizing by time or distance.

## 7.5 Multi-modal routing problem

### 7.5.1 Description

This problem is defined as follows: given a set of origin-destination transport requests, one must optimally route these requests in a multi-modal network including a heterogeneous set of transportation services. These services are generally classified by according to two main characteristics: the departure time and the cost function. By the departure time characteristic, we can differentiate between timetable services (rail and short sea shipping) and time-flexible services (trucks). The cost (and duration) of routes depends on the departure time, the transportation mode, the distance and the waiting time in transshipment nodes. These constraints make the problem even more difficult.

### 7.5.2 Environmental contribution

Negative impacts caused by the transportation activities such as gas emission and noise can be reduced by using cleaner and silent alternative transportation modes. Multi-modal transportation strategies are studied in the operations research literature through the multi-modal routing problem.

### 7.5.3 Related works

In [98], the authors consider a multiobjective multimodal multicommodity flow problem with time windows and piecewise linear concave cost functions. Based on Lagrangian relaxation technique, the problem is broken into a set of smaller and easier subproblems and the subgradient optimization procedure is applied to solve the Lagrangian multipliers problem. Authors in [279] proposed an origin-destination integer multi-commodity flow formulation with non-convex piecewise linear costs and use column generation based heuristic that provides both lower bounds and good quality feasible solutions. The author in [278] deals with two objectives: the cost and the risk and develops a chance-constrained goal programming method to solve the problem.

Multi-modal transportation models need to define the optimization methods from the view of economic and environmental performances. When we are interested in multi-modal transportation, it is essential to make an adapted choice of transportation modes in such a way that the environmental impacts of the transportation system are minimized. In [80], the authors underlined the importance of choosing the transportation modes but little work is known on the calculation of transportation cost, taking environmental impacts into account. In [15], the transportation mode is considered within the framework of a green supply chain, the environmental impacts of the means of transport are integrated into the model as a cost aspect.

In multi-modal transportation, specific characteristics must be taken into account for each transportation mode. For example, the impact of variation in travel time on rail transport links and road transport links is very different, as rail transport mode is not subject to congestion. For determining efficiently the transportation mode for minimizing the environmental impacts, it is important to consider stochastic travel time. The multi-modal transportation network can be assumed in this case to be mixed, where the travel time on some arcs is stochastic.

## 7.6 Waste collection vehicle routing problem

### 7.6.1 Description

Waste Collection Vehicle Routing Problem (WCVRP) can be classified as variation of the VRP but with additional constraints. The major difference between WCVRP and the classical VRP are landfills constraints. When a vehicle is full, it needs to go to the closest available disposal facility. Each vehicle can make multiple disposal trips per day. Three categories of waste are considered in the literature: commercial waste (involves servicing customers such as restaurants and small office buildings), residential waste (involves servicing private homes) and roll-on-roll-off waste (commonly used for construction site waste), these three categories bring about three different waste collection strategies. While works on the VRP consider the major objective of minimizing the travel cost, this problem also considers route compactness (a solution in which many routes cross over each other is less compact than one in which no routes overlap) and work balancing among vehicles.

### 7.6.2 Environmental contribution

In recent years waste management has become an area of concern for municipalities worldwide due to population growth, environmental concerns and the progressive increase in waste management cost. Waste collection is one of its main components. Note here that authors in [324] have discussed the importance attached to waste management and collection in terms of the "green logistics" agenda.

### 7.6.3 Related works

The particularity of residential routes compared to commercial and roll-on-roll-off ones is the mandatory adherence to driving on one side of the street. Unlike drivers on commercial or roll-on-roll-off routes, those on residential routes are permitted to serve only customers on the right side of the street. Very few exceptions are granted for alleys and one-way streets. Commercial and roll-on-roll-off waste collection differ principally from the size of the container.

Routing problems in waste collection applications cannot typically be modelled with a unique routing model. As commercial and roll-on-roll-off waste routing consists of point-to-point collections, it can be modeled by node routing (VRP) models. Commercial waste routing problem can be characterized as a VRP with time windows (VRPTW) (Section 7.2) since commercial waste collection stops may have time windows. A special VRP variant known as rollon-rolloff vehicle routing problem was dedicated to roll-on-roll-off waste. However, residential routes require arc routing models, where costumers are located on the arcs. The periodic vehicle routing problem is also studied in the context of waste collection when collection operations are periodic on the time horizon. As these problems are computationnaly very hard, and can not be solved by optimal (exact) methods in tractice, heuristics are used in this purpose.

**Vehicle routing problem with time windows:** The majority of papers in the literature are case study papers, focusing on results obtained when algorithms are applied to real-world data [294, 321, 349]. Only a few of these papers report computational experience with publicly available waste collection test instances [217]. More references are given in these cited papers.

**The periodic vehicle routing problem:** This problem has a horizon $T$, and there is a frequency for each customer stating how often within this $T$ period this customer must be visited. A solution to the problem consists of $T$ sets of routes that jointly satisfy the demand constraints and the frequency constraints [16].

**The capacitated arc routing problem:** In this problem [129], a fleet of vehicles, all of them located at a central depot and with a known capacity, must serve a set of streets network, with minimum total cost such that the load assigned to each vehicle does not exceed its capacity.

**The rollon-rolloff vehicle routing problem:** In this problem, tractors move large trailers between locations and a disposal facility. The trailers are so large that the tractor can only transport one trailer at a time [47, 73, 274].

### 7.6.4 Summary

Waste collection real-life problems have almost been studied in a off-line context, where it is assumed that all data about the problem are known in advance. However, this is not necessarily the case when some information might not be readily available when the vehicles start their routes. When new information is available as the routes are executed, the problem becomes dynamic, this problem have not attracted yet the attention of waste collection research community.

## 7.7 Dial-a-ride problem

### 7.7.1 Description

The Dial-a-Ride Problem (DARP) consists of designing vehicle routes and schedules for $n$ costumers from their pickup point to their delivery point. The costumer requests the service by calling a central unit and specifying the origin and destination points, the number of passengers and some limitations in service time (the earliest departure time for example). The transport is supplied by a fleet of $m$ identical vehicles based at the same depot. The aim is to plan a set of minimum cost vehicles routes capable of accommodating as many requests as possible, under a set of constraints. The DARP can be static or dynamic. In the first case, the costumer asks for service in advance and the vehicles are routed before the system starts to operate. In the second case, requests are gradually revealed throughout the day and vehicle routes are adjusted in real-time to meet demand.

### 7.7.2 Environmental contribution

The main original target of the DARP is to offer the comfort and flexibility of private cars and taxis at a lower cost. This problem is suited to service sparsely populated areas, to low demand periods or to special classes of passengers with specific requirements (elderly, disabled). In addition, this problem considers indirectly environmental savings, indeed in opposition to individual taxis, the grouped ones can decrease the traffic density particularly in large cities.

### 7.7.3 Related works

The DARP is characterized by multiple objectives such as the maximization of the number of costumers served, the minimization of the number of vehicles used and the maximization of the level of service provided on average to the costumer (costumer waiting time, total time spent in vehicles, difference between actual and desired drop-off times). The DARP can be formulated as multiobjective mixed integer program. Exact algorithms for the single-vehicle DARP have been developed in [122, 307].

Recently, a branch-and-cut algorithm has been proposed in [105]. Heuristics and meta-heuristics are proposed for dealing with the dynamic problem with time-dependent network [160]. For a recent overview of the DARP, see [106].

In [302], the authors propose a heuristic two-phase solution procedure for the dial-a-ride problem with two objectives. The first phase consists of an iterated variable neighborhood search-based heuristic, generating approximate weighted sum solutions and the second phase is a path relinking module, computing additional efficient solutions.

## 7.8 Pick-up and delivery vehicle routing problem

### 7.8.1 Description

In the Pick-up and Delivery Vehicle Routing Problem (PDVRP), a set of routes has to be constructed in order to satisfy transportation requests. A fleet of vehicles is available in a central depot to operate the routes and each vehicle has a given capacity. Each transportation request specifies the size of the load to be transported, the locations where it is picked up and the locations where it is delivered. Each load has to be transported by one vehicle from its set of origins to its set of destinations without any transshipment at other locations. The DARP (Section 7.7) generalizes the PDVRP [106], the main difference between these problems is the human perspective; the level of service criteria is more important in the DARP.

### 7.8.2 Environmental contribution

As mentioned in section 7.6, more and more countries have devoted considerable investments to waste reduction and material recycling. The existence of a backward flow of objects to be collected, stored, disassembled and recycled makes unprofitable to manage separately the forward flow of goods, from the producer to the consumer, and the backward flow of waste or used-up devices, from the consumer to recycling or dumping facilities. In addition, when the reuse of products and materials becomes cheaper than simply disposing them, both of the opposite flows concern the producer, instead of being managed by independent subjects. When pick-ups concern waste, the PDVRP can be considered as a green routing and scheduling problem. This model derive from the development of reverse logistics, which consists of the efficient integration of the forward flow of goods with the backward flow of waste [151]. In their survey [324] underline the importance of reverse logistics in green Logistics, but the transportation aspect was not discussed.

### 7.8.3 Related works

A comprehensive survey on the PDVRP can be found in [301] where different variants of the problem, models and resolution methods are presented. As our knowledge, no work in the literature treat real world pick-up and delivery problem in the context of recuperation of waste for recycling.

### 7.8.4 Discussion

We present here some applications of this problem:

- The door-to-door delivery of mineral water bottles and the simultaneous collection of empty bottles.

- The laundry service for hotels (collecting dirty clothes and delivering clean clothes).
- Medical Waste.

It is important to attach more interest to real problems for evaluate economical and environmental savings induced by these systems. The study of economical impacts of the integration of waste collection with products distribution can encourage industrials to recuperate the unused waste of their products and permits to the reduction of amounts of waste treated by municipalities and environmental saves.

## 7.9 Energy routing problems

### 7.9.1 Description

The Energy Routing Problem (ERP) was defined in the context of electric vehicles routing in [22]. One of the major characteristics of electric vehicles is their ability to recover braking energy to be restored to the battery. This problem consists of designing routes of maximum autonomy for vehicles with rechargeable batteries taking into account energy recuperation during deceleration phases.

### 7.9.2 Environmental contribution

Electric vehicles are generally considered as the cleanest transportation option due to their zero local and potentially minor green house gas emissions.

Advantages: electric vehicles provide zero local gas emission in congested environment, this is why electric vehicles can be deployed to reduce pollution and noise in large cities.

Inconvenient: The global electric vehicle use. Pollution due to batteries and the resource energy used for producing electricity. Example: a good alternative in France, 3/4 of electric production uses nuclear energy.

Efficient EV routing is critical to the operational profitability and customer satisfaction of EV use, especially in light of highly concurrence with fuel of combustion engined vehicles. Whereas the development of clean transportation (hydrogen and fuel cells, biofuel powered vehicles, hybrid vehicles, low-carbon fueled vehicles. . . ) grow extensively, much less research is dedicated to the energy efficient routing problem for clean vehicles (electric vehicles, cycles).

### 7.9.3 Related works

Very scarce works interest to this problem in the literature. Electric vehicles can be used on two different contexts, (1) for the transportation of persons, where maximum autonomous routes have to found for traveling between origin-destination nodes. In [22], the authors formalize the ERP as a generalization of the shortest path problem with hard constraints (which impose that the battery cannot be discharged below zero) and soft constraints (which impose that the battery cannot store more energy than its maximum capacity). A generic shortest path algorithm was proposed to solve the problem. The authors developed a prototypic software system for energy efficient routing where data is obtained by combining geospatial data (of OpenStreetMap) and elevation data (of the NASA Shuttle Radar Topographic Mission), (2) for an alternative transportation mode use in distribution systems, this problem can be formulated as vehicle routing and scheduling problem (section 7.2) with additional objectives and constraints related to energy consumption.

Because the ERP is a relatively new problem, we develop its mathematical formulation and show its interactions with other routing and scheduling problems. This problem can be modeled with a directed graph $G = (N, A)$, where $N = M \cup \{s, t\}$, $M$ is the set of costumers, $s$ and $t$ the source and destination nodes respectively and $A$ represents the set of arcs. A set $K$ of identical vehicles are available, each one has a maximum capacity $Q$. With each arc $(i, j)$ is associated a positive (resp. negative) value $e_{ij}$ indicating consumption (resp. gain) of energy on arc $(i, j)$. The battery of each vehicle has a maximum capacity $C$ and cannot be discharged below zero. The ERP consists of finding a set of minimum cost origin-destination routes, such that each costumer $i \in M$ is visited by exactly one vehicle to satisfy a specific demand $d_i$. The total customers demand satisfied by the same vehicle must not exceed the vehicle capacity. The formulation of the ERP can be given as follows:

$$\min \quad F(x, u) \tag{7.1}$$

$$\text{s.c.} \quad \sum_{k \in K} \sum_{j \in M} x_{ij}^k = 1, \forall i \in N \tag{7.2}$$

$$\sum_{(i,j) \in A} d_i x_{ij}^k \leq Q, \forall k \in K \tag{7.3}$$

$$\sum_{i \in M} x_{si}^k = \sum_{i \in M} x_{it}^k = 1, \forall k \in K \tag{7.4}$$

$$\sum_{j \in N} x_{ij}^k - \sum_{j \in N} x_{ji}^k = 0, \forall i \in M, \forall k \in K \tag{7.5}$$

$$x_{ij}^k (u_j^k - u_i^k - e_{ij}) \geq 0, \forall (i, j) \in A, \forall k \in K \tag{7.6}$$

$$0 \leq u_i^k \leq C, \forall i \in N, \forall k \in K \tag{7.7}$$

$$x_{ij}^k \in \{0, 1\}, \forall (i, j) \in A, \forall k \in K \tag{7.8}$$

$$u_i^k \in \mathbb{N}^+, \forall i \in N, \forall k \in K \tag{7.9}$$

where $x_{ij}^k$ indicate whether arc $(i, j)$ is used or not by vehicle $k$ and $u_i^k$ is the battery remaining storage capacity of vehicle $k$ at node $i$. The objective function (7.1) can be considered as the sum of the remaining battery storage capacity of all vehicles ($\sum_{k \in K} u_t^k$) or the maximum remaining battery storage capacity of all vehicles ($\max_{k \in K} u_k^k$) at the destination node. Constraints (7.2) ensure that each costumer is visited exactly once. Constraints (7.3) ensure that demand of each route is within the capacity limit of the vehicle serving the route. Constraints (7.4)-(7.5) are path constraints. Constraints (7.6)-(7.7) ensure compatible remaining storage capacity at each node for each vehicle. As energy constraints can be considered as particular resource constraints, we can identify some relationships between ERP and other routing and scheduling problems:

- The ERP can be considered as a generalization of the vehicle routing problem with time windows where the objective is to minimize the total travel time of all vehicles or the minimization of the makespan (minimization of the maximum total travel time). $e_{ij}$ represents the travel time on arc $(i, j)$ and the time window at each node is equivalent to $[0, C]$. Hard and soft constraints are equivalent to time windows constraints; Hard constraints express the fact that a vehicle must arrive at the costumer before the end of time window and soft constraints express the fact that if a vehicle arrives before the beginning of the time window, it wait at no cost. In a more general case [22], the EVRP is a generalization of the routing problem with resource constraints [203] where resource consumption represents energy consumption, the amount of available resource is equal to $C$, and the residual resource at node $i$ is equivalent to $u_i$.

- this problem is a special case of a much studied problem in the literature, the Pick-up and Delivery Problem (PDP) [301], where pick-up are not hard constraints: partial pick-ups are allowed and

pick-up is not performed when the vehicle is full. Acceleration (resp. deceleration) phase can be considered as a delivery (resp. pick-up) operation.

Lets $\hat{G} = (\hat{N}, \hat{A})$ be a directed graph where $\hat{N}$ is the set of nodes and $\hat{A}$ is the set of arcs. $\hat{G}$ is constructed using the graph $G$ as follows: let be $P(i)$ the set of predecessor nodes of node $i \in N$, $|P(i)| = l^i$. Each node $i \in N$ is duplicated into $l^i$ nodes $\hat{N}_i = \{i^1, ..., i^{l^i}\}$, so, $\hat{N} = \bigcup_{i \in N} \hat{N}_i$. For each new node, we assign an incoming arc from one predecessor node:

$\hat{A} = \bigcup_{i \in N} \{\hat{A}(j^1, i^1) \bigcup \hat{A}(j^2, i^2) ... \bigcup \hat{A}(j^{l^i}, i^{l^i}) \mid j^1, ... j^{l^i} \in P(i)\}$.

where $A(j^1, i^1) = \bigcup_{k \in \hat{N}_{j^1}} \{(k, i^1)\}$. For each node $i \in N$, we associate to the duplicated nodes $i^1, ..., i^{l^i}$ the weights $q^1 = e_{j^1 i}, ..., q^{l^i} = e_{j^{l^i} i}$. Node $i \in \hat{N}$ represents a pick-up (resp. delivery) node if $q_i < 0$ (resp. $q_i > 0$).

Optimizing the vehicle braking routes, through the maximization of final remaining energy, can result in an increase of the travel time. To make energy vehicle routing effective, the travel time has to be taken into account in the optimization model. As our knowledge, no work considers the travel time in energy routing problem. Such multiobjective problem can be solved using adaptation of multiobjective routing methods [209, 339].

### 7.9.4 Discussion

In practice, the road links have different combinations of energy Consumption/Recuperation (C/R) levels and delays, associated with road furniture such as traffic lights and roundabouts, and road topography and geometry such as inclines. This causes energy C/R variations (resulting from acceleration and deceleration) over links with the same road category and distance. Therefore, instead of constant values, more realistic considerations of energy C/R have to be established. On one hand, energy C/R can be a function of speed variations over links. A number of works in the literature interest to model/estimate travel speed [188], these works can be exploited to estimate energy C/R. On another hand, energy C/R has to be considered as stochastic instead of constant values for assessing the risk of running out of energy before arriving at the destination. No work in the literature tackle these issues yet.

The global critical problem in the electric vehicles promotion is a chicken-and-egg infrastructure dilemma [359]: *consumers will be reluctant to purchase vehicles until a sufficient number of refueling stations has been installed, while vehicle manufactures will not produce vehicles that consumers will not buy, and fuel providers will not invest in a new energy infrastructure until there is sufficient demand for it. Therefore, it is likely that governments will need to play a significant role in promoting any change to alternative fuels, although public support alone will not ensure the success of this transformation.*

## 7.10 Air traffic management

Air traffic management (ATM) consists of two important components: the traffic planning and the traffic control. Traffic planning deals with the balance between demand and the available capacity and traffic control has to guide aircraft safely to their destinations.

### 7.10.1 Environmental contribution

Given the forecast growth in aviation over the next decade there is an urgent need of air traffic control decision-support to address the problem of congestion in the airspace. Air traffic delays due to congestion are a source of unnecessary cost for airline companies and passengers. Delays also have an environmental

cost. Because of congestion, aircraft are often forced to fly far from the cruise altitude and/or the cruise speed for which they are designed, this results in unnecessary fuel burn and gas emissions.

### 7.10.2 Related works

An aircraft conflict occurs when the distance between two or more aircrafts falls below a given threshold. In this case, a minimum separation is required. Aircraft conflict detection and resolution has been widely studied in the literature [228]. In [299], the authors propose an integer programming model that minimizes the maximum deviation in the changes made, by assuming that aircraft can perform either a speed change or a heading change. Authors in [101] studied the traffic control problem by maintaining separation while considering associated fuel costs with any heading deviation or speed changes. Safety requirements are considered as hard constraints that must be maintained. The objective function focuses on minimizing fuel costs, and hence the resulting environmental impact. In [58], the authors address the problem of determining how to reroute aircrafts in the air traffic control system when faced with dynamically changing weather conditions. The overall objective of this problem is the minimization of delay costs.

The multiobjective aspect of the problem was considered in [346] where a multi-objective genetic algorithm has been designed to solve the model. Three objectives have been considered: the minimization of the sum of the flights which exceeds the capacities of all the sectors, the minimization of the sum of the maximum flights which exceeds the capacities of all the sectors and the minimization of the total delay time including the ground delay time and air delay time. In [365] the authors take into account the weather-related flight delays and investigate the problem of generating optimal weather avoidance routes under hazardous weather conditions. The proposed model minimizes the fuel usage, weather conditions, customer comfort and traffic density and is solved using a hybrid ant colony optimization method with a multi-agent approach.

## 7.11 Green vehicle routing and scheduling class problems

The GVRSP was defined in Section 7.1.1 as a classical VRP with additional environmental and social objectives and constraints, this radically changes the problem structure, so different and generally dedicated solution methods are developed. As a consequence, we envision a new era in which optimization systems will not only allocate ressources for optimizing economical costs: they will react and adapt to external events efficiently under environmental constraints and objectives. The GVRSP can be defined as the computation of a set of origin-destination paths optimizing a set of objectives $\mathcal{F}$ by satisfying a set of constraints $C$. We summarize in Table 7.4 some general characteristics of the presented variants of the GVRSP.

### 7.11.1 Classification

Green routing and scheduling class includes a large variety of problems, which differ by their environmental contributions and the associated models and resolution methods. This makes difficult their classification, we present in the next three classifications of GRSP:

- **Environmental contributions**: We distinguish two categories:

  1. Problems which are in nature green routing and scheduling problems: RHM, RS_TDE, DARP, CARP, ERP and ATM. These problems ensure directly (RHM, CARP, ERP) or indirectly (RS_TDE, DARP, ATM) environmental savings.

Table 7.4: Some variants of the Green Vehicle Routing and Scheduling Problem

| Problem | $\mathcal{F}$ | $C$ | Network characteristics | Models and optimization methods | |
|---|---|---|---|---|---|
| | | | | MultiObjective | SingleObjective |
| RHM | Min. risk<br>Min. travel time<br>Max. risk equity<br>Min. travel cost | (1) Origin-destination paths<br>(2) Risk equity<br>(3) Vehicle capacity | Static network<br>Stochastic dynamic network | Weighting method [3, 108, 316, 330, 386]<br>Stochastic dominance [350]<br>Branch-and-bound [215]<br>Goal Programming [163]<br>Insertion heuristic [386] | Shortest path algorithms [6, 208, 227, 263]<br>Lagrangian relaxation [166]<br>Multi-commodity flow model [92]<br>Bilevel flow model [63] |
| TDVRP | Min. travel time<br>Min. travel cost | (1) Origin-destination paths<br>(2) Each costumer must be serviced by one vehicle<br>(3) Vehicle capacity | The travel time or the travel speed on links is a function of the time-of-day | Dynamic programming [224] | Nearest-neighbour heuristic & branch-and-cut algorithm [268]<br>Taboo search heuristic [197]<br>Ant colony optimization [126, 364]<br>Local search algorithm [184] |
| DVRP | Min. travel time<br>Min. travel cost<br>Min. vehicle number | (1) Origin-destination paths<br>(2) Each costumer must be serviced by one vehicle<br>(3) Vehicle capacity | The travel time on links are stochastic varying over the time | Ant colony optimization [211]<br>Hybrid dynamic programming - ant colony optimisation [100] | Tabu search algorithms [365]<br>Genetic algorithms [338]<br>Local search heuristics [304] |
| RTVRP | Min. travel time | (1) Origin-destination paths<br>(2) Each costumer must be serviced by one vehicle<br>(3) Vehicle capacity | The travel time on links are real time travel time | | Genetic algorithm [296] |
| CARP | Min. travel time<br>Min. makespan<br>Min. travel cost | (1) Origin-destination paths<br>(2) Each link must be serviced by one vehicle<br>(3) Vehicle capacity | Static network | Genetic algorithm [232]<br>Memetic algorithm [272]<br>Epsilon-constraint method [168] | Tabu search algorithm [169, 187]<br>Genetic and memetic algorithms [231]<br>Guided local search algorithm [59]<br>Ant colony optimization [233] |
| MMVRP | Min. travel time<br>Max. cleaner transport mean<br>Min. travel cost | (1) Origin-destination paths<br>(2) Vehicle capacity | Static multi-modal network | Chance-constrained goal programming [278]<br>Subgradient optimization [98] | Double-sweep method [72]<br>Column generation [279] |
| DARP | Max. costumers served<br>Min. vehicles number<br>Max. level of service | (1) Origin-destination paths<br>(2) A new costumer is accepted or rejected (3) Vehicle capacity<br>(4) Pairing and ride time | Static network<br>Dynamic network | Variable neighborhood-based heuristic [302] | Dynamic programming [122, 307]<br>Branch-and-cut algorithm [105] |
| PDVRP | Min. travel cost<br>Min. travel time<br>Min. vehicles number | (1) Origin-destination paths<br>(2) Capacity constraints<br>(3) Precedence constraints | Static network | Insertion based algorithm [267] | Genetic algorithm [212] |
| ERP | Min. energy used | (1) Origin-destination paths<br>(2) The battery cannot be discharged below zero<br>(3) Battery energy capacity | The energy consumption/recuperation is evaluted on each link | | Shortest-path algorithm [22] |
| ATM | Min. delays<br>Min. fuel costs<br>Min. Max. deviation | (1) Origin-destination paths<br>(2) Separation constraints<br>(3) Bounds on travel time and distances of the new routes | Multi-dimentional network | Genetic algorithm [346]<br>Hybrid ant colony optimization with a multi-agent approach [365] | Mixed-integer linear program-optimization software [299]<br>Lagrangian relaxation-heuristics [58] |

2. Problems whose some applications are green routing and scheduling problems: MMVRP and PDVRP. The MMVRP can be exploited for choosing the cleanest transportation modes during the transportation process and in the PDVRP, pick-up can concern backward flow of waste that are collected and recycled.

- **Models induced**: We dstinguish two major models. Routing and scheduling problems without intermediate stops (includes THM and an ERP version) and routing and scheduling problems with intermediate stops (includes RS_TDE, CARP, DARP, MMVRP, PDVRP, ATM and an ERP version).

- **Transportation type**: We distinguish four classes: road transportation (THM, RS_TDE, CARP, DARP, PDVRP and ERP), airline transportation (ATM), sea transportation and multimodal transportation (MMVRP).

### 7.11.2 Optimization

Due to the conflicting nature of the criteria in sustainable transportation problems (economic, environmental and social), multiobjective optimization represents a major component. The particularity of these problems is that a unique feasible solution optimizing all the criteria does not exist. To obtain the optimal solutions, one just needs to consider Pareto optimal solutions. We can find in the literature two classes of multiobjective optimization methods, based on the problem elements type: *deterministic methods* are the most studied in the literature and consider that all problem elements are deterministic and *stochastic methods* consider that some elements of the problem are uncertain, these elements are modeled using random variables.

The unpredictable nature of transportation leads to many stochastic elements in the problem like the travel time, speed, traffic congestion, weather conditions, the amount of population present near a route and the effects of an accident. Deterministic methods use approximations of these elements and forcasts and sometimes lead to infeasible schedules and poor decisions. Recent research [186] shown the benefits of adaptability for vehicle routing and scheduling, exploiting stochastic information to produce better solutions.

Generally, the uncertainty can be in the presence or absence of the customers, in the quantity of the their orders, and in the travel and service times. Routing and scheduling when demands are stochastic have been extensively studied when less attention is given to the case of stochastic travel time or stochastic speed. Especially with respect to travel times, variability is generally reduced te be nearly constant within time periods in a day. Such a characterisation of travel times leads to the TDVRP variant (section 7.4.2). The stochastic version of multiobjective routing and scheduling is more studied in the context of transportation of hazardous materials [97, 246, 277, 350, 362], and essentially for the computation of shortest path problems. Very scarce work in the literature interest on the stochastic version of multiobjective vehicle routing and scheduling problem [216, 234].

As we can observe in table 7.4, heuristic and metaheuristic methods are the most used for solving multiobjective problems, we observe also that multiobjective problems are less studied than single objective problems (as our knowledge, no work in the literature leads with the multiobjective RTVRP and the multiobjective ERP). Evolutionary algorithms are one of the most popular methods for solving multiobjective routing problems [209, 210, 337], these methods have been hybridized with local searches, heuristics, and/or exact methods for the problem resolution [210, 337]. Other optimization methods were proposed in the literature for the resolution of multiobjective problems, based on genetic algorithms, lexicographic strategies, ant colony mechanisms, or specific heuristics [209], but a very limited works deal with the stochastic version of the problem.

## 7.12 Conclusion

This report aims at specifying the contribution of Combinatorial Optimization (CO) to environmental transportation. To this purpose a framework for identifying relevant related problems treated in the CO field is proposed, some of the literature has been reviewed and discussed with respect to both models and optimization aspects. It can be observed that during the last few years, CO for transportation problems has extended its scope to include environmental applications.

It can be deduced that relationship between CO and environmental transportation is interactive in the sense that from the complexity of the issues examined stems the need to develop and adapt specific methodological tools. We synthesize in this section the material we have reviewed for the green routing and scheduling by summarizing some fundamental characteristics of this class of problems.

*A recent research area.* Green routing is a relatively recent problem since the consideration of environmental impacts in the models is a new issue. Several small research communities in this field work on their own problems, this causes a lack of common problem definitions, hypothesis, definitions and concepts. Decision-making in environmental transportation can be complex and seemingly intractable, principally because of the inherent trade-offs between socio-political, environmental, ecological, and economic factors. A balance has to be found between the complexity of the real world operation and the level at which the model is developed and also takes into account model accuracy and computational efficiency.

*Evaluation of environmental impacts.* The integration of the environmental costs of transportation is rarely quoted in the literature. Fuel consumption and emissions are complex to estimate and are a function of several variables (type of vehicle, speed, acceleration rates, and meteorological conditions).

*Uncertainty.* Many authors recognized the uncertain nature of some characteristics in transportation area (the number of incidents on a road and the travel time for example), these characteristics can be modeled by means of random variables whose distributions may vary over time. However, variable distributions are hard to determine in many cases due to the scarcity of data. The exact probabilistic expressions are usually too complicated, which results in the use of approximations for optimization. Hence, expertise is needed for understanding well probabilistic modeling to capture the important aspects of the activity, in addition, competence is also needed on optimization techniques to decide which approximations are necessary and which tools to use. Due to uncertainty, the path attribute values may not be simply additive across arcs in the path, causing the criterion to be non-order-preserving. This prevents the use of traditional dynamic programming techniques in the solution method. To deal with this, methods have to be developed that work with non-order-preserving criterion.

*Speed variations.* The existing research and methods (TDVRP and DVRP for example) that address the issue of congestion for decreasing the fuel consumption are limited to calculations that only produce estimates based on the travel time. It was shown in [300] that this does not produce a realistic estimate of CO2 emissions, nor does it produce routes which minimize these emissions. As the consideration of uncertainty is a predominant aspect to be considered in GVRSP, the speed variations is also a major aspect, it is important to consider methods that improve the accuracy of road speeds and incorporate speed variability as a factor in the model for the construction of better routes in both time and fuel efficiency. Since gas emissions depends on speed, acceleration and traffic volume along a link, it is therefore necessary to understand the relationships between these characteristics in order to obtain a good estimate of gas emissions.

*Multi-criteria approaches.* The green vehicle routing problem is a typical multiobjective problem. When we deal with multiple routes (which is the case in practice), multiobjective shortest path problems are of limited efficiency, because we relax the interaction between vehicles (whereas it uses the same network). More adapted models are global routing where a fleet of vehicles (identical or heterogeneous) have to be routed through the network. With environmental issues assuming greater importance, it is desirable to consider global multiobjective methods with uncertain and time-varying criterion as travel time and

speed.

*Real-time information.* There is an increase on utilization of video surveillance, global positioning systems and communication equipment installed on all vehicles. It allows the implementation of more efficient real-time decision-making. A topic of interest in green routing and scheduling is the potential for real-time rerouting in order to react to changing conditions.

# Chapter 8

# The price of equity in the Hazmat Transportation Problem

Fabio Roda, Pierre Hansen, Leo Liberti

*Submitted to CTW International Workshop 2011 on Graphs and Combinatorial Optimization*

*"Equity" is not an easy term to define. In the context of transportation of hazardous materials, different definitions of equity lead to different situations as concerns transportation costs and the risk connected to loss of (quality of) life following catastrophic accidents. Specifically, regional fairness costs lives with respect to individual fairness. We discuss a methodology for trying to peg a monetary cost to this trade-off.*

## 8.1 Introduction

We consider the problem of the transportation of hazardous materials on a road network (Hazardous Materials Transportation Problem). We can figure it this way: there are $N$ trucks which have to transport some kind of dangerous material from one or many production points to one or many garbage dumps and we have to select a set of paths which is optimal from the point of view of risk, cost and equity. The optimization of cost and risk on a network leads quite spontaneously to shortest path and flow problems which are milestones of Operational Research, but equity is somehow unsual and hard to define. We consider and compare two different ideas. The first approach simply requires that all the areas involved in the transportation network share the same level of risk. This is a fair and intuitive idea but it could also lead to "improper" solutions where risk is equal but uniformly high. The second (more interesting) definition of equity we use is inspired by the concept of fairness of J. Rawls [312, 313, 314]. Basically, in this context, the difference principle means that we may introduce disparities only if they advantage the worst-off, namely reduce the risk of the less favourite area (the most exposed to the risk).

The aim of this work is to provide rational elements to be able to estimate the cost of choosing a particular definition of equity (for hazmat trasportation). We investigate the relation between each definition of equity and the cost it generates. This can be used as a first criterion to make a choice.

## 8.2 Mathematical programming formulation

Let $G = (V, A)$ be a directed graph, modelling a road network.
We consider many origin-destination pairs $(s,t) \in C \subseteq V \times V$. For every pair $(s,t)$ there is a commodity to be trasported from a source $s$ to a destination $t$ to respond to a specific demand which we indicate with $d_{st}$. We look for a global route planning given by a multicommodity flow function $x : C \times A \to \mathbb{R}_+$ (the situation involving only one origin and one destination is a special case). Typically we can imagine that the road network covers a geographic area which is divided into zones; in particular each arc (road) belongs to a zone $\zeta \in Z$. For the sake of simplicity we assume that each arc belongs to only one zone. Each arc $(i, j)$ has a positive traversal cost $c_{ij}$, a probability $p_{ij}$ of an accident occurring on that arc, a value of damage (in monetary units) $\Delta_{ij}$ caused by a potential accident on that arc and a capacity $\chi_{ij}$.

1. **Sets**:
   - $C \subseteq V \times V$ is the set of all pairs $(s,t)$;
   - $Z$ is the set of all zones;
   - $\zeta_l \subseteq A$ is a zone $(1 \leq l \leq |Z|)$;
2. **Parameters**:
   - $1 \leq l \leq |Z|$ = zone index
   - $p_{ij}^{st}$ : probability of accident on an arc;
   - $\Delta_{ij}^{st}$ : *damage* (in monetary units) caused by an accident on an arc ;
   - $c_{ij}^{st}$ : cost on an arc;
   - $s$ : source;
   - $t$ : destination (target);
   - $d_{st}$ : demand of commodity $(st)$;

   We call $p_{ij}^{st}\Delta_{ij}^{st}$ *traditional risk* and we indicate it, alternatively, as $r_{ij}^{st}$.
3. **Decision variables**:
   $\forall (i,j) \in A,\ \forall (s,t) \in C\ \ x_{ij}^{st}$ : flow of the commodity $(st)$ on the arc $(i, j)$
4. **Constraints**.
   - (capacity) $\sum_{(st)\in C}\ x_{ij}^{st} \leq c(ij)$
   - (demand) $\sum_{(i)\in V} x_{it}^{st} = d_{st}$
   - (flow conservation) $\forall (st) \in C \quad \sum\limits_{(i,j)\in A} x_{ij}^{st} - \sum\limits_{(j,i)\in A} x_{ij}^{st} = \begin{cases} 1 & \text{if } i = s \\ -1 & \text{if } i = t \\ 0 & \text{otherwise} \end{cases}$
5. **Objective function 1 (Cost):** minimize total cost
$$\min \quad \sum_{(i,j)\in A} \sum_{(s,t)\in C} c_{ij}^{st} x_{ij}^{st} \tag{8.1}$$
6. **Objectives concerning equity, two possible versions:**

- **Objective function 2 (a) (Risk sharing):** minimize the difference of traditional risk between two zones

$$\min \quad \left( \sum_{\forall(\zeta_l,\zeta_m)\in Z\times Z} \left| \left( \sum_{(i,j)\in\zeta_l} \sum_{(s,t)\in C} r_{ij}^{st} x_{ij}^{st} \right) - \left( \sum_{(h,l)\in\zeta_m} \sum_{(s,t)\in C} r_{hl}^{st} x_{ij}^{st} \right) \right| \right) \tag{8.2}$$

- **Objective function 2 (b) (Rawls' principle)**: minimize the traditional risk of the least advantaged zone

$$\min \left( \max_{\zeta\in Z} \quad \sum_{(i,j)\in\zeta} \sum_{(s,t)\in C} r_{ij}^{st} x_{ij}^{st} \right) \tag{8.3}$$

## 8.3 Methodology and Tests

The problem we are considering belongs to the special class of optimization problems called Multicriteria Optimization Problems (MOP) [133, 134, 275]. There are many methods we can apply to MOP, but a full examination of all of them is out of the scope of this work. We briefly recall only the main features of the *$\epsilon$-constraint method* which we use in our investigation. The basic idea of this method consists in the trasformation of all the objectives in constraints, out of one which is minimized (or maximized). Varying $\epsilon_i$, alternative solutions are obtained (even if it is known that it is difficult to chose proper values for the vector $\epsilon$ and arbitrarily ones produce no feasible solutions).

We consider first small instances of the problem based on networks composed by a kept down number of nodes and involving a few zones and shipments and only one origin and destination. We used the AMPL modelling environment and the off-the-shelf CPLEX 10.1 solver running on a 64-bit 2.1 GHz Intel Core2 CPU with 4GB RAM. The results we got using an instance composed by 15 nodes distributed in 2 zones, with 10 shipments show that both kind of equity have a negative impact on the cost and make it grow, which is the awaited outcome. The aim is to establish which one makes it increase most. In order to solve this question we introduce some methodological expedients. In fact, even if we solved either $C_1$ and $C_2$ applying the *$\epsilon$-constraint method*, we can not simply compare the cost for a fixed equal threshold of $\epsilon$ because it has a different meaning in the two situations. We have to normalize the comparison to "equal levels" of equity and to map the cost to the *share* of equity instead of its absolute value. For example, we compare the cost we get when we have the peak of equity in Risk Sharing and Rawls sense, then when we get the 99% of equity (independently from the different corresponding values of $\epsilon$ which are different in $C_1$ and $C_2$), the 98% ... and so on. Thus we establish first the maximal possible level of equity and then we (can) define the values corresponding to its fractions. We measure the increment of cost while equity varies from its possible minimum to its possible maximum and we map it on the share of equity. We discover that the raise of cost induced by equity in the sense of Rawls is weaker that the one induced by the "naive" one. We report some sample results (in the format [Equity Share; Cost of Risk Sharing; Cost of Rawls' Principle] : [0;40;40], [50;41;41], [60;42;42], [70;43;43], [80;45;44], [90;50;45], [95;53;49], [96;55;52], [97;57;54], [98;58;55], [99;60;58], [100;67;67]). Fig.1 shows the corresponding plot. The tests are partial since we use small artificial instances. We plan to use real data and different multi objectives methods in future work to corroborate our conclusions.

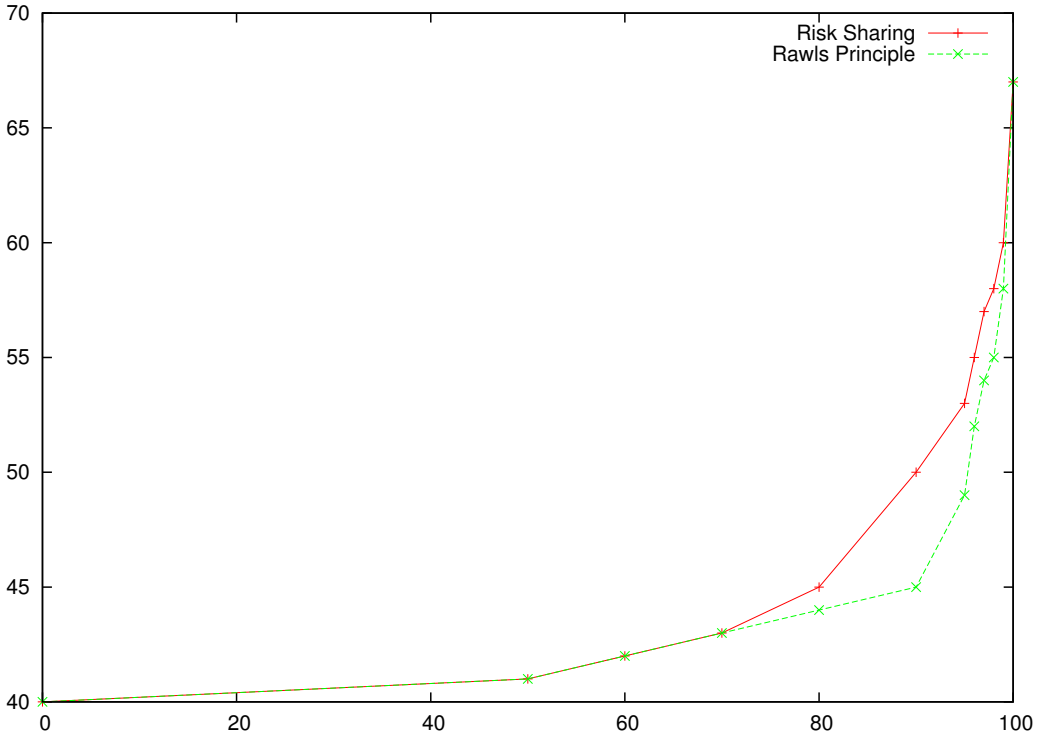



Figure 8.1: Comparison between Risk Sharing and Rawls' Principle

# Chapter 9

# Routing of hazardous materials

Nora Touati Moungla, Vincent Jost

*Published in Proceedings of Toulouse Global Optimization workshop 2010, p.153-156*

*The routing of vehicles represents an important component of many distribution and transportation systems and has been intensively studied in the operations research literature. In this report, particular consideration is given to routing models for the transportation of hazardous materials. This problem has received a large interest in recent years, this results from the increase in public awareness of the dangers of hazardous materials and the enormous amount of hazardous materials being transported. We describe here some major differences between routing of hazardous materials and the classical vehicle routing problems. We review some general models and optimization techniques and propose several direction for future research.*

## 9.1 Introduction

Hazardous Materials Management (HMM) has received a large interest in recent years, this results from the increase in public awareness of the dangers of hazardous materials and the enormous amounts of hazardous materials being transported. HMM is concerned with four main subjects: risk analysis [141], routing and scheduling [159], facility location [144, 159] and treatment and disposal of waste [292]. We focus here on Routing of Hazardous Materials (RHM). The main target of this problem is to select routes from a given origin 's' to a given destination 't' such that the risk for the surrounding population and the environment is minimum, without producing excessive economic costs. The risk associated with the hazardous materials makes these problems more complicated by its assessment, the collection of the related data and the resolution of the associated formulations.

In analyzing the routing of hazardous materials problem, it is important to include multiple objectives, this results from the presence of multiple interested parties or stakeholders. In this case, it is not possible to identify a single "best" route, generally "Pareto optimal" routes represent the available tradeoffs explicitly. Another important aspect of the transportation of hazardous materials is uncertainty, this is a result of the difficulty of risk measurement and the lack of data. We present in this report some characteristics of routing of hazardous materials problem and describe some models and the most used resolution methods for this problem.

## 9.2 Routing for hazardous materials

The difference between RHM and other routing problems is mainly the risk. We present in this section some methodologies in RHM problem consisting essentially of modeling and resolution frameworks.

## 9.3 The risk

Because of the risk, a network related to this problem is generally different from other transportation networks. In RHM, the arcs do not necessarily connect junctions, in the case where a road between two junctions goes through a set of different density population regions, this arc is divided into a set of arcs in such a way that these new arcs have the same incident probability and the same consequences.

Although the fact that the major target of the RHM problem is the minimization of the risk, there is no universally accepted definition of risk [143]. It is pointed in [159] that the evaluation of risk in transportation of hazardous materials generally consists of the evaluation of the probability of an undesirable event, the exposure level of the population and the environment, and the degree of the consequences (e.g., deaths, injured people, damages). In practice, these probabilities are difficult to obtain due to the lack of data and generally, the analysis is reduced to consider the risk as the expected damage or the population exposure.

As the risk is a part of the objective function, it is quantified with a path evaluation function [143]. Some risk models (the conditional risk model for example, see [332]) lead to non-linear binary integer formulations which can not optimized using a simple labeling algorithm (the associated models violate a path selection optimality principle). Generally, these models are based on approximations which lead to tractable formulations. Three axioms are introduced in [143] for prescribing properties for path evaluation functions. These axioms allow to check the monotonicity of the links attribute and to check if the path selection model verifies Bellman's optimality principle.

## 9.4 Models and optimization

RHM problem is multi-objective in nature, nevertheless, some papers deal with single-objective problems. These models often fail to handle the conflict between transportation risk and cost. A number of multi-objective models have been proposed in the literature. With multiple objectives, all objectives usually cannot be optimized simultaneously. Generally, a set of alternative (Pareto-optimal) solutions are computed. As the number of Pareto-optimal solutions can be exponential as a function of the network size, one might wish to compute a subset of these solutions "approximating well" the set of all Pareto-optimal solutions.

Resolution methods of hazardous materials routing can be classified in two categories. The "local routing" which consists in selecting routes between only one origin-destination pair and transporting a single commodity at a time. The "global routing" where different hazardous materials have to be shipped simultaneously among different origin-destination pairs.

### 9.4.1 Local routing of hazardous materials

The one origin-destination routing consists of selecting a route between a given origin-destination nodes for a given hazardous materials, transport mode and vehicle type.

Weighting methods are widely used due to their simplicity and computational efficiency. They are

based on optimizing a weighted linear combination of the objectives. This can be done using the classical shortest path algorithm. The drawback of these methods is that they are able to identify only a subset of Pareto-optimal solutions. Since some solutions of interest might be ignored, a method which can identify them is desirable. However, in some cases when the decision maker is able to express additional a priori knowledge and preferences on the objectives, the problem can be reduced to a single objective optimization problem.

The goal programming formulations offers considerable flexibility to the decision maker, the purpose is to minimize the deviation between the achievement of objectives (goals) and their aspiration level (the acceptable achievement level for the objective). This method is able to compute Pareto-optimal solutions that can not be obtained with the weighting method.

### 9.4.2 Global routing of hazardous materials

A non-negligible work in the literature focuses on the selection of a single commodity route between only one origin-destination pair. In practice, a more adapted model is global routing. When many vehicles have to be routed between the same origin-destination nodes, these vehicles are routed on the same path, so, the risk associated to regions surrounding this path could be hight. In this case, we sometimes wish to distribute the risk in an equitable way over the population.

#### Risk equity

Different techniques was proposed to handle equity on the transportation network. In [166], the authors guarantee equity by constraining the differences of the risks associated to every pair regions, to be less than or equal to a given threshold. The computation of routes with a fairly distributed risk consists in generating dissimilar origin-destination paths. The dissimilar path generation problem has been dealt in the literature in many ways, we cite the Iterative Penalty Method, the Gateway Shortest-Paths method, the Minimax Method and the p-dispersion Method.

#### Multi-commodity flow models

The transportation of hazardous materials can be naturally modeled by a multi-commodity flow model [92]. Given an origin-destination pair and the amount of commodities to be transported between such an origin and a destination, the multi-commodity minimum cost flow model finds the optimal distribution of such quantities minimizing the total transportation cost.

## 9.5 Uncertainty in routing of hazardous material

It is important to classify the nature of uncertainty in transportation of hazardous materials problems. In particular, three types of uncertainty concern the amount of population present near a route, traffic congestion and weather conditions. The effect of the release of hazardous materials and the travel time can be modeled by means of random variables whose distributions may vary over time [350] over a stochastic and time-dependent network [97]. Optimal routing on time-dependent networks can be classified into three categories [142]:

• *A priori optimization*: Optimal routes are definitively computed before the travel begins, the random arc travel time is reduced to its expected value and a standard shortest path problem is applied.

• *Adaptive route selection*: The optimal route depends on intermediate informations concerning past travel time, road congestion and weather conditions. The adaptive road specifies the road link to be chosen at each intermediate node, as a function of the arrival time at the node [293]. The multi-objective version of the adaptive route selection was proposed in [276].

• *Adaptive route selection with real-time updates*: In this case, recent technologies such as automatic vehicle location and mobile communication systems permits to guide the route of vehicles based on real-time informations. Estimation of future values of some network attributes such as travel times, incident probabilities and population in the impact area are updated using the real-time informations.

## 9.6 Conclusion

Transportation of hazardous materials is a complex and seemingly intractable problem, principally because of the inherent trade-offs between social, environmental, ecological, and economic factors. An model for the routing of hazardous materials problem not only needs to be accurate but also technically and computationally tractable. There is no common conceptual model for the RHM problem. Works in this field take generally into account different considerations (economic, environmental and social) and significant simplifications are necessary to obtain tractable models. Several challenging directions for future research can be stemmed from this problem:

• A common resolution approach for multi-objective shortest path problems consists of computing Pareto optimal solutions. As the number of Pareto optimal solutions can grow exponentially with the size of the network, one can propose to the decision maker a subset of Pareto optimal solutions representing a good approximation of the Pareto optimal solutions set or compute preferred solutions by exploiting some preferences of the decision maker.

• An important issue of the RHM is the treatment of the stochastic phenomena, indeed, some or all attributes of this problem are stochastic.

• Considerable advances are needed to appropriately treat the stochastic phenomena when some transformation of measures (cost and risk) are nonlinear.

• New technological advances in communication systems and Global Positioning System (GPS) are challenging researchers to develop routing models and robust optimization procedures that are able to respond quickly to changes in the data.

• Most studies on RHM in the literature deal with road transportation mode [142]. Although rail transport is a safer transportation mode (automatic control system, cross less populated zones), and more capacitated, it has received less attention.

# Chapter 10

# Evolutionary optimization for the multiobjective risk-equity constrained routing problem

Dimo Brockhoff and Nora Touati-Moungla

*Submitted to CTW International Workshop 2011 on Graphs and Combinatorial Optimization*

*This report proposes a methodology to find a routing plan for transporting hazardous materials that attempts to balance costs with safety and green requirements.*

## 10.1 Introduction

The transportation of hazardous materials (*hazmat* from now on) has received a large interest in recent years, which results from the increase in public awareness of the dangers of hazmats and the enormous amount of hazmats being transported [92]. The main target of this problem is to select routes from a given origin-destination pair of nodes such that the risk for the surrounding population and the environment is minimized—without producing excessive economic costs.

When solving such a problem by minimizing both cost and the total risk, typically several vehicles share the same (short) routes which results in high risks associated to regions surrounding these paths whereas other regions are not affected. In this case, one may wish to distribute the risk in an equitable way over the population and the environment. Several studies consider this additional minimization of the equity risk, but most of them consist of a *single* origin-destination hazmat routing for a specific hazmat, transport mode and vehicle type (see for example [6, 92]).

A more realistic *multi*-commodity flow model was proposed in [92] where each commodity is considered as one hazmat type. The objective function is formulated as the sum of the economical cost and the cost related to the consequences of an incident for each material. To deal with risk equity, the costs are defined as functions of the flow traversing the arcs which imposes an increase of the arc's cost and risk when the number of vehicles transporting a given material increases on the arc.

The majority of all hazmat routing studies deal with a single-objective scenario although the problem itself is multiobjective in nature and it is important to study the trade-offs among the objectives. Evolutionary Multiobjective Optimization (EMO) algorithms are able to compute a set of solutions showing

these trade-offs within a single algorithm run which is the reason why we propose to use them for the problem of hazmat routing in this study (Sec. 10.3). Before, we formalize the routing of hazmat problem with three objectives (minimize total routing cost, total routing risk *and* risk equity) as a multicommodity flow model as in [92] since this model is the most realistic one permitting to manage several hazmat types simultaneously (Sec. 10.2).

## 10.2 The multiobjective risk-equity constrained routing problem

Let the transportation network be represented as a directed graph $G = (N, A)$, with $N$ being the set of nodes and $A$ the set of arcs. Let $C$ be the set of commodities, given as a set of point-to-point demands to transport a certain amount of hazmats. For any commodity $c \in C$, let $s^c$ and $t^c$ be the source node and the destination node respectively, and let $V^c$ be the number of available trucks for the transportation of commodity $c$. Each commodity is associated with a unique type of hazmat. We assume that the risk is computed on each arc of the network and is proportional to the number of trucks traversing such an arc. We consider a set $Q$ of regions, and we define $r_{ij}^{cq}$ as the risk imposed on region $q \in Q$ when the arc $(i, j) \in A$ is used by a truck for the transportation of hazmat of type $c$. We remark that we employ a notion of *spread risk*, in that an accidental event on arc $(i, j)$ within region $q \in Q$ may strongly affect another region $q' \in Q$. With each arc $(i, j) \in A$ a cost $c_{ij}^c$ is associated, involved by the travel of a truck of commodity $c$ on this arc.

The problem of transporting hazmat is multiobjective in nature: one usually wants to minimize the *total cost* of transportation, the *total risk* of transportation imposed on all regions and the *distributed risk*, which can be defined as a measure of risk that is shared among different regions. More specifically, for a given solution, each region $q \in Q$ will be affected by a risk $\omega_q$ which depends on the transportation patterns in all other regions. The third objective will then be $\max_{q \in Q} \omega_q$, and has to be minimized.

We introduce the integer variable $y_{ij}^c$ for the number of trucks that use arc $(i, j)$ for transporting commodity $c$. We assume a fixed number of trucks and that all trucks have the same load. The proposed model is defined as follows:

$$\min \quad \sum_{c \in C} \sum_{(i,j) \in A} c_{ij}^c y_{ij}^c \tag{10.1}$$

$$\min \quad \sum_{c \in C} \sum_{q \in Q} \sum_{(i,j) \in A} r_{ij}^{cq} y_{ij}^c \tag{10.2}$$

$$\min \quad \max_{q \in Q} \left\{ \sum_{c \in C} \sum_{(i,j) \in A} r_{ij}^{cq} y_{ij}^c \right\} \tag{10.3}$$

$$s.t. \quad \sum_{j \in \delta^+(i)} y_{ij}^c - \sum_{j \in \delta^-(i)} y_{ji}^c = q_i^c \qquad \forall i \in N, c \in C \tag{10.4}$$

$$y_{ij}^c \in \{0, 1, 2, \ldots\} \qquad \forall (i, j) \in A, c \in C. \tag{10.5}$$

The first objective in (10.1) is a cost function and is to be minimized. The second objective, given by (10.2), minimizes the total risk on all regions and objective (10.3) minimizes the maximum risk imposed on all regions. Constraints (10.4) are conservation constraints, where $\delta^-(i) = \{j \in N : (j, i) \in A\}$ and $\delta^+(i) = \{j \in N : (i, j) \in A\}$ are the direct successors and predecessors of node $i$, and $q_i^c = V^c$ if $i = s^c$, $q_i^c = -V^c$ if $i = t^c$ and $q_i^c = 0$ otherwise.

## 10.3 Evolutionary multiobjective optimization

Evolutionary algorithms (EAs) and Evolutionary Multiobjective Optimization (EMO) algorithms in particular are general-purpose randomized search heuristics and as such well suited for problems where the objective function(s) can be highly non-linear, noisy, or even given only implicitly, e.g., by expensive simulations [115, 103]. Since the third objective in the above problem formulation is nonlinear, we

propose to use an EMO algorithm for the multiobjective risk-equity constrained routing problem here. Our EMO algorithm follows the standard iterative/generational cycle of an EA of mating selection, variation, objective function evaluation, and environmental selection and is build upon the state-of-the-art selection scheme in HypE [44] as implemented in the PISA platform [67]. The variation operators as well as the representation of the solutions, however, have to be adapted to the problem at hand in the following way in order to fulfill the problem's constraints at all times.

### 10.3.1 Representation

We choose a variable length representation as it has been theoretically shown to be a good choice for multiobjective shortest paths problems [191]: A solution is thereby represented by a list of paths of variable lengths with one path per truck. For the moment, we consider a fixed amount of trucks for each commodity and therefore a fixed number of paths through the network. In order to have every variable length path represent an uninterrupted path from source to destination at any time (see the constraints in (11.3)), we ensure all paths to always start with the source $s^c$ for the corresponding commodity $c$, ensure with the variation operator that all neighbored vertices in the path are connected by an arc, and complete each path by the shortest path between the path's actual end node and the commodity's destination node $t^c$.

### 10.3.2 Initialization

Initially, we start with a single solution where the paths $p$ for all trucks are empty ($p = (s^c)$). This corresponds to the situation where all trucks choose the shortest $s^c$–$t^c$ path for their assigned commodity—implying the smallest possible overall cost but a high risk along the used route(s). Nevertheless, the initial solution is already Pareto-optimal and is expected to be a good starting point for the algorithm.

### 10.3.3 Variation

As mutation operator, we suggest to shorten or lengthen the path of one or several trucks. In order to generate a new solution $s'$ from $s$, for each truck path, we draw a binary value $b \in \{0, 1\}$ uniformly at random and create the new path $p'$ from the old one $p = (v_1 = s^c, v_2, \ldots, v_l)$ as in [191]:

- if $b = 0$ and $l =: \text{length}(p) \geq 2$, set $p' = (s^c, \ldots, v_{l-1})$
- if $b = 1$ and $|V_{\text{rem}} = \{v \in V \,|\, (v_l, v) \in A\}| \neq \emptyset$, choose $v_{l+1}$ from $V_{\text{rem}}$ uniformly at random and set $p' = (v_0, \ldots, v_l, v_{l+1})$.
- otherwise, use the same path $p$ also in the new solution $s'$.

## 10.4 Conclusions

The transportation of hazmats is an important optimization problem in the field of sustainable development and in particular the equitable distribution of risks is of high interest. Within this study, we formalize this transportation problem as the minimization of three objectives and propose to use an evolutionary algorithm to cope with the non-linear equity risk objective.

The third objective function of our problem can be rewritten by minimizing the additional variable $z$ as third objective and adding the constraints $\forall q \in Q : z \geq \sum_{c \in C} \sum_{(i,j) \in A} r_{ij}^{cq} f_{ij}^c$. Although this equivalent formulation makes the problem linear (with additional linear constraints), classical algorithms are

expected to have difficulties with this formulation as well and our algorithm is supposed to be more efficient in the current formulation due to the fewer number of constraints. Note that, for the moment, the proposed EMO algorithm exists on paper only and an actual implementation has to prove in the future which additional algorithm components (such as problem-specific initialization, recombination operators, or other exact optimization (sub-)procedures) are necessary to generate solutions of sufficient quality and whether adaptively changing the number and capacity of trucks is beneficial.

# Chapter 11

# A branch-and-price algorithm for the risk-equity constrained routing problem

Pietro Belotti, Vincent Jost, Leo Liberti and Nora Touati-Moungla

*Accepted in the International Network Optimization Conference 2011 (LNCS Proceedings)*

*We study a multi-criteria variant of the problem of routing hazardous material on a geographical area subdivided in regions. The two objective functions are given by a generally defined routing cost and a* risk equity *equal to the maximum, over each region, of the risk* perceived *within a region. This is a multicommodity flow problem where integer variables are used to define the number of trucks used for the routing. This problem admits a straightforward path formulation, and we propose a branch-and-price problem where, for each node of the branch-and-bound tree, column generation is used to obtain a lower bound. Experimental results on a set of instances are reported.*

## 11.1 Introduction

The transportation of hazardous materials (*hazmat* from now on) has received a large interest in recent years, this results from the increase in public awareness of the dangers of hazmats and the enormous amount of hazmats being transported. The main target of this problem is to select routes from a given origin-destination pair of nodes such that the risk for the surrounding population and the environment minimum, without producing excessive economic costs. The risk associated to the hazmats makes this problems more complicated by it assessment [143], the related data collection and the resolution of the induced formulations [159, 348].

A non-negligible work in the literature focuses on local routing, where a single commodity routes between one origin-destination pair have to be selected. In practice, a more adapted model is global routing where many vehicles have to be routed between the same origin-destination nodes (see [92, 142] for a survey on these models). In global routing, vehicles are routed on the same path, so the risk associated to regions surrounding this path could be high; In this case, one may wish distribute the risk in an equitable way over the population. Routing of hazmat shipments can be defined as the problem of finding minimum cost/risk routes while spreading equitably the risk in any zone of the network area, this problem is a "many-to-many" routing problem with multiple origins and destinations. In the sequel, we

refer to this problem as Risk-Equity Constrained Routing Problem (RECRP). The research in the latter include two main classes:

1. Methods based on the generation of dissimilar origin-destination paths: In [7, 227, 263], the authors used different dissimilarity indexes between paths by considering common parts of the paths. The main principle of these methods consists in two phases: (1) the computation of an initial set of routes $R$, (2) the selection of a subset of dissimilar paths $D \subset R$. A drawback of these methods is that aiming to compute dissimilar paths by considering common parts of the paths can be useless. In fact, it can happen that two paths, without any part in common, are parallel but very close, these paths would be considered by this model as having maximum dissimilarity.

2. Methods based on the uniformly risk distribution among all the zones of the geographical crossed region: In [167] the authors propose an integer programming formulation for the RECRP, they handle equity by imposing that the difference between the risks imposed on any two arbitrary zones does not exceed a given threshold $\epsilon$. The problem was solved using heuristic that repeatedly solves single-trip problems: a Lagrangian dual approach with a gap-closing procedure is used to optimally solve single-trip problems. The major drawback of using this model is that the paths computed can impact the same zones when $\epsilon$ is large and the problem can be infeasible when $\epsilon$ is small.

   In [92] was proposed a multi-commodity flow model for the RECRP where each commodity can be considered as one hazmat type. The capacity associated to each arc for a given commodity can be considered as the maximum amount of risk tolerated along the considered arc resulting from the transportation of a unit of hazmat type and the capacity associated to each arc for the global commodities can be considered as the total maximum risk tolerating on the given arc. The objective function is formulated as the sum of the economical cost and the cost related to the consequences of an incident for each material $k$. This model is well adapted to manage simultaneously different materials that have to be routed over the network, but with this model one can compute a spatially dissimilar path set, and, therefore, does not guarantee the equity target. To deal with risk equity [92], the costs were defined as functions of the flow traversing the arcs, this imposes an increase of the arc's cost and risk when the number of vehicles transporting a material $k$ increases on the arc, a linear version of this model was proposed. This model can allow to prevent arc saturation, by discouraging the use of a few arcs to transport a commodity $k$ between its origin and destination nodes.

Our problem is similar to that proposed in [92]. We consider the problem where a set of given quantities of hazmats has to be routed over a transportation network from specific origin points to specific destination points. Our goal is the minimization of the total routing cost *and* of the risk equity, the latter broadly defined as the risk shared by a set of regions that compose the geographical area under consideration. Thus our focus is a multi-criteria optimization problem which we describe more in detail below. The originality of our work is the integration of the objective of minimization of the maximum of risk imposed on all regions during the transportation activity into the multi-commodity flow model which is solved using a Branch-and-Price algorithm.

### 11.1.1 Description of the problem

Let the transportation network be represented as a directed graph $G = (N, A)$, with $N$ being the set of $n$ nodes and $A$ the set of $m$ arcs. Let $C$ be the set of commodities, given as a set of point-to-point demands to transport a certain amount of hazmats. For any commodity $c \in C$, let $s^c$ and $t^c$ be respectively the source node and the destination node, and let $D^c$ be the amount of hazmats to be shipped, by means of a set of trucks of given capacity $F_c$, from $s^c$ to $t^c$. We for now assume that each commodity is associated with a unique type of hazmat.

We assume that the risk is computed on each arc of the network and is proportional to the flow traversing such an arc. We consider a set $Q$ of regions, each given as subsets $N_q$ of nodes for each $q \in Q$ of the transportation network, and we define $r_{ij}^{cq}$ as the risk imposed on region $q \in Q$ when the arc $(i, j) \in A$ is used for the transportation of one unit of hazmats of type $c$. We remark that we employ a notion of *spread risk*, in that an accidental event on arc $(i, j)$ within region $q \in Q$ may strongly affect another region $q' \in Q$.

### 11.1.2 Multiple objective functions

The problem of transporting hazmat is multi-objective in nature: one usually wants to minimize two (or more) objectives, namely the total cost of transportation, computed as a function of the amount of hazmat transported throughout the network and the trucks used for the transportation, and the *distributed risk*, which can be defined as a measure of risk that is shared among different regions. More specifically, for a given solution each region $q \in Q$ will be affected by a risk which is dependent on the transportation patterns in all other regions, and which can be summarized by a quantity $\omega_q$. The second objective will then be $\max_{q \in Q} \omega_q$, and has to be minimized.

### 11.1.3 An optimization model

We introduce a flow variable $f_{ij}^c$ defining the portion of commodity $c$ being transported on arc $(i, j)$. These variables are subject to flow conservation constraints

$$\sum_{j \in \delta^+(i)} f_{ij}^c - \sum_{j \in \delta^-(i)} f_{ji}^c = b_i^c \qquad \forall i \in N, c \in C$$

where $\delta^-(i)$ and $\delta^+(i)$ are the forward and backward star of $i$, i.e.,

$$\delta^-(i) = \{j \in N : (j, i) \in A\}, \qquad \delta^+(i) = \{j \in N : (i, j) \in A\},$$

and

$$b_i^c = \begin{cases} 1 & \text{if } i = s^c \\ -1 & \text{if } i = t^c \\ 0 & \text{otherwise.} \end{cases}$$

Also, $y_{ij}^c$ defines the number of trucks to be used on arc $(i, j)$ for commodity $c$. The link between variables $f$ and $y$ is given by the constraint

$$D_c f_{ij}^c \leq F_c y_{ij}^c \qquad \forall (i, j) \in A, c \in C.$$

The first objective is a function of both $f$ and $y$ variables and is to be minimized: $\sum_{c \in C} \sum_{(i,j) \in A} (\alpha_{ij}^c f_{ij}^c + \beta_{ij}^c y_{ij}^c)$, with $\alpha$ and $\beta$ suitable cost coefficients which we assume nonnegative. We define the risk $\omega_q$ imposed on a region $q \in Q$ as a linear combination of the flow variables:

$$\omega_q := \sum_{c \in C} \sum_{(i,j) \in A} r_{ij}^{cq} f_{ij}^c$$

and add a new variable $z := \max_{q \in Q} \omega_q$, which therefore is subject to the constraints

$$z \geq \sum_{c \in C} \sum_{(i,j) \in A} r_{ij}^{cq} f_{ij}^c \qquad \forall q \in Q.$$

The $y$ variables represent trucks that transport hazmat from each source to each destination, and are therefore subject to flow conservation constraints. We write such constraints here for each commodity

and for all intermediate nodes of each commodity, as the source and destination flow balance is redundant here (i.e., it is strictly dependent on the flow variables $f$):

$$\sum_{j\in\delta^+(i)} y^c_{ij} - \sum_{j\in\delta^-(i)} y^c_{ji} = 0 \qquad \forall i \in N \setminus \{s^c, t^c\}, \forall c \in C$$

The optimization model is therefore as follows:

$$\begin{aligned}
\min \quad & \textstyle\sum_{c\in C}\sum_{(i,j)\in A}(\alpha^c_{ij} f^c_{ij} + \beta^c_{ij} y^c_{ij}) && && (11.1)\\
\min \quad & z && && (11.2)\\
s.t. \quad & \textstyle\sum_{j\in\delta^+(i)} f^c_{ij} - \sum_{j\in\delta^-(i)} f^c_{ji} = b^c_i, && \forall i \in N, c \in C && (11.3)\\
& \textstyle\sum_{j\in\delta^+(i)} y^c_{ij} - \sum_{j\in\delta^-(i)} y^c_{ji} = 0 && \forall i \in N \setminus \{s^c, t^c\}, \forall c \in C && (11.4)\\
& D_c f^c_{ij} \le F_c y^c_{ij} && \forall (i,j) \in A, c \in C && (11.5)\\
& z \ge \textstyle\sum_{c\in C}\sum_{(i,j)\in A} r^{cq}_{ij} f^c_{ij} && \forall q \in Q && (11.6)\\
& f^c_{ij} \in [0,1] && \forall (i,j) \in A, c \in C && (11.7)\\
& y^c_{ij} \in \mathbb{Z} && \forall (i,j) \in A, c \in C. && (11.8)
\end{aligned}$$

Notice that constraints (11.4) and (11.5) guarantee that a sufficient number of trucks is allocated for each commodity regardless of the flow of hazmat. The path formulation described below is unable to provide such a guarantee and will therefore have to be modified.

### 11.1.4 A path formulation

The above *arc-flow* formulation is polynomial in $|N|$, $|A|$, $|Q|$, and $|C|$, but its size can make it impractical to solve real-world instances of our problem. A common approach is to use a *path-flow* formulation [57]. In these formulations, for each commodity $c$ a variable is associated with every *path* from $s^c$ to $t^c$. We denote by $\mathcal{P}^c$ the set of paths from $s^c$ to $t^c$ for a commodity $c \in C$ and by $\mathcal{P}^c_{ij}$ the set of paths in $\mathcal{P}^c$ containing arc $(i,j) \in A$. A new path variable $f_p$, $\forall p \in \mathcal{P}^c, \forall c \in C$, represents the portion of commodity transported on path $p$. We drop the upper bound on $f_p$ variables as it is redundant, i.e., it is easy to prove that it is satisfied by any optimal solution.

As for the flow of hazmat, in this formulation the number of trucks, previously denoted by variables $y^c_{ij}$, might be dependent on path variables $f_p$. They are by definition the number of trucks to be used on arc $(i,j) \in A$ for commodity $c \in C$. Unless (but this does not seem viable) in practice every truck is required to drive on a single arc (i.e. there is a transfer of hazmat from a truck to another at every node of $G$), intuition suggests that each of them drives on the whole path $p$, hence there should be a variable $y_p$ that counts the number of trucks is simply related to variable $f_p$ as follows:

$$F_c y_p \ge D_c f_p \qquad \forall p \in \mathcal{P}^c, c \in C. \tag{11.9}$$

This constraint substitutes the flow conservation constraint (11.4) and the "capacity" constraint (11.5). Let us write this path formulation for completeness:

$$\begin{aligned}
\min \quad & \textstyle\sum_{c\in C}(\sum_{p\in\mathcal{P}^c} \alpha_p f_p + \sum_{(i,j)\in A} \beta_{ij} y^c_{ij}) && && (11.10)\\
\min \quad & z && && (11.11)\\
s.t. \quad & \textstyle\sum_{p\in\mathcal{P}^c} f_p \ge 1 && \forall c \in C && (11.12)\\
& F_c y_p - D_c f_p \ge 0 && \forall p \in \mathcal{P}^c, c \in C && (11.13)\\
& z - \textstyle\sum_{c\in C}\sum_{p\in\mathcal{P}^c} r^q_p f_p \ge 0 && \forall q \in Q && (11.14)\\
& f_p \ge 0 && \forall p \in \mathcal{P}^c, c \in C && (11.15)\\
& y^c_p \in \mathbb{Z} && \forall p \in \mathcal{P}^c, c \in C, && (11.16)
\end{aligned}$$

where $\alpha_p = \sum_{(i,j)\in p} \alpha_{ij}^c$ are cost coefficients on the path $p \in \mathcal{P}^c$ and $r_p^q = \sum_{(i,j)\in p} r_{ij}^{cq}$ is the risk imposed on region $q \in Q$ when the path $p \in \mathcal{P}^c$ is used for the transportation of one unit of hazmats of type $c$. Constraint (11.12) is the path-flow counterpart of the flow conservation constraint (11.3) and requires that, regardless of the set of paths used, each commodity is fully routed. Constraints (11.13) and (11.14) are straightforward extensions of (11.5) and (11.6) respectively, given that the flow of commodity $c \in C$ on arc $(i, j) \in A$ is equal to $\sum_{p\in\mathcal{P}_{ij}^c} f_p$.

When restricting to a single-objective optimization problem, this model is an integer multicommodity flow problem. Regardless of considering only one objective, problem (11.11)-(11.16) contains $V = \sum_{c\in C} |\mathcal{P}^c|$ variables, which can be exponential in $|N|$. Therefore, solving it by introducing all path variables is in general impractical using the usual combinatorial optimization methods.

Column generation algorithms are very well suited for solving this kind of problems [123]. They use a relatively small initial set of columns to solve a problem, and iteratively introduce a new column when necessary to improve the objective function. Specifically, given a set of columns with negative reduced cost (among those that haven't been considered yet), one can introduce one or more such variables and apply a primal simplex method to resolve the amended problem. The problem with an initially small subset of columns is called the *restricted master problem*, while the problem of finding a variable (column) with negative reduced cost is called the *pricing problem*.

Constraint (11.13) introduces a major issue in the problem. In principle, introducing $y$ variables indexed on paths rather than arcs and commodities allows to further reduce the number of columns, as we only need $1 + 2\sum_{c\in C} |\mathcal{P}^c|$ variables. However, analogously to columns, we do not want to have exponentially many rows (there are exponentially many paths). The above constraint could be *dynamically* generated, hence instead of column generation we would need *row-column* generation. One huge problem here is that to dynamically generate paths one needs to know all dual variables $\sigma_p$, for each $p \in \mathcal{P}^c$ and for all $c \in C$, to solve a pricing problem, and most of these dual variables are *not* available since we didn't generate all of them.

One possible way to deal with this is to use *surrogate constraints*: rather than impose all such constraints or generate them dynamically, we consider a *cover* of such set of constraints and impose conic combinations thereof (see for instance [297]). More specifically, for each $(i, j) \in A$, consider all constraints (11.9) summed up for all paths containing $(i, j)$. We obtain

$$F_c \sum_{p\in\mathcal{P}_{ij}^c} y_p \geq D_c \sum_{p\in\mathcal{P}_{ij}^c} f_p \qquad \forall (i, j) \in A, c \in C. \tag{11.17}$$

The problem has now $|C|(1 + m) + |Q|$ rows and $1 + 2\sum_{c\in C} |\mathcal{P}^c|$ variables. Since we relax all of the path constraints (11.9), the model (11.10)-(11.16) constitutes a relaxation of (11.1)-(11.8). Column generation can be applied safely now, although it has to be applied to both $f$ and $y$ variables, and it converges to a *dual feasible* solution which gives a lower bound but not necessarily an optimal solution of the continuous relaxation of (11.1)-(11.8).

Suppose an integer solution is found as the optimal solution of the LP relaxation (solved with column generation). If at least one of the constraints (11.13) is violated, we are stuck with a solution that has no physical value but that cannot be proven primal infeasible unless a constraint is added. What we can do is therefore to create a second branching rule which discriminates between integer feasible solutions and eliminates the integer (but infeasible) solution just found. We will detail this procedure later in this report, and instead provide insight on how to generate variables.

### 11.1.5 Column generation applied to one objective only.

We consider from now on a continuous relaxation of (11.11)-(11.16) amended by the surrogate constraints:

$$\begin{aligned}
\min \quad & z & & (11.18)\\
s.t. \quad & \textstyle\sum_{p\in\mathcal{P}^c} f_p \geq 1 & \forall c \in C & (11.19)\\
& F_c \textstyle\sum_{p\in\mathcal{P}^c_{ij}} y_p \geq D_c \sum_{p\in\mathcal{P}^c_{ij}} f_p & \forall (i,j)\in A, c\in C & (11.20)\\
& z - \textstyle\sum_{c\in C}\sum_{p\in\mathcal{P}^c} r^q_p f_p \geq 0 & \forall q \in Q & (11.21)\\
& f_p \geq 0 & \forall p\in\mathcal{P}^c, c\in C & (11.22)
\end{aligned}$$

We associate the dual variables vector $\mu \in \mathbb{R}^{|C|}_+$ with constraints (11.12), $\sigma \in \mathbb{R}^{m|C|}_+$ with constraints (11.20), and $\lambda \in \mathbb{R}^{|Q|}_+$ with constraints (11.14). We first analyze this problem considering the single objective (11.11). Let us define the subset of paths $\bar{\mathcal{P}}^c \subset \mathcal{P}^c, \forall c \in C$. The restricted master problem (RMP from now on) of (11.11)-(11.16), generated on a restricted subset of variables $f_p, p \in \bar{\mathcal{P}}^c, c \in C$, is as follows:

$$\begin{aligned}
\min \quad & z & & (11.23)\\
s.t. \quad & \textstyle\sum_{p\in\bar{\mathcal{P}}^c} f_p \geq 1 & \forall c \in C & (11.24)\\
& F_c \textstyle\sum_{p\in\mathcal{P}^c_{ij}} y_p - D_c \sum_{p\in\bar{\mathcal{P}}^c_{ij}} f_p \geq 0 & \forall c\in C, (i,j)\in A & (11.25)\\
& z - \textstyle\sum_{c\in C}\sum_{p\in\bar{\mathcal{P}}^c} r^q_p f_p \geq 0 & \forall q \in Q & (11.26)\\
& f_p \geq 0 & \forall p\in\mathcal{P}^c, c\in C. & (11.27)
\end{aligned}$$

It is barely worth noting here that (11.23)-(11.27) is a restriction of the continuous relaxation of (11.11)-(11.16), which therefore provides neither a lower nor an upper bound. Only by applying column generation to (11.23)-(11.27), i.e., by iteratively amending columns with negative reduced cost, can we find a lower bound of (11.11)-(11.16). The reduced cost of variables $f_p$, for each $c \in C, p \in \mathcal{P}^c$, is as follows:

$$\begin{aligned}
h(f_p) = \quad & -\mu_c + D_c \textstyle\sum_{(i,j)\in p} \sigma^c_{ij} + \sum_{q\in Q} r^q_p \lambda_q\\
= \quad & -\mu_c + D_c \textstyle\sum_{(i,j)\in p} \sigma^c_{ij} + \sum_{q\in Q}\sum_{(i,j)\in p} r^{cq}_{ij} \lambda_q.
\end{aligned} \tag{11.28}$$

Suppose an optimal primal solution $(\bar{f}, \bar{y}, \bar{z})$ and an optimal dual solution $(\bar{\mu}, \bar{\sigma}, \bar{\lambda})$ is given. At each iteration of the column generation algorithm, we look for a negative reduced cost variable by solving the problem:

$$\min_{c\in C, p\in\mathcal{P}^c} h(f_p),$$

which provides the column with most negative reduced cost. The *pricing problem* consists of finding the path $p$ that minimizes (11.28), and is equivalent to solving a *shortest path* problem on a graph $G$ where each arc $(i,j) \in A$ has weight $w_{ij} = D_c\bar{\sigma}^c_{ij} + \sum_{q\in Q} r^{cq}_{ij}\bar{\lambda}_q$. The path must have an origin-destination pair among those defined by the commodities in $C$. Suppose that, for the shortest path obtained, $-\bar{\mu}_c + D_c L_c + \sum_{q\in Q} r^q_p \bar{\lambda}_q < 0$. Then variable $f_p$ has a negative reduced cost and can be introduced in the model.

One may also look for a negative reduced cost variable for *each* commodity, and add at most $|C|$ such variables. Although this usually speeds up convergence in terms of number of iterations, adding many column every time slows the primal simplex used to obtain a new solution. We obtain $|C|$ origin-destination shortest path problems, therefore the pricing problem becomes $|C|$ times slower — this is negligible given that most of the CPU time is usually spent on the primal simplex.

Notice that $y$ variables do not need to be generated for the risk-objective problem: they only appear in the surrogate constraint, which makes them completely useless given that their value can be decided from an optimal $f$. This only happens if we consider the second objective function, while the first does contain those variables and would force us to generate them as well. Actually, no $f$ variable is needed either as long as the $y$ variables are only contained in the capacity constraint. The next subsection should shed light on this and introduce another use for $y$ variables.

### 11.1.6 Risk on trucks

Another consideration is on risk equity associated to trucks: is the risk (especially the perceived one) only related to the real quantity, or portion, of hazmat transported, or is it also related on the trucks? If both quantity of hazmat and number of trucks should be considered, then the risk equity constraint would change. In this case, we could probably use a parameter $s_{ij}^{cq}$ with an analogous meaning to that of parameter $r$, i.e., the influence of one truck driving through $(i, j)$, transporting commodity $c \in C$, on region $q$, and modify (11.26) as follows:

$$z \geq \sum_{c \in C} \sum_{p \in \bar{\mathcal{P}}^c} (r_p^q f_p + s_p^q y_p) \qquad \forall q \in Q.$$

where, similarly to $r$, we define $s_p^q := \sum_{(i,j) \in p} s_{ij}^{cq}$. This provides a motivation for the generation of both $f$ and $y$ variables. In fact, now the procedure to generate $y$ variable can be defined as one that aims at finding a path $p$ such that the reduced cost of the corresponding $y_p$ is minimum:

$$\min_{c \in C, p \in \mathcal{P}^c} h(y_p) = \min\{-D_c \sum_{(i,j) \in p} \sigma_{ij}^c + \sum_{q \in Q} s_p^q \lambda_q : p \in \mathcal{P}\}$$

which provides a more difficult problem given that now the shortest path has to be found on a network with possibly both positive and negative weights.

## 11.2 Branch-and-price for single objective problems

In order to find an optimal integer solution to problem (11.10)-(11.16), the column generation approach outlined above must be coupled with a branch-and-bound algorithm. This class of algorithms, better known as *branch-and-price*, solve each branch-and-bound node by applying column generation on each lower bounding (continuous) subproblem [50]. For the single objective routing problem, we outline below an implementation of a branch-and-price, which we have implemented in AMPL.

### 11.2.1 Branching rules

If only integer variables $y_{ij}^c$ are not dynamically generated (but this no longer seems to be the case), the branching rule is rather simple: consider an optimal solution $(\bar{f}, \bar{y}, \bar{z})$ obtained after column generation at a branch-and-bound node. If, for all $c \in C$ and $(i, j) \in A$, we have $\bar{y}_{ij}^c \in \mathbb{Z}$, then the node can be fathomed as the solution is integer feasible. Otherwise, we select an arc $(i, j) \in A$ and a commodity $c \in C$ such that $\bar{y}_{ij}^c \notin \mathbb{Z}$ and generate two new branch-and-bound nodes with the amended constraints $y_{ij}^c \leq \left\lfloor \bar{y}_{ij}^c \right\rfloor$ and $y_{ij}^c \geq \left\lceil \bar{y}_{ij}^c \right\rceil$, respectively.

If we use $y_p$ variables instead, we need to take special care in branching rules: given that these variables are generated, the branching rules have dual variables that need to be taken into account in the pricing problem. Furthermore, simple branching rules would not work and the branch-and-bound algorithm would not converge: the branching rule $y_p \leq k$, with $k \in \mathbb{Z}$, does not impose anything on the pricing problem, which might generate another variable that uses the same path as $p$ with reduced cost. Another issue is making sure that the pricing problem remains a shortest path problem. One common branching rule for these cases is that used by Barnhart et al. [50]:

I am currently working on another class of branching rules, which we didn't test yet [52]. The report is still in draft version, but if we see that Barnhart's branching rule doesn't work we can try them.

### 11.2.2 Accelerating techniques

A common drawback of column generation techniques is the so-called *tailing off*, i.e., the tendency of the objective function of the RMP to decrease very slowly although at least one column is generated at each iteration. Stabilization or acceleration techniques for column generation can be used to this purpose [131, 351]. The simplest of them requires perturbing the RMP in such a way that large changes of the dual variables within an iteration are penalized.

### 11.2.3 Computational experiments

Ideas to be implemented and tested independently:

- implement the path formulation and the column generation in AMPL (I have some examples of it and a shortest path module)
- accelerate using duMerle's method
- implement a simple branch-and-bound algorithm in AMPL (not easy...)
- mix the BB and the accelerated column generation

## 11.3 Column generation in multi-criteria problems

To do:

- Search the literature for previous work. From a brief search not much comes up, which is not so surprising given that column generation is a method to test optimality on a single objective, while trying to characterize or generate all Pareto-optimal solution is orthogonal to this.
- Given a multi-objective problem $\min c^\top x, \min d^\top x : Ax \geq b$, consider the two parametric single objective problems

$$\begin{array}{lll} z_\alpha = \min & c^\top x & \\ & d^\top x \leq z_\beta & \\ & Ax \geq b & \end{array} \qquad \begin{array}{lll} z_\beta = \min & d^\top x & \\ & c^\top x \leq z_\alpha & \\ & Ax \geq b & \end{array}$$

  and consider the following (trivial?) question: when the column generation does not improve either linear relaxations, i.e., neither $z_\alpha$ nor $z_\beta$ improve, do we have a Pareto-optimal solution? This approach is analogous to a well-known method by Konno [223] for bilinear programming.
- Apply Lagrangian relaxation to either problem and see what the Lagrangian duals look like.

# Part III

# Innovative Methods for Optimization

# Chapter 12

# Multi-Objective Differential Evolution with Adaptive Control of Parameters and Operators

Ke Li, Álvaro Fialho, Sam Kwong

*Manuscript in preparation*

*Differential Evolution (DE) is a simple yet powerful evolutionary algorithm, whose performance highly depends on the setting of some parameters. In this report, we propose an adaptive DE algorithm for multi-objective optimization problems. Firstly, a novel tree neighborhood density estimator is proposed to enforce a higher spread between the non-dominated solutions, while the Pareto dominance strength is used to promote a higher convergence to the Pareto front. These two metrics are then used by an original replacement mechanism based on a three-step comparison procedure; and also to port two existing adaptive mechanisms to the multi-objective domain, one being used for the autonomous selection of the operators, and the other for the adaptive control of DE parameters CR and F. Experimental results confirm the superior performance of the proposed algorithm, referred to as Adap-MODE, when compared to two state-of-the-art baseline approaches, and to its static and partially-adaptive variants.*

## 12.1 Introduction

Differential Evolution, proposed by Storn and Price [305], is a popular and efficient population-based, direct heuristic for solving global optimization problems in continuous search spaces. The main benefits brought by DE are its simple structure, ease of use, fast convergence speed and robustness, which enables it to be widely applied to many real-world applications. For the generation of new solutions (trial vectors), each individual (target vector) is combined with others by means of different forms of weighted sums (mutation strategies). Originally, in case the newly generated solution has a better fitness value than its corresponding parent, it replaces its parent in the population for the next generation. The aim of these iterations is basically to find a proper direction for the search process towards the optimum, by following the quality distribution of the solutions in the current population.

One of the possible application domains of DE are the Multi-objective Optimization Problems (MOPs), which exist everywhere in real-world applications, such as engineering, financial, and scientific computing. The main difficulty in these cases lies in providing a way to compare the different solutions,

as the involved multiple criteria might compete with one another, besides possibly not being directly comparable. Multi-Objective Evolutionary Algorithms (MOEAs) tackle this issue by searching for the set of optimal trade-off solutions, the so-called Pareto optimal set: the aim is not only to approach the Pareto optimal front as closely as possible, but also to find solutions that are distributed over the Pareto optimal front as uniformly as possible, in order to better satisfy all the different objectives considered. Needless to say, to be applied to MOPs, the DE original scheme needs to be adapted according to the mentioned aims.

Many different types of DE variants proposed to tackle MOPs can be found in the literature, such as GDE3 [230], and DEMO [317]. We refer the reader to [112] for a recent comprehensive survey of DE, including its application to MOPs. But the performance of DE largely depends on the definition of some parameters. Besides the crossover rate CR, and the mutation scaling factor F, there is the need of choosing which mutation strategies, from the many available ones, should be used for the generation of new solutions, and at which rate each of the chosen strategies should be applied. The setting of these parameters is usually a crucial and very time-consuming task: the optimal values for them do not only depend on the problem at hand, but also on the region of the search space that is being explored by the current population, while solving the problem. Following the intuition of the Exploration versus Exploitation (EvE) balance, exploration tends to be more beneficial in the early stages of the search (consequently more exploratory mutation strategies, high values for F and CR), while more exploitation should be promoted when getting closer to the optimum (respectively, more fine-tuning operators, and a smaller value for F).

A prominent paradigm to automate the setting of these parameters on-line, i.e., while solving the problem, is the so-called Adaptive parameter control. It constantly adapts the values of the parameters based on feedbacks received from the search process. Some algorithms have been recently proposed for the on-line adaptation of CR and F, and for the autonomous control of which of the strategies should be applied at each instant of the search, the latter being commonly referred to as Adaptive Operator Selection (AOS). Some DE algorithms using adaptive methods can be found in the literature, such as SaDE [311], JADE [374], jDE [83] and ISADE [206]. Regarding DE for MOPs, there also exists some pioneering works, such as JADE2 [373] and OW-MOSaDE [192]. However, to the best of our knowledge, the employment of both adaptive parameter control of CR and F, and adaptive operator (mutation strategy) selection, is still relatively scarce in the domain of MOPs.

In this work, we employ an adaptive parameter control of CR and F slightly different from the one employed by the JADE method [374], which adapts their values based on the recent success rate of the search process; and an AOS mechanism inspired from the PM-AdapSS-DE method [165], which uses the Probability Matching mechanism to select between the available mutation strategies, based on the normalized relative fitness improvements brought by their recent applications. The main contribution of this work lies in the porting of these adaptive methods to the multi-objective domain. More specifically, a novel method is proposed to partially evaluate the fitness of the solutions, referred to as Tree Neighborhood Density (TND) estimator. The aggregation of the TND with the Pareto Dominance Strength (brought from the SPEA2 [380] method) is the information used by the AOS mechanism to keep its operator preferences up-to-date, and by a novel replacement mechanism based on a three-step comparison scheme. Lastly, the output of this replacement mechanism defines the success rates used for the adaptive parameter control of CR and F. The resulting algorithm, referred to as Adaptive Multi-Objective DE (Adap-MODE), is assessed in the light of a set of multi-objective benchmark functions, and shows to achieve significantly better results than other state-of-the-art approaches (NSGA-II [118] and GDE3 [230]) and than its static and partially-adaptive variants in most of the cases.

The remainder of this report is organized as follows. Firstly, the background and some related work are briefly reviewed in Section 12.2. Then, our proposed algorithm is described in detail in Section 12.3. After that, some experimental results are analyzed in Section 12.4. Finally, Section 12.5 concludes this report and gives possible directions for further work.

## 12.2 Related Work

The performance of an Evolutionary Algorithm (EA) strongly depends on the setting of some of its parameters. Section 12.2.1 will briefly overview the different ways of doing parameter setting in EAs, focusing on the kind of approach used in this work, referred to as Adaptive Parameter Control. Then, Section 12.2.2 will survey more specifically the Adaptive Operator Selection (AOS) paradigm.

### 12.2.1 Parameter Setting in Evolutionary Algorithms

There are different ways of doing parameter setting in EAs, as acknowledged by the well-known taxonomy proposed by Eiben et al. in [135]. In the higher level, there is the separation between Parameter Tuning and Parameter Control methods. Parameter Tuning methods set the parameters off-line, based on statistics over several runs; besides being computationally expensive, it provides a single parameter setting, that remains static during all the run. Parameter Control methods continuously adapt the parameters on-line, i.e., while solving the problem; these methods are further sub-divided into three branches, as follows.

The Deterministic methods adapt the parameter values according to pre-defined (deterministic) rules; but the definition of these rules already defines a complex optimization problem *per se*, besides hardly adapting to different problems. The Self-Adaptive methods adapt the parameter values *for free*, by encoding them within the candidate solution and letting the evolution take care of their control; in this case, however, the search space of the parameter values is aggregated to that of the problem, what might significantly increase the overall complexity of the search process. Lastly, the Adaptive methods control the parameter values based on feedback received from the previous search steps of the current optimization process.

In this work, we use an adaptive method very similar to the one proposed in the JADE algorithm [374], which controls the values of DE crossover rate CR and mutation scaling factor F based on the recent success rate (more details in Section 12.3.4). Another example of adaptive method proposed for the same aim is the SaDE [311] algorithm. Furthermore, another kind of adaptive method is also used in our algorithm, the AOS, surveyed in the following.

### 12.2.2 Adaptive Operator Selection

A recent paradigm, referred to as Adaptive Operator Selection (AOS), proposes the autonomous control of which operator (or mutation strategy in the case of DE) should be applied at each instant of the search, while solving the problem, based on their recent performance. A general AOS method usually consists of two components: the Credit Assignment scheme defines how each operator should be rewarded based on the impacts of its recent applications on the search progress; and the Operator Selection mechanism decides which of the available operators should be applied next, according to their respective empirical quality estimates, which are built and constantly updated by the rewards received. Each of these components will now be briefly overviewed in turn.

#### Credit Assignment

The most common way of assessing the impact of an operator application is the fitness improvement achieved by the offspring generated by its application, with respect to a baseline individual. In [165], the fitness improvement with respect to its parent is considered, while [113] use as baseline individual the best individual of the current population.

Based on this impact assessment, different ways of assigning credit to the operators can be found, in addition to the common average of the recent fitness improvements. In [361], a statistical technique rewards the operators based on their capability of generating outlier solutions, arguing that rare but highly beneficial improvements might be more important than frequent small improvements. Along the same line, in [148] each operator is rewarded based on the extreme (or maximal) fitness improvement recently achieved by it. In the quest for a more robust rewarding, in [147] a rank-based scheme is proposed. In multi-modal problems, however, the diversity is also important; in [269], both diversity variation and fitness improvement are combined to evaluate the operator application.

#### Operator Selection

The Operator Selection mechanism usually keeps an empirical quality estimate for each operator, built by the received rewards, which is used to guide its selection. The most popular method for Operator Selection is referred to as Probability Matching (PM) [345]: basically, the probability of selecting each operator is proportional to its empirical quality estimate with respect to the others; this is the method used in this work, more details in Section 12.3.3.

Other more complex Operator Selection methods worth to be mentioned are: the Adaptive Pursuit (AP) [345], originally proposed for learning automata, employs a winner-takes-all strategy to enforce a higher exploitation of the best operator; and the Dynamic Multi-Armed Bandit (DMAB) [148], which tackles the Operator Selection problem as yet another level of the Exploration vs. Exploitation dilemma, efficiently exploiting the current best operator, while minimally exploring the others, inspired from the multi-armed bandit paradigm.

## 12.3 Adaptive Multi-Objective DE

The general framework of the proposed adaptive Differential Evolution (DE) algorithm for multi-objective problems is illustrated in Fig. 12.1. As can be seen, it is divided into three modules. In the middle, there is the main cycle of the DE algorithm, represented here by only three steps for the sake of brevity: once after every generation, the fitness (see Section 12.3.1) of each offspring is evaluated by the sum of its Pareto Dominance (PD) strength and its Tree Neighborhood Density (TND). While the PD enforces convergence towards the Pareto front, the TND promotes diversification between the non-dominated solutions. These two measures are separately used by the Replacement mechanism, that decides which of the individuals should be maintained for the next generation by means of an original three-step comparison procedure (Section 12.3.2).

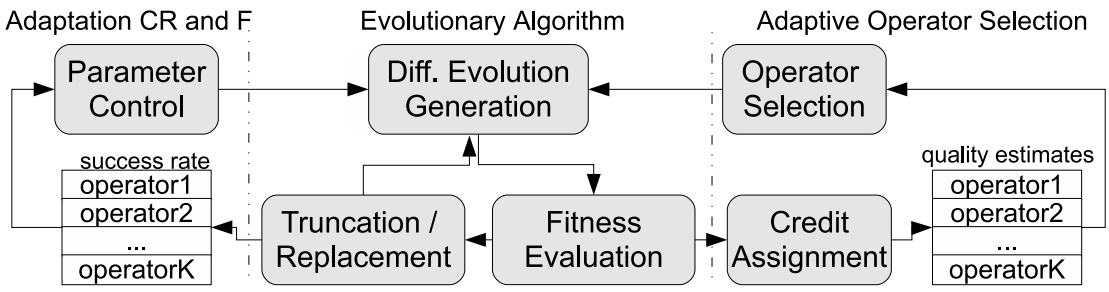



Figure 12.1: The framework of the proposed adaptive Differential Evolution algorithm

Two adaptive mechanisms are employed in parallel. On the right side, there is the AOS module, inspired from the PM-AdapSS-DE algorithm [165]. And on the left side, there is the Adaptive Parameter Control module slightly modified from the JADE algorithm [374]. Both adaptive mechanisms are described, respectively, in Sections 12.3.3 and 12.3.4. Although these are adaptive mechanisms brought from the literature, it is worth noting that in this work they are originally ported to the multi-objective domain, by receiving inputs based on the special aggregation between the PD and the novel TND measures.

### 12.3.1 Fitness Evaluation

In multi-objective optimization, the aims of the search can be said to be two-fold. On the one hand, the solutions found should approach as much as possible to the Pareto front. On the other hand, the non-dominated solutions should be distributed over the Pareto front as uniformly as possible, in order to have satisfiable solutions for all the different objectives. In this work, we use the Pareto Dominance (PD) strength proposed in the SPEA2 algorithm [380] to enforce the first issue (convergence). For the second issue, we propose a novel measure to promote spread between the non-dominated solutions, referred to as the Tree Neighborhood Density (TND). The fitness of each individual is assessed by an aggregation of these two criteria, as described in the following.

#### Pareto Dominance Strength

In order to calculate the Pareto Dominance (PD) strength, we use the mechanism proposed in the SPEA2 algorithm [380]. The only difference is that the external archive to store elite individuals is not implemented here. Briefly, a strength value $S(i)$ is assigned to each individual $i$ in the population $P$, representing the number of solutions it dominates. If solely based on this criterion, the fitness of each individual $i$, referred to as $PD(i)$ here, would be calculated as:

$$PD(i) = \sum_{j \in P, j \succ i} S(j) \tag{12.1}$$

i.e., the sum of the strengths of all the individuals that dominate individual $i$. Intuitively, the smaller the better, with $PD(i) = 0$ corresponding to a non-dominated solution; whereas a large $PD(i)$ means that the individual $i$ is dominated by many others.

#### Tree Neighborhood Density

As previously mentioned, the Tree Neighborhood Density (TND) is a novel estimation proposed to enforce a higher level of spread between the non-dominated solutions. For the sake of a clearer discussion, some definitions and terminologies are firstly given as follows.

**Tree crowding density:** Let $T$ be a minimum spanning tree connecting all the individuals of population $P$. For any individual $i$ in $P$, let $d_i$ be the degree of $i$ in $T$, i.e., the number of edges of $T$ connected to $i$; and let these edges be $\{l_{i,1}, l_{i,2}, \ldots, l_{i,d_i}\}$. The tree crowding density of $i$ is estimated as:

$$T_{crowd}(i) = \sum_{j=1}^{d_i} l_{i,j}/d_i \tag{12.2}$$

**Tree neighborhood:** Let $r_i = \max\{l_{i,1}, l_{i,2}, \ldots, l_{i,d_i}\}$. A circle centered in individual $i$, with radius $r_i$, is defined as the tree neighborhood of $i$.

**Membership of individual on the tree neighborhood:** Let the Euclidean distance between individuals $i$ and $j$ be denoted as $dist_{i,j}$. The individual $j$ is considered as a member of the tree neighborhood of $i$ if and only if $dist_{i,j} \leq r_i$ (denoted as $i \rhd_T j$).

Based on these definitions, the calculation procedure for the Tree Neighborhood Density (TND) is implemented as follows:

1. The Euclidean distance between each individual of the population $P$ with the other $NP - 1$ individuals is calculated;
2. A minimal spanning tree $T$ connecting all individuals is generated;
3. The tree crowding density for each individual $i$ in $T$ is assessed, and the corresponding tree neighborhood is generated;
4. For each individual $i$, the degrees of the individuals pertaining to its tree neighborhood are summed:
$$sumdegrees(i) = \sum_{j \in U} d_j, \text{ where } U = \{j | j \in P, i \rhd_T j\} \tag{12.3}$$
5. Then, the Tree Neighborhood Density of individual $i$ is calculated as:
$$TND(i) = \frac{\sum_{j \in U}(1/Tcrowd_j)}{sumdegrees(i)} \tag{12.4}$$
6. Finally, the TND values of all individuals are normalized:
$$nTND(i) = \frac{TND(i) - TND_{min}}{TND_{max} - TND_{min}}. \tag{12.5}$$
where $nTND(i)$ is the normalized TND of individual $i$, and $TND_{max}$ and $TND_{min}$ indicate, respectively, the maximum and minimum TND in the current population.

In the same way as for the $PD$ measure, the smaller TND the better. The underlying motivation for its proposal can be explained as follows. The whole set of solutions in the population can be regarded as a connected graph, with the Euclidean minimum spanning tree of this graph being an optimized structure that reflects the distribution of the solutions of the current population in the search space. Then, for a given individual, the corresponding neighborhood can be defined by the other individuals connected to it, and finally, the crowdedness of this neighborhood can be said to represent its density.

**Aggregated Fitness Evaluation**

Based on the aforementioned discussion, the fitness value (to be minimized) of each individual $i$ is calculated as the sum of both criteria:

$$f(i) = PD(i) + nTND(i) \tag{12.6}$$

It is worth noting that only the TND measure is normalized between 0 and 1. Hence, evolution proceeds by firstly minimizing PD, i.e., approaching the Pareto front; and then, as soon as some non-dominated solutions (i.e., with $PD = 0$) are found, $nTND$ becomes significant in the fitness evaluation, and a higher spread between the non-dominated solutions is promoted.

### 12.3.2 Replacement Mechanism

At each generation, each of the $NP$ parental solutions is used to generate other $NP$ offspring solutions. In the original DE algorithm, the offspring replaces its parent in the next generation if it has a better fitness value. In the case of multi-objective optimization, a different replacement mechanism is needed in order to incorporate the already mentioned properties of this kind of problem. To this aim, a three-step comparison method is proposed in this work, as follows.

Starting from the mixed population of size $2 \times NP$, containing the $NP$ parental and the $NP$ offspring individuals, firstly, the Pareto dominance relationship is considered: each pair (parent, offspring) is compared at a time, and the dominated one is immediately rejected.

In case the mixed population is still bigger than $NP$, the replacement mechanism proceeds to the second step, which uses the non-dominated sorting method proposed in the NSGA-II algorithm [118]. Briefly, at each round, the non-dominated individuals of the mixed population are chosen to survive to the next generation, and are removed from the mixed population. This is done iteratively up to the completion of the population for the next generation (i.e., $NP$ chosen individuals after the first and second steps), or until there are no less than $NP$ individuals with assigned rank values in the population.

If there are still individuals to be filtered for the next generation, the third step finally considers the TND values. At each iteration, the individual that has the lowest TND (i.e., the most crowded individual) is maintained, until the exact number of individuals for the completion of the new population is achieved.

### 12.3.3 Adaptive Operator Selection

As surveyed in Section 12.2.2, to implement the AOS paradigm, there is the need of defining two elements, the Credit Assignment and the Operator Selection mechanisms. The approaches used in this work will be now detailed in turn.

#### Credit Assignment: Normalized Relative Fitness Improvement

The Credit Assignment scheme is inspired from the one used in the PM-AdapSS-DE algorithm [165]; the differences are the use of a different and normalized calculation of the relative fitness improvements (which showed to perform better after some preliminary experiments) and in the already described fitness evaluation, specially designed for multi-objective optimization.

The impact of each operator application $i$ is evaluated as the normalized relative fitness improvement $\eta_i$ achieved by it, measured as:

$$\eta_i = \frac{|pf_i - cf_i|}{|f_{best} - f_{worst}|} \tag{12.7}$$

where $f_{best}$ (respectively $f_{worst}$) is the fitness value of the best (respectively the worst) solution in the current population; $pf_i$ and $cf_i$ are the fitness values of the (parent) target vector and its offspring, respectively. As in [165], in case no improvement is achieved i.e., $pf_i - cf_i \geq 0$, $\eta_i$ is set to zero.

All the normalized relative fitness improvements achieved by the application of operator (mutation strategy in this case) $a \in \{1, \ldots, K\}$ during each generation $g$ are stored in a specific set $R_a$. Following [165], at the end of each generation $g$, a unique credit (or reward) is assigned to each operator, calculated as the average of all the normalized relative fitness improvements achieved by it:

$$r_a(g) = \sum_{i=1}^{|R_a|} \frac{R_a(i)}{|R_a|}. \tag{12.8}$$

**Operator Selection: Probability Matching**

The Operator Selection mechanism used is the Probability Matching (PM) [345]. Formally, let the strategy pool be denoted by $S = \{s_1, \ldots, s_K\}$ where $K > 1$. The probability vector $P(g) = \{p_1(g), \ldots, p_K(g)\}(\forall t : p_{min} \leq p_i(g) \leq 1; \sum_{i=1}^{K} p_i(g) = 1)$ represents the selection probability of each operator at generation $g$. At the end of every generation, the PM technique updates the probability $p_a(g)$ of each operator $a$ based on the received reward $r_a(g)$, as follows. Firstly, the empirical quality estimate $q_a(g)$ of operator $a$ at generation $g$ is updated as [345]:

$$q_a(g+1) = q_a(g) + \alpha \left[ r_a(g) - q_a(g) \right] \tag{12.9}$$

where $\alpha \in (0, 1]$ is the adaptation rate; the selection probability is updated as:

$$p_a(t+1) = p_{min} + (1 - K \cdot p_{min}) \frac{q_a(g+1)}{\sum_{i=1}^{K} q_i(g+1)}. \tag{12.10}$$

where $p_{min} \in (0, 1)$ is the minimal selection probability value of each operator, used to ensure that all the operators have a minimal chance of being selected. The rationale for this minimal exploration is that the operators that are currently performing badly might become useful at a further moment of the search [345].

### 12.3.4 Adaptive Parameter Control of CR and F

The parameter adaptation method used here is similar to that used in the JADE algorithm [374]. Let $CR_i^a$ denote the crossover rate for the individual $i$ using operator $a \in \{1, \ldots, K\}$. At each generation, $CR_i^a$ is independently generated according to a normal distribution with mean $\mu_{CR}^a$ and standard deviation 0.1:

$$CR_i^a = norm(\mu_{CR}^a, 0.1) \tag{12.11}$$

being regenerated whenever it exceeds 1. All successful crossover rates at generation $g$ for operator $a$ are stored in a specific set denoted as $S_{CR}^a$. The mean $\mu_{CR}^a$ is initialized to a user defined value and updated after each generation as:

$$\mu_{CR}^a = (1 - c) \cdot \mu_{CR}^a + c \cdot mean(S_{CR}^a) \tag{12.12}$$

where $c$ is a constant and $mean(S_{CR}^a)$ is the arithmetic mean of values in $S_{CR}^a$.

An analogous adaptation mechanism is used for the scaling factor $F_i^a$. After some preliminary experiments, a difference with respect to the JADE algorithm [374] at this point is that the mean value $\mu_F^a$ is calculated by the root-mean-square of the values in $S_F^a$, instead of Lehmer mean.

## 12.4 Performance Comparison

In this section, three different empirical comparisons are presented. Firstly, the proposed Adap-MODE is compared with two state-of-the-art MOEAs, namely, NSGA-II [118] and GDE3 [230]. Then, in order

to assess the benefits brought by the combined use of both adaptive parameter control modules, Adap-MODE is compared with four static variants, each using one of the four mutation strategies and a fixed values for control parameters ($CR = 0.5, F = 1.0$). Lastly, we compare Adap-MODE with its "partially-adaptive" variants, namely, the same MODE but using only AOS (and $CR = 0.5, F = 1.0$), and the same MODE but using only the adaptive parameter control of CR and F (and the mutation strategies being uniformly selected). This latter is done in order to evaluate the gain achieved by the combination of both modules, compared with each of the modules being independently applied.

### 12.4.1 Experimental Settings

For the sake of a fair empirical comparison, the parameters of the two state-of-the-art MOEAs are set as in the respective original papers. For the NSGA-II [118], $\eta_c = \eta_m = 20, p_c = 0.9, p_m = 1/D$, with $D$ representing the dimension of the problem; and for GDE3 [230], $CR = 0.5, F = 1.0$. For the parameters of the proposed Adap-MODE method, the PM adaptation rate is set to $\alpha = 0.3$ and minimal probability $p_{min} = 0.05$, as in [165]; and the parameter $c$ for the adaptive parameter control of CR and F is set to 0.1, as in [374], with CR and F being both initialized to 0.2. Lastly, the DE population size is set to 100.

In this work, the AOS mechanism implemented in Adap-MODE is used to select between the following four DE mutation strategies: (1) DE/rand/1/bin, (2) DE/current-to-rand/1/bin, (3) DE/rand/2/bin, and (4) DE/rand-to-best/2/bin. These are the same strategies used in some previous works [311, 165]; no theoretical or empirical analysis was preliminary performed for their choice. It is worth highlighting that the AOS scheme is generic: any other set of mutation strategies could be considered here.

In order to compare the performance of the proposed and baseline approaches, ZDT [377] and DTLZ [120] test suites are considered as benchmark functions. The maximum number of generations is set to 300 for ZDT, and to 500 for DTLZ.

Two assessment metrics are used to quantitatively evaluate the performance of each algorithm at the end of each run, averaged over 50 runs. The Uniform Assessment (UA) metric [248] is used to evaluate the spread of the solutions, while the Hyper-Volume (HV) [382] is a comprehensive performance indicator. Generally, for the values of both UA and HV, the larger the better.

### 12.4.2 Experimental Results

The comparative results, for each of the are presented in Tables 12.1 to 12.3. Following the central limit theorem, we assume that the sample means are normally distributed; therefore, the paired $t$-test statistical test at 95% confidence level is adopted to compare the significance between two competing algorithms, with the † indicating that Adap-MODE is significantly better than all its competitors in the corresponding Table, and ‡ representing that the best competitor significantly outperforms Adap-MODE. Moreover, the best results for each metric on each problem function are highlighted in **boldface**.

Starting with the comparison between Adap-MODE and the two state-of-the-art MOEAs, namely NSGA-II and GDE3, the results are presented in Table 12.1. These results clearly show that Adap-MODE is the best choice when compared to its competitors: it achieves the best results in 23 out of the 24 performance metrics, performing significantly better in 22 of them. The only exception is for the UA metric in the DTLZ4 problem, in which NSGA-II wins. It is worth noting that Adap-MODE performs around two times better than its competitors w.r.t. the uniformity metric UA in most of the functions, what might be largely attributed to the use of the proposed tree neighborhood density estimator by the fitness assignment.

Table 12.2 compares the performance of Adap-MODE with four static variants of it, each using one of the four available mutation strategies, without any adaptive parameter control. From these results, it becomes clear that there is no single mutation strategy that is the best over all the functions. For example,

Table 12.1: Comparative results of NSGA-II, GDE3 and Adap-MODE

| | | NSGA-II | GDE3 | Adap-MODE | S |
|---|---|---|---|---|---|
| ZDT1 | UA | 4.433e-1/3.56e-2 | 2.359e-1/4.42e-2 | **8.080e-1/1.62e-2** | † |
| | HV | 3.65960/3.00e-4 | 3.65990/3.55e-4 | **3.66193/3.15e-5** | † |
| ZDT2 | UA | 4.391e-1/4.68e-2 | 2.551e-1/4.98e-2 | **8.069e-1/1.89e-2** | † |
| | HV | 3.32618/3.21e-4 | 3.32673/3.06e-4 | **3.32853/4.19e-5** | † |
| ZDT3 | UA | 4.252e-1/4.49e-2 | 2.069e-1/4.17e-2 | **7.660e-1/1.98e-2** | † |
| | HV | 4.80650/5.13e-2 | 4.81433/1.95e-4 | **4.81463/4.81e-4** | † |
| ZDT4 | UA | 4.173e-1/4.69e-2 | 2.403e-1/4.64e-2 | **8.055e-1/1.85e-2** | † |
| | HV | 3.65413/4.04e-3 | 3.63033/1.84e-1 | **3.66201/5.33e-4** | † |
| ZDT6 | UA | 4.529e-1/4.86e-2 | 2.226e-1/4.67e-2 | **7.896e-1/2.27e-2** | † |
| | HV | 3.03090/1.51e-3 | 3.04029/2.67e-4 | **3.04183/1.62e-5** | † |
| DTLZ1 | UA | 3.742e-1/4.44e-2 | 5.256e-1/3.58e-2 | **8.246e-1/1.48e-2** | † |
| | HV | 0.967445/1.95e-3 | 0.965469/9.45e-4 | **0.973582/2.75e-4** | † |
| DTLZ2 | UA | 3.688e-1/3.78e-2 | 4.868e-1/3.30e-2 | **8.236e-1/1.84e-2** | † |
| | HV | 7.33017/2.70e-2 | 7.31392/9.05e-3 | **7.40523/1.14e-2** | † |
| DTLZ3 | UA | 3.353e-1/7.92e-2 | 4.857e-1/4.09e-2 | **8.304e-1/1.72e-2** | † |
| | HV | 6.41853/1.80e+0 | 5.85267/2.37e+0 | **7.32465/5.76e-1** | † |
| DTLZ4 | UA | **4.404e-1/9.19e-2** | 2.532e-1/3.91e-2 | 2.654e-1/2.99e-2 | ‡ |
| | HV | 6.90792/7.55e-1 | 5.46000/1.10e+0 | **7.02943/5.46e-1** | † |
| DTLZ5 | UA | 3.930e-1/4.63e-2 | 4.379e-1/3.85e-2 | **7.866e-1/1.82e-2** | † |
| | HV | 6.10048/1.42e-3 | 6.08543/1.83e-3 | **6.10548/4.40e-3** | |
| DTLZ6 | UA | 2.939e-1/5.19e-2 | 2.652e-1/4.33e-2 | **7.759e-1/2.18e-2** | † |
| | HV | 5.86932/7.09e-2 | 6.10187/2.12e-3 | **6.10732/4.88e-3** | † |
| DTLZ7 | UA | 4.102e-1/3.96e-2 | 4.491e-1/3.70e-2 | **7.723e-1/1.86e-2** | † |
| | HV | 13.15151/8.55e-2 | 13.19772/9.29e-2 | **13.46486/7.43e-2** | † |

for the ZDT2 function, strategy 2 is the best in terms of HV, while strategy 1 is the winner for ZDT3. It is also worth noting that strategy 4 performs worst, while strategies 1 and 3 are the most competitive ones. This kind of situation motivates the use of the AOS paradigm. And indeed, Adap-MODE remains the best option in most of the functions, while achieving very similar performance in others.

Table 12.2: Comparative results of Adap-MODE and its pure versions, following the same order of the problems as in Table 12.1

| | Str.1 | Str.2 | Str.3 | Str.4 | Adap-MODE | S |
|---|---|---|---|---|---|---|
| UA | 7.9e-1/1.9e-2 | 7.9e-1/1.9e-2 | 7.4e-1/2.6e-2 | 4.1e-1/5.2e-2 | **8.1e-1/1.6e-2** | † |
| HV | 3.662/3.4e-5 | 3.662/3.4e-5 | 3.656/2.2e-3 | 1.902/3.9e-1 | **3.662/3.1e-5** | |
| UA | 7.9e-1/1.5e-2 | 8.0e-1/1.9e-2 | 7.2e-1/2.7e-2 | 3.6e-1/6.3e-2 | **8.1e-1/1.9e-2** | |
| HV | 3.328/3.4e-5 | **3.328/3.9e-5** | 3.319/4.6e-3 | 1.905/2.4e-1 | 3.328/4.2e-5 | ‡ |
| UA | **7.7e-1/2.1e-2** | 7.5e-1/2.7e-2 | 5.1e-1/8.0e-2 | 3.4e-1/2.7e-2 | 7.6e-1/1.9e-2 | |
| HV | **4.815/6.4e-5** | 4.814/1.6e-3 | 4.775/1.3e-2 | 1.781/3.4e-1 | 4.814/4.8e-4 | ‡ |
| UA | **8.1e-1/1.7e-2** | 8.0e-1/1.6e-2 | 8.0e-1/1.9e-2 | 3.3e-1/3.9e-2 | 8.0e-1/1.8e-2 | |
| HV | 3.636/1.0e-1 | 3.662/3.8e-5 | 3.649/8.6e-2 | 0.0/0.0 | **3.662/5.3e-4** | |
| UA | 7.9e-1/1.9e-2 | 8.1e-1/2.0e-2 | **8.2e-1/1.9e-2** | 7.6e-1/4.6e-2 | 7.9e-1/2.2e-2 | ‡ |
| HV | 3.042/1.7e-5 | 3.042/2.4e-5 | **3.042/1.5e-5** | 3.041/3.1e-3 | 3.042/1.6e-5 | |
| UA | **8.3e-1/2.1e-2** | 8.2e-1/1.7e-2 | 8.2e-1/1.5e-2 | 4.7e-1/4.3e-2 | 8.2e-1/1.5e-2 | |
| HV | 0.97/1.0e-3 | 0.97/5.6e-4 | 0.969/7.1e-4 | 0.0/0.0 | **0.973/2.7e-4** | † |
| UA | 8.1e-1/1.8e-2 | 8.0e-1/2.0e-2 | 8.0e-1/1.9e-2 | 7.9e-1/2.5e-2 | **8.2e-1/1.8e-2** | |
| HV | 7.348/1.4e-2 | 7.337/1.4e-2 | 7.335/7.8e-3 | 7.303/8.8e-3 | **7.405/1.1e-2** | † |
| UA | 8.0e-1/1.8e-2 | 3.5e-1/3.9e-2 | 3.4e-1/3.2e-2 | 4.0e-1/3.9e-2 | **8.3e-1/1.7e-2** | † |
| HV | 6.538/2.0 | 0.0/0.0 | 0.0/0.0 | 0.0/0.0 | **7.324/5.7e-1** | † |
| UA | 2.5e-1/3.8e-2 | 2.4e-1/3.3e-2 | 2.5e-1/3.0e-2 | 2.3e-1/3.0e-2 | **2.6e-1/2.9e-2** | † |
| HV | 5.58/1.1 | 6.639/4.8e-1 | 6.359/7.7e-1 | 5.971/1.1 | **7.029/5.4e-1** | † |
| UA | 7.4e-1/2.2e-2 | 7.2e-1/2.3e-2 | 7.2e-1/2.1e-2 | 7.3e-1/2.8e-2 | **7.8e-1/1.8e-2** | † |
| HV | 6.073/3.4e-3 | 6.067/3.4e-3 | 6.065/3.9e-3 | 6.052/5.3e-3 | **6.105/4.4e-3** | † |
| UA | 7.9e-1/2.0e-2 | 7.9e-1/2.1e-2 | **7.9e-1/1.7e-2** | 7.6e-1/2.5e-2 | 7.7e-1/2.2e-2 | ‡ |
| HV | 6.107/4.4e-3 | 6.106/4.2e-3 | **6.108/5.4e-3** | 5.764/1.0 | 6.107/4.9e-3 | |
| UA | 7.6e-1/1.8e-2 | 7.7e-1/1.6e-2 | 7.4e-1/2.2e-2 | 5.1e-1/1.3e-1 | **7.7e-1/1.8e-2** | |
| HV | 13.412/5.6e-2 | 13.427/4.8e-2 | 13.346/7.3e-2 | 7.735/3.7 | **13.46/7.4e-2** | † |

The last comparative results, shown in Table 12.3, presents the performance of Adap-MODE compared with its "partially"-adaptive variants, one using only the AOS, and the other using only the adaptive parameter control of CR and F. From these results, it is not clear which of the adaptive modules is the most beneficial for the performance of Adap-MODE: at some functions, the "AOS only" method is better than the "parameter control only" one, while in others the opposite occurs. But these results clearly demonstrate that the combined use of both adaptive modules is better than their sole use, what is shown

by the fact that Adap-MODE significantly outperforms them in most of functions, in terms of both UA and HV.

Table 12.3: Comparative results of Adap-MODE, Adap-MODE with AOS only and Adap-MODE with parameter control only

| | | CR/F(fixed)+AOS | CR/F(adapt.)+Unif.OS | Adap-MODE | S |
|---|---|---|---|---|---|
| ZDT1 | UA | 7.860e-1/2.08e-2 | 7.851e-1/2.42e-2 | **8.080e-1/1.62e-2** | † |
| | HV | 3.66162/2.97e-4 | 3.66066/2.69e-4 | **3.66193/3.15e-5** | † |
| ZDT2 | UA | 7.809e-1/2.02e-2 | 7.793e-1/1.71e-2 | **8.069e-1/1.89e-2** | † |
| | HV | 3.32840/3.13e-4 | 3.32612/5.27e-4 | **3.32853/4.19e-5** | † |
| ZDT3 | UA | 7.538e-1/2.83e-2 | 7.487e-1/1.52e-2 | **7.660e-1/1.98e-2** | † |
| | HV | 4.81448/1.18e-3 | 4.81228/1.18e-3 | **4.81463/4.81e-4** | |
| ZDT4 | UA | **8.127e-1/2.30e-2** | 7.486e-1/6.12e-2 | 8.055e-1/1.85e-2 | |
| | HV | 3.64150/1.43e-1 | 3.65409/4.26e-2 | **3.66201/5.33e-4** | † |
| ZDT6 | UA | 7.626e-1/2.34e-2 | **8.078e-1/2.34e-2** | 7.896e-1/2.27e-2 | ‡ |
| | HV | 3.04179/3.22e-5 | 3.04183/4.93e-5 | **3.04183/1.62e-5** | |
| DTLZ1 | UA | **8.247e-1/1.80e-2** | 8.200e-1/1.73e-2 | 8.246e-1/1.48e-2 | |
| | HV | 0.969925/5.41e-4 | 0.917842/1.25e-1 | **0.973582/2.75e-4** | † |
| DTLZ2 | UA | 8.096e-1/2.01e-2 | 8.224e-1/1.56e-2 | **8.236e-1/1.84e-2** | |
| | HV | 7.33762/1.10e-2 | 7.40368/9.20e-3 | **7.40523/1.14e-2** | |
| DTLZ3 | UA | 6.365e-1/1.44e-1 | 8.289e-1/1.42e-2 | **8.304e-1/1.72e-2** | |
| | HV | 7.13704/3.70e-1 | 4.59535/2.92e+0 | **7.32465/5.76e-1** | † |
| DTLZ4 | UA | 2.092e-1/3.32e-2 | 9.814e-2/4.33e-3 | **2.654e-1/2.99e-2** | † |
| | HV | 6.78321/6.03e-1 | 4.66216/1.09e+0 | **7.02943/5.46e-1** | † |
| DTLZ5 | UA | 7.334e-1/2.26e-2 | 7.792e-1/1.95e-2 | **7.866e-1/1.82e-2** | † |
| | HV | 6.07005/3.69e-3 | **6.10649/3.78e-3** | 6.10548/4.40e-3 | |
| DTLZ6 | UA | 7.739e-1/2.32e-2 | **7.876e-1/2.00e-2** | 7.759e-1/2.18e-2 | ‡ |
| | HV | **6.10841/5.67e-3** | 6.10640/4.14e-3 | 6.10732/4.88e-3 | |
| DTLZ7 | UA | 7.621e-1/1.83e-2 | 7.634e-1/1.70e-2 | **7.723e-1/1.86e-2** | † |
| | HV | 13.42436/6.19e-2 | 13.43145/7.25e-2 | **13.46486/7.43e-2** | † |

## 12.5 Conclusion

In this report, we propose a new DE algorithm for multi-objective optimization that uses two adaptive mechanisms in parallel: the Adaptive Operator Selection mechanism, to control which operator should be applied at each instant of the search; and the Adaptive Parameter Control, that adapts the values of the DE parameters CR and F while solving the problem. A tree neighborhood density estimator is proposed and, combined with the Pareto dominance strength measure, is used in order to evaluate the fitness of each individual. Additionally, a novel replacement mechanism is proposed, based on a three-step comparison procedure. As a consequence, the adaptive methods employed by the proposed algorithm, inspired from recent literature, are originally ported to the multi-objective domain.

Numerical experiments demonstrate that the proposed Adap-MODE is capable of efficiently adapting to the characteristics of the region that is currently being explored by the algorithm, by efficiently selecting appropriate operators and their corresponding parameters. Adap-MODE is shown to outperform two state-of-the-art MOEAs, namely NSGA-II [118] and GDE3 [230], in most of the functions. It also performs significantly better, in most of the functions, than the same MODE with static parameters, and than the partially-adaptive variants using each of the two adaptive modules.

But there is still a lot of space for improvements. Firstly, for the fitness evaluation, more sophisticated schemes to control the balance between both convergence and spread could be analyzed. Regarding the AOS implementation, other schemes have already shown to perform better than PM in the literature and should also be analyzed in the near future, such as the Adaptive Pursuit [345] and the Dynamic Multi-Armed Bandit [148]; a more recent work, that also use bandits, reward the operators based on ranks [147], thus achieving a much higher robustness w.r.t. different benchmarking situations. In the same way, there are different alternatives for the adaptive parameter control of CR and F that could be further explored.

Another issue that deserves further exploration is related to the (hyper) parameters of the adaptive modules. In the case of Adap-MODE, the AOS requires the definition of the adaptation rate $\alpha$ and the minimum probability $p_{min}$, while the adaptive parameter control requires the setting of $c$. In this work, these parameters were set as in the original references, but further analysis of their sensitivity should be done. Ideally, Adap-MODE and the other methods used as baseline should also be all compared again, after a proper off-line tuning phase. Another important baseline would be the same MODE with off-line tuned CR, F, and mutation application rates.

Lastly, the extra computational time resulting from the use of these adaptive schemes should be further analyzed; although it is true to say that, in real-world problems, the fitness evaluation is usually the most computationally expensive step, all the rest becoming negligible.

# Chapter 13

# Analyzing the Impact of Mirrored Sampling and Sequential Selection in Elitist Evolution Strategies

Anne Auger, Dimo Brockhoff, Nikolaus Hansen

*Accepted in Foundations of Genetic Algorithms (FOGA) 2011*

*This report presents a refined single parent evolution strategy that is derandomized with mirrored sampling and/or uses sequential selection. The paper analyzes some of the elitist variants of this algorithm. We prove, on spherical functions with finite dimension, linear convergence of different strategies with scale-invariant step-size and provide expressions for the convergence rates as the expectation of some known random variables. In addition, we derive explicit asymptotic formulae for the convergence rate when the dimension of the search space goes to infinity. Convergence rates on the sphere reveal lower bounds for the convergence rate of the respective step-size adaptive strategies. We prove the surprising result that the* (1+2)*-ES with mirrored sampling converges at the same rate as the* (1+1)*-ES without and show that the tight lower bound for the* (1+$\lambda$)*-ES with mirrored sampling and sequential selection improves by 16% over the* (1+1)*-ES reaching an asymptotic value of about* $-0.235$.

## 13.1 Introduction

Evolution Strategies (ESs) are robust search algorithms designed to minimize objective functions $f$ that map a continuous search space $\mathbb{R}^{\mathbb{N}}$ into $\mathbb{R}$. In a $(1 \overset{+}{,} \lambda)$-ES, $\lambda$ candidate solutions, the offspring, are created from a single parent, $\underline{X}_k \in \mathbb{R}^{\mathbb{N}}$. The $\lambda$ offspring are generated by adding $\lambda$ *independent* random vectors $(\underline{\mathcal{N}}_k^i)_{1 \le i \le \lambda}$ to $\underline{X}_k$. Then, the *best* of the $\lambda$ offspring $\underline{X}_k + \underline{\mathcal{N}}_k^i$ in case of comma selection or of the $\lambda$ offspring plus parent in case of plus selection is *selected* to become the next parent $\underline{X}_{k+1}$. The (1+1)-ES is arguably the most local, and the locally fastest, variant of an evolution strategy.

Derandomization of random numbers is a general technique where the independent samples are replaced by dependent ones with the objective of accelerating algorithm convergence. Derandomization by means of antithetic variables for isotropic samples was first introduced within general ESs in [344]. Mirrored samples, as used in this report, are a special case, where the number of independent events is reduced by a factor of two only. In [343], the sequence of uniform random numbers used for

sampling a multivariate normal distribution was replaced by scrambling-Halton and Sobol sequences. These sequences achieved consistent improvements of CMA-ES (covariance matrix adaptation evolution strategy) mainly on unimodal test functions, typically with ≤30% speed-up and most pronounced in dimension 2. The improvements are however difficult to attribute to a cause for at least two reasons. First, in CMA-ES with $\mu > 1$, quasi-random numbers possibly introduce a (strong) bias on the step-size. For mirrored samples and Sobol sequences, we have verified this bias empirically (shown for mirrored sampling in [85]). The bias can improve convergence rates,[1] but violates the demand on a stochastic search algorithm to be unbiased [181, 175]. Second, random rotations of the quasi-random vector sets in [343] lead to a significant loss of the advantage. The investigated functions were however unrotated. This makes the identity as initial covariance matrix, represented in the given coordinate system and in connection with the quasi-random numbers, presumably a choice that is unintentionally biased towards the function testbed.

Consequently, it remains to be investigated to what extend the improvements can be attributed to a bias on the variance of the sum of selected vectors (leading to the bias on the step-size), to a coordinate system dependency, or to the quasi-random structure itself.

Our own experiments with derandomizations beyond mirroring, similar to those in [344], revealed the most pronounced effects (unsurprisingly) by mirroring and in small populations. We have not considered algorithms that are—by themselves or in combination with CMA-ES—biased or not rotationally invariant.

*Mirrored sampling* is a derandomization technique similar to antithetic variables that was recently introduced within $(1+\lambda)$ and $(1,\lambda)$-ESs [85]. In addition, mirrored sampling has been coupled with *sequential selection*, a modification of the $(1,\lambda)$ and $(1+\lambda)$ selection schemes where the offspring are evaluated sequentially and the iteration is concluded as soon as one offspring is better than its current parent [85].

Sequential selection and mirrored sampling have been implemented within the CMA-ES and extensively empirically studied on 54 noiseless [178] and noisy [179] functions in a series of papers [27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 38]. In summary, the variants with mirrored mutation and sequential selection improved their baseline algorithms (without these two ideas) on almost all functions for almost all target values where the combination of the two concepts was never statistically significantly worse than the standard algorithms. In particular for the elitist (1+1)-CMA-ES, additional mirrored mutation and sequential selection improved the performance by about 17% on the non-separable ellipsoid function, by about 20% on the ellipsoid, the discus, and the sum of different powers functions, and by 12% on the sphere function while no statistically significant worsening of the performance was reported [29].

So far, theoretical investigations of mirrored sampling and sequential selection is restricted to comma selection [85]. Convergence rates of the scale-invariant step-size $(1,\lambda)$-ES with mirrored sampling and sequential selection on spherical functions have been derived and lower bounds for the convergence of the different strategies were compared. Those results hold for finite dimensions of the search space. In this report, we aim at generalizing those theoretical results to plus selection: we extend finite dimension convergence proofs to plus selection and complement those results with asymptotic estimates of the convergence rates when the dimension goes to infinity.

The paper is structured as follows. In Section 13.2, we describe the $(1 \dagger \lambda)$-ES with mirrored sampling and sequential selection and derive general properties. In Section 13.3, we derive the linear convergence of the $(1+\lambda)$-ES with mirrored sampling and sequential selection with scale-invariant step-size on spherical functions. We express the convergence rate in terms of the expectation of a random variable. In addition, we establish that the (1+1)-ES and the $(1+2_m)$-ES with mirrored sampling exhibit the same convergence rate. In Section 13.5, we derive some simple expressions for the asymptotic normalized convergence rate of the different algorithms, where asymptotic refers to the dimension tending to infinity.

[1] For mirrored sampling this most probably happens if random vectors with different lengths are realized, which is the case in particular in small dimensions.

In Section 13.6, we numerically simulate the convergence rates for different dimensions and appraise quantitatively the improvements brought by mirrored sampling and sequential selection.

**Notations:** In this report, a multivariate normal distribution with mean vector zero and covariance matrix identity will be called *standard* multivariate normal distribution. The first vector $(1, 0, \ldots, 0)$ of the canonical basis will be denoted $\underline{e}_1$.

## 13.2 $(1 \overset{+}{,} \lambda)$-ES with Mirrored Sampling and Sequential Selection

In this section, we introduce the $(1 \overset{+}{,} \lambda)$-ES with mirrored sampling and sequential selection and derive general theoretical results on those algorithms.

### 13.2.1 Algorithm Description

Mirrored mutations and sequential selection have been introduced in [85] and are two independent ideas for improving simple local search strategies such as $(1 \overset{+}{,} \lambda)$-ESs. Algorithm 7 shows the pseudocode of a combination of both concepts within the $(1+\lambda)$-ES and the $(1, \lambda)$-ES. Note that we describe the algorithms without specifying which sampling distribution is used—though most of the time, for Evolution Strategies, multivariate normal distributions are used. However, since we will derive some results that are independent of the choice of the sampling distribution, we keep the description general and indicate when a standard multivariate normal distribution is required.

**Mirrored sampling:** The idea behind mirrored sampling is to derandomize the generation of new sample points. Instead of using two independent random vectors to create two offspring, with mirrored sampling only a single random vector instantiation is used to create two offspring: one by adding and the other by subtracting the vector from the current search point. The two instantiations are called *mirrored* or *symmetric* with respect to the parent $\underline{X}_k$ at iteration $k$ if they take place in the same iteration. For odd $\lambda$, every other iteration, the first offspring uses the mirrored last vector from the previous iteration, see $j$ in Algorithm 7. Consequently, in the $(1+1_{\mathrm{m}})$-ES, a mirrored sample is used if and only if the iteration index is even (given *skip mirror* in Algorithm 7 is false).

When evaluating a sampled solution and its mirrored counterpart, sometimes unnecessary function evaluations are performed: for example, on unimodal objective functions with convex sub-level sets, $\{x \mid f(x) \leq c\}$ for $c \in \mathbb{R}$, such as the sphere function, $f(x) = \|x\|^2$, the mirrored solution $\underline{X}_k - \underline{\mathcal{N}}$ is always worse than the parent $\underline{X}_k$ if $\underline{X}_k + \underline{\mathcal{N}}$ was better than $\underline{X}_k$, see Fig. 13.1 and Proposition 13.2.4 below. Setting *skip mirror* to true in Algorithm 7 prevents these mirrored samples from being realized.[2]

Note that in the $(1+\lambda_{\mathrm{m}})$-ES, two mirrored offspring are entirely dependent and, in a sense, complementary, similar to antithetic variables for Monte-Carlo numerical integration [173]. Mirrored sampling is also similar to using a symmetric difference quotient instead of the standard one-sided difference quotient. The technique has been applied to Evolutionary Gradient Search (EGS) with good success [19].

**Sequential selection:** In sequential selection, the offspring are evaluated one by one, compared to their parent, and the iteration is concluded immediately if one offspring is better than its parent. Sequential selection has been introduced in the context of comma selection, where it aims to combine the robustness advantage of comma selection with the speed advantage of elitist plus selection [85]. Sequential selection and mirrored sampling are independent of each other and can be employed separately within ESs, see Algorithm 7. We will see that sequential selection, in the elitist context, essentially comes down to the (1+1)-ES. The $(1+\lambda)$-ES variant employing sequential selection is denoted by $(1+\lambda^{\mathrm{s}})$-ES.

**Combining mirrored sampling and sequential selection:** Sequential selection has been combined

[2] In the $(1+1_{\mathrm{m}})$-ES, these unnecessary mirrored solutions fall closely together with the previous parental solution.

**Algorithm 7**: Pseudocode for the $(1+\lambda)$-ES and the $(1,\lambda)$-ES with all combinations with/without mirrored sampling and/or sequential selection. $\underline{X}_k \in \mathbb{R}^{\mathbb{N}}$ denotes the current search point and $\sigma_k$ the current step-size at iteration $k$. $(\underline{\mathcal{N}}_m)_{m\in\mathbb{N}}$ is a sequence of random vectors. In this report, *skip mirror* is true whenever sequential selection is true.

**given**: $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$, $\underline{X}_0 \in \mathbb{R}^{\mathbb{N}}$, $\sigma_0 > 0$, $\lambda \in \mathbb{N}^+$, $(\underline{\mathcal{N}}_m)_{m\in\mathbb{N}}$

$m \leftarrow 0$ number of random samples used
$j \leftarrow 0$ use previous sample if $j$ is even
$k \leftarrow 0$ iteration counter for notational consistency
**while** stopping criterion not fulfilled **do**
  $i \leftarrow 0$ offspring counter
  **while** $i < \lambda$ **do**
    $i \leftarrow i + 1$, $j \leftarrow j + 1$
    **if** *mirrored sampling* and $j \equiv 0 \pmod 2$ **then**
      $\underline{X}_k^i = \underline{X}_k - \sigma_k \underline{\mathcal{N}}_m$ use previous sample
    **else**
      $m \leftarrow m + 1$
      $\underline{X}_k^i = \underline{X}_k + \sigma_k \underline{\mathcal{N}}_m$
    **if** $f(\underline{X}_k^i) \leq f(\underline{X}_k)$ **then**
      **if** *skip mirror* **then**
        $j \leftarrow 0$ continue with a fresh sample
      **if** *sequential selection* **then**
        **break**
  **end while**
  **if** *plus selection* **then**
    $\underline{X}_{k+1} = \operatorname{argmin}\{f(\underline{X}_k), f(\underline{X}_k^1), \ldots, f(\underline{X}_k^i)\}$
  **else**
    $\underline{X}_{k+1} = \operatorname{argmin}\{f(\underline{X}_k^1), \ldots, f(\underline{X}_k^i)\}$
  $\sigma_{k+1} = \text{update}(\sigma_k)$
  $k \leftarrow k + 1$ iteration counter
**end while**

with mirrored sampling in the $(1,\lambda)$-ES and, although not explicitly mentioned, *skip mirror* was in this case always applied [85]. Both concepts complement each other well for the $(1,\lambda)$-ES. Sequential selection tends to reduce the realized population size to a minimum and mirrored sampling improves the performance in particular for very small population sizes. In this report as well, *skip mirror* is set to true when sequential selection and mirrored sampling are combined. With sequential selection, it is important that independent and mirrored offspring are evaluated alternately in order to profit immediately from the increased probability of the mirrored offspring being better after the independently drawn offspring was worse than the parent [85].

The $(1+\lambda)$-ES with mirrored sampling and sequential selection is denoted as $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES where the superscript refers to sequential selection and the subscript to mirrored sampling with *skip mirror* set to true. All results in this report refer to strategies where *skip mirror* is true when sequential selection is applied.

### 13.2.2 General Properties of Mirrored Sampling and Sequential Selection

In this section, we derive general results on evolution strategies using mirrored sampling and sequential selection. Let us first recognize that the $(1+\lambda^{\mathrm{s}})$-ES is essentially a (1+1)-ES with smaller iteration counter. In both strategies, the parent is updated if and only if the currently sampled offspring is better (but see also Remark 13.2.2). Now, we also establish for mirrored sampling that $(1+1_{\mathrm{ms}})$-ES and $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES

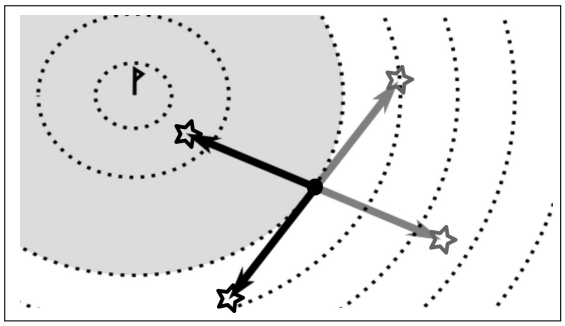


Figure 13.1: Mirrored sampling on an objective function with convex sub-level sets. The shaded region represents the set of solutions with a better objective function value than the parent solution (black dot). *Not both mirrored offspring can be better than the parent solution at the same time*: shown are two examples of an offspring (black) and its mirrored version (gray) where either one or both offspring are worse than the parent.

all evaluate the same points, provided they use the same (constant or scale-invariant) step-size and the same random instance for generating the offspring.

**13.2.1 Proposition**
*The* $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$*-ES is for any* $\lambda \geq 1$ *the same algorithm—with possibly different iteration counter, given the same random vectors and the same step-sizes are used (for example the step-size* $\sigma_k$ *is either constant or scale-invariant, i.e.,* $\sigma_k = \sigma\|X_k\|$ *with* $\sigma$ *constant).*

*Proof.* We prove that the state of the algorithm (apart from the iteration counter) does not depend on $\lambda$. Independently of $\lambda$, because of sequential selection applied in combination with plus selection, any new evaluated offspring is sampled from the *best ever* evaluated point so far. Since step-size only depends on the parent or is constant, the same offspring will be sampled provided the random samples used are also independent of $\lambda$. However, the samples used are taken one by one from $(\underline{\mathcal{N}}_m, -\underline{\mathcal{N}}_m)_{m\in\mathbb{N}}$ where because *skip mirror* is true, some mirrored vectors $-\underline{\mathcal{N}}$ are skipped. But the decision of whether or not to skip the mirroring of a sample is also independent of $\lambda$ since it only depends on a comparison between the single last offspring and the parent. □

Due to this result, the notations $(1+1_{\mathrm{ms}})$-ES, $(1+2^{\mathrm{s}}_{\mathrm{ms}})$-ES and $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES refer, in this report, all to the same strategy. However, this might not be the case in practice.

**13.2.2 Remark**
*In practice, the behavior of ESs with sequential selection depends on* $\lambda$*, because the step-size is typically updated at the end of each iteration and therefore more often with small* $\lambda$*.*

**13.2.3 Remark**
*For* $\mu = 1$*, sequential selection and/or mirroring have been combined with CMA-ES and extensively studied with plus and comma selection and different step-size rules [27, 28, 29, 30, 31, 32, 33, 35, 36, 37, 38].*

We derive now some results on objective functions with convex sub-level sets. We first establish that on objective functions with convex sub-level sets two mirrored offspring cannot be both better than their parent (see also Fig. 13.1).

**13.2.4 Proposition**
*Let* $f$ *be an objective function with convex sub-level sets, then two mirrored offspring cannot be simultaneously strictly better than their parent.*

*Proof.* Considering the convex sub-level set, given by the parent solution, and the tangent hyperplane in this solution, the two mirrored offspring can never lie on the same side of the tangent hyperplane, see Fig. 13.1. At the same time, due to the convexity of the sub-level set and the definition of the tangent hyperplane, all solutions that are better than the parent solution lie on the same side of the tangent hyperplane such that not both mirrored offspring can have better objective function values at the same time. □

A consequence of Proposition 13.2.4 is that on objective functions with convex sub-level sets, sequential selection applied with two mirrored offspring has no effect on the sequence of *accepted* solutions. In this case, sequential selection combined with skip mirror only reduces the number of evaluated solutions.

**13.2.5 Corollary (Identical trace for $\lambda = 2$)**
*On objective functions with convex sub-level sets, the* $(1+2_{\mathrm{m}})$*-ES and the* $(1+2^{\mathrm{s}}_{\mathrm{ms}})$*-ES deliver the same sequence of parental solutions (given they use the same random vectors and step-sizes). The same holds for the* $(1, 2_{\mathrm{m}})$*-ES and the* $(1, 2^{\mathrm{s}}_{\mathrm{ms}})$*-ES.*

*Proof.* We consider the iteration step $k$ and assume that $m = k$ at the beginning of the inner while loop. According to Proposition 13.2.4, it can never happen that both offspring are better than the parent and only two remaining cases need to be investigated. (i) In case of both offspring being worse than the parent, both plus-selection algorithms will keep the parent whereas the better of the two offspring is taken as the new parent by both comma-strategy algorithms. (ii) In case that one of the two offspring is not worse than the parent, the other must be worse and all algorithms will take the better offspring as their new parent solution—there is only a difference between the algorithms if the first offspring is not worse than the parent. Only in this case, the algorithms with sequential selection will directly accept the first offspring as next parent while the other variants evaluate unnecessarily the second (worse) offspring as well. Since either both offspring are evaluated or the sample associated to the non-evaluated offspring is skipped, in the next iteration, a fresh sample $\underline{N}_{m+1}$ will be used for the first offspring thus $m = k + 1$ at the beginning of the next inner while loop. □

Because sequential selection evaluates sometimes only one solution per iteration, the corollary implies that on functions with convex sub-level sets, the $(1+2^{\mathrm{s}}_{\mathrm{ms}})$-ES (or $(1, 2^{\mathrm{s}}_{\mathrm{ms}})$-ES) will converge faster than the $(1+2_{\mathrm{m}})$-ES (or $(1, 2_{\mathrm{m}})$-ES) in case of convergence and diverge faster in case of divergence.

We can additionally establish that for all strategies with two offspring and sequential selection, the number of offspring evaluated per iteration is the same, independent of mirroring and elitism:

**13.2.6 Lemma**
*Assume the* $(1 + 2^{\mathrm{s}})$, $(1, 2^{\mathrm{s}})$, $(1+2^{\mathrm{s}}_{\mathrm{ms}})$, $(1, 2^{\mathrm{s}}_{\mathrm{ms}})$*-ESs start at iteration $k$ from the same parent $\underline{X}_k$, sample the same first offspring $\underline{X}_k + \underline{N}$, and optimize the same objective function. Then the number of evaluated offspring at iteration $k$ will be the same for all strategies.*

*Proof.* In all the cases, the number of evaluated offspring will be 1 if $\underline{X}_k + \underline{N}$ is not worse than $\underline{X}_k$ and 2 otherwise. □

Finally, we find that for $\lambda$ to infinity, comma strategies using sequential selection without or with mirroring converge to the (1+1)-ES or the $(1+1_{\mathrm{ms}})$-ES, respectively:

**13.2.7 Lemma (Equivalence of $(1, \infty^{\mathrm{s}})$-ES and $(1+1)$-ES)**
*Using scale-invariant or constant step-size, the* $(1, \infty^{\mathrm{s}})$*-ES is equivalent to the* (1+1)*-ES and the* $(1, \infty^{\mathrm{s}}_{\mathrm{ms}})$*-ES is equivalent with the* $(1+1_{\mathrm{ms}})$*-ES.*

*Proof.* The proof follows directly from the algorithm descriptions, similar to the proof of Proposition 13.2.1. □

## 13.3 Linear Convergence and Lower Bounds

Evolution Strategies are rank-based search algorithms and as such cannot exhibit a faster convergence than linear [342]. We here define linear convergence as the logarithm of the distance to the optimum decreasing linearly with the increasing number of function evaluations. An example of linear convergence is illustrated in Fig. 13.2 for three different instances of the (1+1)-ES with scale-invariant step-size. Formally, for the (1+1)-ES, let $\underline{X}_k$ be the estimate of the solution at iteration $k$. Almost sure (a.s.) linear convergence takes place if there exists a constant $c \neq 0$, such that

$$\frac{1}{k} \ln \frac{\|\underline{X}_k\|}{\|\underline{X}_0\|} \to c \ \ a.s. \tag{13.1}$$

Literally, *convergence* of $\underline{X}_k$ takes place only if $c < 0$ and for $c > 0$ the term *divergence* is more appropriate. If the above expression goes to zero, the strategy might still converge sub-linearly [42]. In this report, we analyze algorithms that do not have a constant number of function evaluations per iteration and we will use the following generalization of (13.1) that accounts for the actual number of function evaluations: let $T_k$ be the number of function evaluations performed until iteration $k$. Almost sure linear convergence takes place if there exists a constant $c \neq 0$, such that

$$\frac{1}{T_k} \ln \frac{\|\underline{X}_k\|}{\|\underline{X}_0\|} \to c \ \ a.s. \tag{13.2}$$

For the (1+1)-ES both equations are equivalent and for the $(1 \overset{+}{,} \lambda)$-ES we have $T_k = k\lambda$. The constant $c$ is called the *convergence rate* and it corresponds to the slope of the curves in Fig. 13.2. The dynamics and thus the convergence rate of a step-size adaptive ES obviously depends on the step-size rule. The fastest convergence rates for adaptive step-size ESs are in general reached for a specific step-size rule in the so-called *scale-invariant step-size* ES where the step-size $\sigma_k$ at time $k$ is proportional to the distance to the optimum. Assuming the optimum w.l.o.g. in 0, the scale-invariant step-size is $\sigma_k = \sigma\|\underline{X}_k\|$ for $\sigma > 0$ on spherical functions $g(\|x\|)$ for $g \in \mathcal{M}$ where $\mathcal{M}$ denotes the set of functions $g : \mathbb{R} \mapsto \mathbb{R}$ that are strictly increasing [204]. ESs with scale-invariant step-sizes are artificial algorithms as they use the distance to the optimum to adapt the step-size. However, they are interesting to study as (1) they are a realistic approximation of step-size adaptive isotropic ESs where $\|\underline{X}_k\|/\sigma_k$ is usually a stable Markov Chain, here modeled as a constant, and (2) they achieve, for the right choice of the constant, optimal convergence rates. In addition, the simplification of $\|\underline{X}_k\|/\sigma_k$ being a constant induces in general much simpler theoretical analysis. We now state formally the linear convergence of a (1+1)-ES with scale-invariant step-size and give an implicit expression for the convergence rate:

**13.3.1 Theorem (Linear convergence of (1+1)-ES [204])**
*The (1+1)-ES with scale-invariant step-size ($\sigma_k = \sigma\|\underline{X}_k\|$) converges linearly on the class of spherical functions $g(\|x\|)$, $g \in \mathcal{M}$, and*

$$\lim_{k\to\infty} \frac{1}{k} \ln \frac{\|\underline{X}_k\|}{\|\underline{X}_0\|} = \mathrm{CR}_{(1+1)}(\sigma), \tag{13.3}$$

*with*

$$\mathrm{CR}_{(1+1)}(\sigma) = -\frac{1}{2} E\Big[ \ln^- \Big(1 + \underbrace{2\sigma[\underline{\mathcal{N}}]_1}_{\text{gain if negative}} + \underbrace{\sigma^2\|\underline{\mathcal{N}}\|^2}_{\text{loss}}\Big)\Big] \ ,$$

*where $\ln^-$ is the negative part of the function $\ln$, i.e., $\ln^-(x) = -\min(\ln(x), 0)$, $\underline{\mathcal{N}}$ is a standard multivariate normal distribution and $[\underline{\mathcal{N}}]_1$ is the projection of $\underline{\mathcal{N}}$ onto the first coordinate $\underline{e}_1$.*

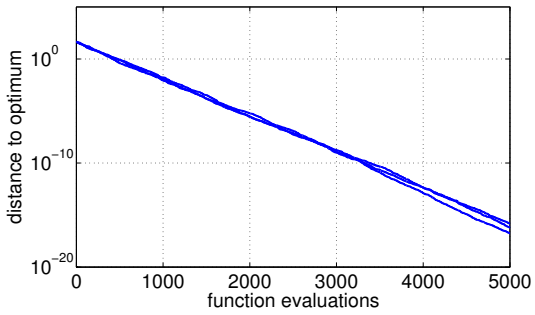



Figure 13.2: Distance to optimum versus number of function evaluations for three different instances of a (1+1)-ES minimizing the sphere function $g(\|x\|), g \in \mathcal{M}$ and with scale-invariant step-size $\sigma_k = \sigma\|\underline{X}_k\|$ for $\mathbb{N} = 20$ and $\sigma = 0.6/\mathbb{N}$. Linear decrease is observed, the convergence rate corresponds to the slope of the curves.

The function $\sigma \mapsto \mathrm{CR}_{(1+1)}(\sigma)$ gives for each $\sigma > 0$ the convergence rate of the (1+1)-ES with step-size $\sigma_k = \sigma\|\underline{X}_k\|$. The function has been studied in [204] and is plotted in Fig 13.4 (left) for different dimensions. The minimum of $\sigma \mapsto \mathrm{CR}_{(1+1)}(\sigma)$ gives for a given dimension the lower bound for the convergence rate of (1+1)-ES with offspring sampled with a standard multivariate normal distribution and *any step-size adaptation mechanism* on *any objective function* as formally stated now:

**13.3.2 Theorem (Lower Bound for (1+1)-ES [204])**
*Let $f : \mathbb{R}^{\mathbb{N}} \mapsto \mathbb{R}$ be a measurable objective function and $x^* \in \mathbb{R}^{\mathbb{N}}$. Assume that at each iteration $k$, the standard multivariate normal distribution used to sample the offspring is independent of $\sigma_k$ and $\underline{X}_k$ and that $E[|\ln\|\underline{X}_0 - x^*\|\|] < \infty$, then the convergence of the step-size adaptive (1+1)-ES is at most linear and*

$$\inf_{k\in\mathbb{N}} E[\ln\|\underline{X}_k - x^*\|/\|\underline{X}_0 - x^*\|] \geq \inf_{\sigma} \mathrm{CR}_{(1+1)}(\sigma) \ .$$

**Objective for the rest of the paper:** In the rest of the paper, we investigate the linear convergence of mirrored and sequential variants of the $(1+\lambda)$-ES with scale-invariant step-size. As for the (1+1)-ES, the minimum of the convergence rate in $\sigma$ will represent lower bounds for the convergence rate of step-size adaptive methods with a standard multivariate normal sampling on any objective function. Before tackling the linear convergence of the different variants, we explain the main proof idea behind the linear convergence proofs.

**How to prove linear convergence of scale-invariant step-size ESs?** We sketch the proof idea in the case of the (1+1)-ES and we will explain in the core of the paper how to generalize this proof in particular for the case of a non-constant number of evaluation per iteration. The first step of the proof expresses the left-hand side (LHS) of (13.1) as a sum of $k$ terms exploiting standard properties of the logarithm function:

$$\frac{1}{k}\ln\frac{\|\underline{X}_k\|}{\|\underline{X}_0\|} = \frac{1}{k}\sum_{i=0}^{k-1}\ln\frac{\|\underline{X}_{i+1}\|}{\|\underline{X}_i\|} \ . \tag{13.4}$$

We then exploit the isotropy of the sphere function, the isotropy of the standard multivariate normal distribution and the scale-invariant step-size rule to prove that all terms $\ln(\|\underline{X}_{i+1}\|/\|\underline{X}_i\|)$ are independent identically distributed (i.i.d.). A law of large numbers[3] (LLN) therefore implies that the right-hand side (RHS) of (13.4) converges when $k$ goes to infinity to $E[\ln(\|\underline{X}_{i+1}\|/\|\underline{X}_i\|)]$ almost surely.

[3]Verifying some technical conditions such that the expectation and the variance of $\ln(\|\underline{X}_{i+1}\|/\|\underline{X}_i\|)$ are finite.

## 13.4 Convergence Rate on Spherical functions in Finite Dimension

In this section, we analyze the linear convergence of the $(1+2_{\mathrm{m}})$-ES and the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES for a fixed dimension $\mathbb{N}$ of the search space. Before to establish the main results, we derive some technical results and introduce some useful definitions.

### 13.4.1 Preliminary Results and Definitions

We establish first a lemma that simplifies the writing of the acceptance event of mirrored offspring.

**13.4.1 Lemma**
*Let $\underline{X}_{\underline{e}_1} = \underline{e}_1 + \sigma\underline{\mathcal{N}}$ and $\underline{X}^{\mathrm{m}}_{\underline{e}_1} = \underline{e}_1 - \sigma\underline{\mathcal{N}}$ be two mirrored offspring sampled from the parent $\underline{e}_1 = (1, 0, \ldots, 0)$. On spherical functions, the acceptance event $\{\|\underline{e}_1 + \sigma\underline{\mathcal{N}}\| \leq 1\}$ can be written as $\{2[\underline{\mathcal{N}}]_1 + \sigma\|\underline{\mathcal{N}}\|^2 \leq 0\}$. Similarly, the acceptance event of $\underline{X}^{\mathrm{m}}_{\underline{e}_1}$ satisfies $\{\|\underline{e}_1 - \sigma\underline{\mathcal{N}}\| \leq 1\} = \{-2[\underline{\mathcal{N}}]_1 + \sigma\|\underline{\mathcal{N}}\|^2 \leq 0\}$.*

*Proof.* We remark first that $\|\underline{e}_1 + \sigma\underline{\mathcal{N}}\| \leq 1$ is equivalent to $\|\underline{e}_1 + \sigma\underline{\mathcal{N}}\|^2 \leq 1$. We now develop $\|\underline{e}_1 + \sigma\underline{\mathcal{N}}\|^2$ as $1 + 2\sigma[\underline{\mathcal{N}}]_1 + \sigma^2\|\underline{\mathcal{N}}\|^2$ and we immediately obtain that $1 + 2\sigma[\underline{\mathcal{N}}]_1 + \sigma^2\|\underline{\mathcal{N}}\|^2 \leq 1$ is equivalent to $2\sigma[\underline{\mathcal{N}}]_1 + \sigma^2\|\underline{\mathcal{N}}\|^2 \leq 0$. We proceed similarly for the acceptance event of $\underline{X}^{\mathrm{m}}_{\underline{e}_1}$. □

In the sequel, we will need to use the indicator function of the acceptance events of mirrored offspring sampled from $\underline{e}_1$. For that reason we define the random variables $W_1$ and $W_1^{\mathrm{m}}$ in the following way:

**13.4.2 Definition**
*Let $W_1 = 2[\underline{\mathcal{N}}]_1 + \sigma\|\underline{\mathcal{N}}\|^2$ and $W_1^{\mathrm{m}} = -2[\underline{\mathcal{N}}]_1 + \sigma\|\underline{\mathcal{N}}\|^2$.*

We can now express the indicator of the acceptance event of $\underline{X}_{\underline{e}_1}$ as

$$1_{\{\underline{X}_{\underline{e}_1} \text{ is better than } \underline{e}_1\}} = 1_{\{W_1 \leq 0\}} \ , \tag{13.5}$$

and the indicator of the acceptance of $\underline{X}^{\mathrm{m}}_{\underline{e}_1}$ as

$$1_{\{\underline{X}^{\mathrm{m}}_{\underline{e}_1} \text{is better than } \underline{e}_1\}} = 1_{\{W_1^{\mathrm{m}} \leq 0\}} \ . \tag{13.6}$$

Using the expression of $W_1$ and a straightforward derivation, we find the following alternative expression for the convergence rate of the (1+1)-ES:

$$\mathrm{CR}_{(1+1)}(\sigma) = \frac{1}{2} E\left[\ln(1 + \sigma W_1 1_{\{W_1 \leq 0\}})\right] \ . \tag{13.7}$$

We now establish two technical lemmas that will be useful to prove the equality of the convergence rate of the (1+1)-ES and the $(1+2_{\mathrm{m}})$-ES.

**13.4.3 Lemma**
*Let $\underline{\mathcal{N}}$ be a standard multivariate normal distribution, the following equality holds*

$$E\left[\ln(1 + (2\sigma[\underline{\mathcal{N}}]_1 + \sigma^2\|\underline{\mathcal{N}}\|^2)1_{\{2[\underline{\mathcal{N}}]_1 + \sigma\|\underline{\mathcal{N}}\|^2 \leq 0\}}\right] = E\left[\ln(1 + (-2\sigma[\underline{\mathcal{N}}]_1 + \sigma^2\|\underline{\mathcal{N}}\|^2)1_{\{-2[\underline{\mathcal{N}}]_1 + \sigma\|\underline{\mathcal{N}}\|^2 \leq 0\}}\right] \tag{13.8}$$

*or, using the notations $W_1$ and $W_1^{\mathrm{m}}$*

$$E\left[\ln(1 + \sigma W_1 1_{\{W_1 \leq 0\}})\right] = E\left[\ln(1 + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}} \leq 0\}})\right] \tag{13.9}$$

*Proof.* Since $\underline{\mathcal{N}}$ is a standard multivariate normal distribution, $-\underline{\mathcal{N}}$ follows the same distribution as $\underline{\mathcal{N}}$ and thus (13.8) follows. □

**13.4.4 Lemma**
*The following equation holds*

$$E\left[\ln\left(1+\sigma W_1 1_{\{W_1\leq 0\}}+\sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}}\leq 0\}}\right)\right] = 2E\left[\ln\left(1+\sigma W_1 1_{\{W_1\leq 0\}}\right)\right] \quad (13.10)$$

*where* $W_1 = 2[\underline{\mathcal{N}}]_1 + \sigma\|\underline{\mathcal{N}}\|^2$ *and* $W_1^{\mathrm{m}} = -2[\underline{\mathcal{N}}]_1 + \sigma\|\underline{\mathcal{N}}\|^2$ *with* $\underline{\mathcal{N}}$ *a random vector following a standard multivariate normal distribution.*

*Proof.* According to Proposition 13.2.4, two mirrored offspring cannot be simultaneously better than their parent on the sphere function. Since $\{W_1 \leq 0\}$ and $\{W_1^{\mathrm{m}} \leq 0\}$ are the acceptance events of mirrored offspring started from $\underline{e}_1$ on the sphere function ((13.5) and (13.6)), we know that they are incompatible such that $1_{\{W_1\leq 0\}}$ and $1_{\{W_1^{\mathrm{m}}\leq 0\}}$ are not simultaneously equal to 1. Consequently

$$\ln\left(1+\sigma W_1 1_{\{W_1\leq 0\}}+\sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}}\leq 0\}}\right) = \ln(1+\sigma W_1 1_{\{W_1\leq 0\}}) + \ln(1+\sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}}\leq 0\}}) \ .$$

Using the linearity of the expectation, we obtain that

$$E\left[\ln\left(1+\sigma W_1 1_{\{W_1\leq 0\}}+\sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}}\leq 0\}}\right)\right] = E\left[\ln(1+\sigma W_1 1_{\{W_1\leq 0\}})\right] + E\left[\ln(1+\sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}}\leq 0\}})\right] \ .$$

We now use Lemma 13.4.3 and obtain that the RHS of the last equation equals $2E\left[\ln(1+\sigma W_1 1_{\{W_1\leq 0\}})\right]$. Hence the result. □

## 13.4.2 Convergence Rate for the $(1 + 2_{\mathrm{m}})$-ES

In this section, we prove the linear convergence of the $(1+2_{\mathrm{m}})$-ES with scale-invariant step-size and prove the surprising result that the convergence rate of the $(1+2_{\mathrm{m}})$-ES equals the convergence rate of the $(1+1)$-ES. In a $(1+2_{\mathrm{m}})$-ES with scale-invariant step-size, two mirrored offspring $\underline{X}_k + \sigma\|\underline{X}_k\|\underline{\mathcal{N}}$ and $\underline{X}_k - \sigma\|\underline{X}_k\|\underline{\mathcal{N}}$ are sampled from the parent $\underline{X}_k$ where $\underline{\mathcal{N}}$ is a standard multivariate normal distribution independent of $\underline{X}_k$ and of the past (we omit the dependence in $k$ for the sampled vectors for the sake of readability). Since on the sphere function, the offspring cannot be simultaneously better than their parent (see Proposition 13.2.4), the update equation for $\|\underline{X}_k\|$ reads:

$$\begin{aligned}\|\underline{X}_{k+1}\| = {} & \|\underline{X}_k + \sigma\|\underline{X}_k\|\underline{\mathcal{N}}\| \times 1_{\{\|\underline{X}_k+\sigma\|\underline{X}_k\|\underline{\mathcal{N}}\|\leq\|\underline{X}_k\|\}} + \\ & \|\underline{X}_k - \sigma\|\underline{X}_k\|\underline{\mathcal{N}}\| \times 1_{\{\|\underline{X}_k-\sigma\|\underline{X}_k\|\underline{\mathcal{N}}\|\leq\|\underline{X}_k\|\}} + \\ & \|\underline{X}_k\| \times 1_{\{\|\underline{X}_k+\sigma\|\underline{X}_k\|\underline{\mathcal{N}}\|>\|\underline{X}_k\|,\|\underline{X}_k-\sigma\|\underline{X}_k\|\underline{\mathcal{N}}\|>\|\underline{X}_k\|\}} \ . \quad (13.11)\end{aligned}$$

Before to prove the linear convergence of the $(1+2_{\mathrm{m}})$-ES with scale-invariant step-size, we need to establish the following lemma:

**13.4.5 Lemma**
*Let* $Z_k$ *be the sequence of random variables*

$$Z_k = \frac{1}{2}\ln\Big[\|\underline{Y}_k + \sigma\underline{\mathcal{N}}\|^2 1_{\{\|\underline{Y}_k+\sigma\underline{\mathcal{N}}\|\leq 1\}} + \|\underline{Y}_k - \sigma\underline{\mathcal{N}}\|^2 1_{\{\|\underline{Y}_k-\sigma\underline{\mathcal{N}}\|\leq 1\}} + 1_{\{\|\underline{Y}_k+\sigma\underline{\mathcal{N}}\|>1,\|\underline{Y}_k-\sigma\underline{\mathcal{N}}\|>1\}}\Big]$$

*where* $\underline{Y}_k = \underline{X}_k/\|\underline{X}_k\|$ *with* $\underline{X}_k$ *defined with* (13.11). *Then* $Z_k$ *are independent identically distributed as*

$$\mathbb{Z}^{(1+2_{\mathrm{m}})} = \frac{1}{2}\ln\left[1 + \sigma W_1 1_{\{W_1\leq 0\}} + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}}\leq 0\}}\right] \ .$$

*Moreover* $E[|\mathbb{Z}^{(1+2_{\mathrm{m}})}|] < \infty$.

*Proof.* Because of the isotropy of the distribution of $\underline{\mathcal{N}}$ and of the sphere function, in distribution $Z_k$ equals

$$Z_k \stackrel{d}{=} \frac{1}{2}\ln\Big[\|\underline{e}_1 + \sigma\underline{\mathcal{N}}\|^2 1_{\{\|\underline{e}_1+\sigma\underline{\mathcal{N}}\|\leq 1\}} + \|\underline{e}_1 - \sigma\underline{\mathcal{N}}\|^2 1_{\{\|\underline{e}_1-\sigma\underline{\mathcal{N}}\|>1\}} + 1_{\{\|\underline{e}_1+\sigma\underline{\mathcal{N}}\|>1,\|\underline{e}_1-\sigma\underline{\mathcal{N}}\|>1\}}\Big] \quad (13.12)$$

where we have replaced $\underline{Y}_k$ by $\underline{e}_1$. The independence of $Z_k$ comes from the fact that $\underline{\mathcal{N}}$ is independent of $\underline{Y}_k$ and from the isotropy of the sphere. The detailed proof of those two points is rather technical and we refer to [34, Lemma 1 and Lemma 2] to see how to have a fully formal proof. We are now going to simplify the following term

$$\|\underline{e}_1 + \sigma\underline{\mathcal{N}}\|^2 1_{\{W_1\leq 0\}} + \|\underline{e}_1 - \sigma\underline{\mathcal{N}}\|^2 1_{\{W_1^{\mathrm{m}}\leq 0\}} + 1_{\{W_1>0,W_1^{\mathrm{m}}>0\}}$$

that comes into play in the RHS of (13.12). Developing $\|\underline{e}_1 + \sigma\underline{\mathcal{N}}\|^2$ as $1 + 2\sigma[\underline{\mathcal{N}}]_1 + \sigma^2\|\underline{\mathcal{N}}\|^2$ and $\|\underline{e}_1 - \sigma\underline{\mathcal{N}}\|^2$ as $1 - 2\sigma[\underline{\mathcal{N}}]_1 + \sigma^2\|\underline{\mathcal{N}}\|^2$, we can simplify the previous equation into

$$1_{\{W_1\leq 0\}} + \sigma W_1 1_{\{W_1\leq 0\}} + 1_{\{W_1^{\mathrm{m}}\leq 0\}} + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}}\leq 0\}} + 1_{\{W_1>0,W_1^{\mathrm{m}}>0\}} \ .$$

Since $1_{\{W_1\leq 0\}} + 1_{\{W_1^{\mathrm{m}}\leq 0\}} + 1_{\{W_1>0,W_1^{\mathrm{m}}>0\}} = 1$ we can simplify the previous equation into

$$1 + \sigma W_1 1_{\{W_1\leq 0\}} + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}}\leq 0\}} \ .$$

Injecting this in (13.12), we obtain the result. The proof of the fact that $E[|\mathbb{Z}^{(1+2_{\mathrm{m}})}|] < \infty$ comes from the proof of the integrability of $\ln[1 + W_1 1_{\{W_1\leq 0\}}]$ that has been shown in detail in [204]. □

We are now ready to prove the linear convergence of the $(1+2_{\mathrm{m}})$-ES and express its convergence rate as the expectation of the random variable $Z^{(1+2_{\mathrm{m}})}$ introduced in the previous lemma divided by 2.

**13.4.6 Theorem**
*For the* $(1+2_{\mathrm{m}})$*-ES with scale-invariant step-size on the class of spherical functions* $g(\|x\|)$, $g \in \mathcal{M}$, *linear convergence holds and*

$$\lim_{k\to\infty} \frac{1}{T_k}\ln\frac{\|\underline{X}_k\|}{\|\underline{X}_0\|} = \mathrm{CR}_{(1+2_{\mathrm{m}})}(\sigma) \quad (13.13)$$

*where*

$$\mathrm{CR}_{(1+2_{\mathrm{m}})}(\sigma) = \frac{1}{2}E[Z^{(1+2_{\mathrm{m}})}] = \frac{1}{4}E\left[\ln\left(1 + \sigma W_1 1_{\{W_1\leq 0\}} + \sigma W_1^{\mathrm{m}} 1_{\{W_1^{\mathrm{m}}\leq 0\}}\right)\right] \quad (13.14)$$

*where* $W_1 = 2[\underline{\mathcal{N}}]_1 + \sigma\|\underline{\mathcal{N}}\|^2$ *and* $W_1^{\mathrm{m}} = -2[\underline{\mathcal{N}}]_1 + \sigma\|\underline{\mathcal{N}}\|^2$ *with* $\underline{\mathcal{N}}$ *a random vector following a standard multivariate normal distribution.*

*Proof.* We start from (13.11), square it, normalize the equation by $\|\underline{X}_k\|$ and take the logarithm. We obtain

$$\frac{1}{2}\ln\frac{\|\underline{X}_{k+1}\|^2}{\|\underline{X}_k\|^2} = \frac{1}{2}\ln\Big[\|\underline{Y}_k + \sigma\underline{\mathcal{N}}\|^2 1_{\{\|\underline{Y}_k+\sigma\underline{\mathcal{N}}\|\le 1\}} + \|\underline{Y}_k - \sigma\underline{\mathcal{N}}\|^2 1_{\{\|\underline{Y}_k-\sigma\underline{\mathcal{N}}\|\le 1\}} + 1_{\{\|\underline{Y}_k+\sigma\underline{\mathcal{N}}\|>1,\|\underline{Y}_k-\sigma\underline{\mathcal{N}}\|>1\}}\Big]$$

where $\underline{Y}_k = \underline{X}_k/\|\underline{X}_k\|$. According to Lemma 13.4.5, by isotropy of the standard multivariate normal distribution, the random variables in the RHS of the previous equation are independent and identically distributed as

$$\mathbb{Z}^{(1+2_\mathrm{m})} = \frac{1}{2}\ln\Big[1 + \sigma W_1 1_{\{W_1\le 0\}} + \sigma W_1^\mathrm{m} 1_{\{W_1^\mathrm{m}\le 0\}}\Big]$$

In addition, by Lemma 13.4.5, $E[|\mathbb{Z}^{(1+2_\mathrm{m})}|] < \infty$, and we can thus apply the LLN for independent random variables to

$$\frac{1}{T_k}\ln\frac{\|\underline{X}_k\|}{\|\underline{X}_0\|} = \frac{1}{2k}\ln\frac{\|\underline{X}_k\|}{\|\underline{X}_0\|} = \frac{1}{4k}\sum_{i=0}^{k-1}\ln\frac{\|\underline{X}_{i+1}\|^2}{\|\underline{X}_i\|^2}$$

and we obtain (13.14). □

Putting together (13.7), Lemma 13.4.4 and the expression of the convergence rate of the $(1+2_\mathrm{m})$-ES found in the previous theorem, we immediately obtain that the $(1+1)$-ES and the $(1+2_\mathrm{m})$-ES converge at the same rate. This result is stated in the following corollary.

**13.4.7 Corollary**
*On the class of spherical functions, the* $(1+1)$*-ES and* $(1+2_\mathrm{m})$*-ES with scale-invariant step-size converge at the same convergence rate, i.e.*

$$\mathrm{CR}_{(1+1)}(\sigma) = \mathrm{CR}_{(1+2_\mathrm{m})}(\sigma) \text{ for all } \sigma \ .$$

We close this section with a geometrically based argumentation for the corollary. Consider the tangent hyperplane at the parent location that divides the space into two half spaces and only one of the half spaces contains better solutions. The $(1+1)$-ES samples isotropically into both half spaces integrating over the entire space. The $(1+2_\mathrm{m})$-ES samples one offspring into one half space and the second one into the other. Together, the offspring integrate over exactly the same region as the single offspring in the $(1+1)$-ES. The worse offspring is never successful, while the better offspring realizes twice the expected improvement of the offspring in the $(1+1)$-ES.

### 13.4.3 Convergence Rate for the $(1 + \lambda^\mathrm{s}_\mathrm{ms})$-ES

In this section, we analyze the convergence rate of the $(1+\lambda^\mathrm{s}_\mathrm{ms})$-ES. According to Proposition 13.2.1, for all $\lambda$, the $(1+\lambda^\mathrm{s}_\mathrm{ms})$-ES with scale-invariant step-size evaluate the same points in the search space provided they use the same independent random sequence $(\underline{\mathcal{N}}_m)_{m\in\mathbb{N}}$. Therefore, also the convergence rate of the $(1+\lambda^\mathrm{s}_\mathrm{ms})$-ES is independent of $\lambda$. Note that this is true because we investigate the convergence rate defined as log-progress *per function evaluation* and not per iteration. Though we could think that the easiest algorithm to analyze is the $(1+1_\mathrm{ms})$-ES, we investigate the $(1+2^\mathrm{s}_\mathrm{ms})$ for which iterations are independent—contrary to the $(1+1_\mathrm{ms})$—allowing thus to apply *directly* the LLN for *independent* random variables.

**13.4.8 Theorem**
*For the* $(1+\lambda^\mathrm{s}_\mathrm{ms})$*-ES with scale-invariant step-size on the class of spherical functions* $g(\|x\|)$*,* $g \in \mathcal{M}$*, linear convergence holds and for all* $\lambda$

$$\lim_{k\to\infty}\frac{1}{T_k}\ln\frac{\|\underline{X}_k\|}{\|\underline{X}_0\|} = \frac{2}{2-p_s(\sigma)}\,\mathrm{CR}_{(1+1)}(\sigma) \tag{13.15}$$

*where* $p_s(\sigma) = \Pr(2[\underline{\mathcal{N}}]_1 + \sigma\|\underline{\mathcal{N}}\|^2 \leq 0)$ *is the probability that the offspring* $\underline{X}_{\underline{e}_1} = \underline{e}_1 + \sigma\underline{\mathcal{N}}$ *is better than its parent* $\underline{e}_1$ *where* $\underline{\mathcal{N}}$ *is a standard multivariate normal distribution.*

*Proof.* We have seen in Proposition 13.2.1 that the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES with scale-invariant step-size evaluates the same points for all $\lambda$. Therefore for all $\lambda$, the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ESs with scale-invariant step-size have the same convergence rate. Let us analyze the $(1+2^{\mathrm{s}}_{\mathrm{ms}})$-ES. Let us write $\frac{1}{T_k}\ln\frac{\|\underline{X}_k\|}{\|\underline{X}_0\|}$ as $A_kB_k$ with $A_k = k/T_k$ and $B_k = \frac{1}{k}\ln(\|\underline{X}_k\|/\|\underline{X}_0\|)$. We are going to handle both terms separately. For $B_k$, we exploit Corollary 13.2.5 where we have seen that, starting from the same parent, the $(1+2_{\mathrm{m}})$-ES and $(1+2^{\mathrm{s}}_{\mathrm{ms}})$-ES have the same parent for the next iteration for objective functions with convex sub-level sets. Thus the sequence of parents $\underline{X}_k$ is the same for a $(1+2_{\mathrm{m}})$-ES and a $(1+2^{\mathrm{s}}_{\mathrm{ms}})$-ES and thus the expected relative improvement *per iteration* will be the same for both algorithms. By Corollary 13.4.7, we have that $B_k$ goes to $2\,\mathrm{CR}_{(1+1)}(\sigma)$ (the factor 2 comes from the normalization by evaluations for the convergence rate of the $(1+2_{\mathrm{m}})$-ES). For the term $A_k$, we denote by $\Lambda_i$ the number of offspring evaluated at iteration $i$. Then, $T_k = \Lambda_1 + \ldots + \Lambda_k$ and $1/A_k = \frac{1}{k}\sum_{i=1}^{k}\Lambda_i$. Similarly to [34, Lemma 8], the $\Lambda_k$ are independent and identically distributed. In addition, according to Lemma 13.2.6, the number of evaluated offspring for the $(1+2^{\mathrm{s}}_{\mathrm{ms}})$ is the same as for the $(1, 2^{\mathrm{s}})$, we can therefore use the result shown in [29, Lemma 8] and obtain that $1/A_k$ converges almost surely to $2 - p_s(\sigma)$.

Therefore $A_k$ times $B_k$ converges to

$$\frac{2}{2 - p_s(\sigma)}\,\mathrm{CR}_{(1+1)}(\sigma) \quad .\square$$

□

We see in (13.15) that the convergence rate of the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES is expressed as the product of the convergence rate of the (1+1)-ES times $2/(2-p_s(\sigma))$. The term $2/(2-p_s(\sigma))$—which is always larger or equal one—is the gain brought by sequential selection. Indeed, as sketched in the proof of the theorem, the gain brought by sequential selection in strategies with two offspring (with mirrored or non-mirrored sampling) always equals $2/(2 - p_s(\sigma))$.

We give a useful expression for the success probability $p_s(\sigma)$ for a single offspring on the sphere function.

**13.4.9 Lemma**
*For all* $\sigma > 0$*, we have*

$$p_s(\sigma) = \Pr\left([\underline{\mathcal{N}}]_1 \leq -\frac{d}{2}\sigma \underbrace{\frac{\|\underline{\mathcal{N}}\|^2}{d}}_{\text{close to } 1}\right) \tag{13.16}$$

*Proof.* The lemma follows from the definition of $p_s(\sigma) = \Pr\left(2[\underline{\mathcal{N}}]_1 + \sigma\|\underline{\mathcal{N}}\|^2 \leq 0\right)$ □

The expression suggests that $\sigma \propto 1/d$ achieves a fairly $d$-independent success probability. A typical, close to optimal value is $\sigma \approx 1.2/d$ with $p_s \approx 1/4$ and $2/(2 - p_s) \approx 1.16$.

Finally, we can give the upper bound for the speed-up brought by sequential selection, when $\lambda = 2$.

**13.4.10 Corollary (Speed-up for $\lambda = 2$)**
*The upper bound for the speed-up brought by sequential selection for* $\lambda = 2$ *is given by*

$$\frac{2}{2 - p_s} < \frac{4}{3} = 1.333\ldots \tag{13.17}$$

*Proof.* From Lemma 13.4.9 we find for $\sigma > 0$ that $p_s < \Pr([\underline{\mathcal{N}}]_1 \leq 0) = 1/2$ which implies (13.17). For $\sigma = 0$ we have no speed-up. □

This upper bound holds equally well for savings by sequential selection whether or not *skip mirror* is applied.

## 13.5 Asymptotic Convergence Rates

So far, we have proven the linear convergence of some scale-invariant step-size ESs *for a fixed dimension of the search space*. In this section, we want to study how the finite dimension convergence rates derived previously behave when the dimension goes to infinity. We have observed that the convergence rate of the $(1+1_{\text{ms}})$-ES is a function of the convergence rate of the $(1+1)$-ES and of the probability of success $p_s$. We therefore study those two quantities asymptotically in order to obtain the asymptotic behavior of $\text{CR}_{(1+1_{\text{ms}})}$. Both asymptotic estimates were already (less rigorously) derived in [315].

### 13.5.1 Asymptotic Probability of Success

We first derive the limit of the probability of success $p_s(\sigma/d)$ when $d$ goes to infinity.

**13.5.1 Lemma**
*For all $\sigma > 0$*

$$\begin{aligned}\lim_{\mathbb{N}\to\infty} p_s\left(\frac{\sigma}{d}\right) &= \Pr([\underline{\mathcal{N}}]_1 \leq -\sigma/2) \\ &= \Phi\left(-\frac{\sigma}{2}\right) \end{aligned} \tag{13.18}$$

*where $\Phi$ is the cumulative distribution of a standard normal distribution, i.e. $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$ or, with the error function* erf, $\Phi(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$.

*Proof.* We start from the expression of $p_s(\sigma/d)$:

$$p_s(\sigma/d) = \Pr\left(2[\underline{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\underline{\mathcal{N}}\|^2 \leq 0\right) \tag{13.19}$$

$$= E\left[1_{\{2[\underline{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\underline{\mathcal{N}}\|^2 \leq 0\}}\right] \tag{13.20}$$

From the LLN, we know that

$$\lim_{\mathbb{N}\to\infty} \frac{1}{\mathbb{N}}\|\underline{\mathcal{N}}\|^2 = \lim_{\mathbb{N}\to\infty} \frac{1}{\mathbb{N}} \sum_{i=1}^{d} \mathcal{N}_i^2 = 1$$

almost surely, where $\mathcal{N}_i$ are i.i.d. standard normal distributions that are the coordinates of the vector $\underline{\mathcal{N}}$. Thus

$$2[\underline{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\underline{\mathcal{N}}\|^2 \xrightarrow[\mathbb{N}\to\infty]{} 2[\underline{\mathcal{N}}]_1 + \sigma$$

and therefore we have that

$$1_{\{2[\underline{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\underline{\mathcal{N}}\|^2 \leq 0\}} \xrightarrow[\mathbb{N}\to\infty]{} 1_{\{2[\underline{\mathcal{N}}]_1 + \sigma \leq 0\}} a.s.$$

Since $1_{\{2[\underline{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\underline{\mathcal{N}}\|^2\}} \leq 1$, we can apply the Lebesgue dominated convergence theorem that implies that

$$E\left[1_{\{2[\underline{\mathcal{N}}]_1 + \frac{\sigma}{d}\|\underline{\mathcal{N}}\|^2 \leq 0\}}\right] \xrightarrow[\mathbb{N}\to\infty]{} E\left[1_{\{2[\underline{\mathcal{N}}]_1 + \sigma \leq 0\}}\right] \quad .$$

We can rewrite the RHS of the last equation as

$$E\left[1_{\{2[\underline{\mathcal{N}}]_1+\sigma\leq 0\}}\right] = \Pr(2[\underline{\mathcal{N}}]_1 + \sigma \leq 0) = \Pr([\underline{\mathcal{N}}]_1 \leq -\sigma/2) \ .$$

Moreover, $\Pr([\underline{\mathcal{N}}]_1 \leq -\sigma/2) = \Phi(-\sigma/2)$. □

### 13.5.2 Asymptotic Convergence Rate of the (1+1)-ES

We will derive now the asymptotic convergence rate of the (1+1)-ES with scale-invariant step-size and find that it coincides with the negative of the well-known progress rate of the (1+1)-ES [315]. We first need to derive the following technical lemma:

**13.5.2 Lemma**
*Let $\mathcal{N}$ be a standard normal distribution, the following equation holds*

$$E[\mathcal{N}1_{\{\mathcal{N}\leq -\sigma/2\}}] = -\frac{1}{\sqrt{2\pi}}\exp(-\frac{\sigma^2}{8}) \ , \tag{13.21}$$

*for all $\sigma > 0$.*

*Proof.* In a first step we write the LHS of (13.21) using the density of a normal distribution

$$E[\mathcal{N}1_{\{\mathcal{N}\leq -\sigma/2\}}] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{-\sigma/2} x\exp(-\frac{x^2}{2})dx \ . \tag{13.22}$$

By integrating the RHS of (13.22) we obtain the result. □

We are now ready to derive the limit of the convergence rate of the (1+1)-ES.

**13.5.3 Theorem**
*Let $\sigma > 0$, the convergence rate of the (1+1)-ES with scale-invariant step-size on spherical functions satisfies at the limit*

$$\lim_{\mathbb{N}\to\infty} d \times \mathrm{CR}_{(1+1)}\left(\frac{\sigma}{d}\right) = \frac{-\sigma}{\sqrt{2\pi}}\exp\left(-\frac{\sigma^2}{8}\right) + \frac{\sigma^2}{2}\Phi\left(-\frac{\sigma}{2}\right) \tag{13.23}$$

*where $\Phi$ is the cumulative distribution of a normal distribution.*

*Proof.* We are going to investigate the almost sure limit of the random variable inside the RHS of

$$CR_{(1+1)}(\sigma/d) = \frac{1}{2}E\left[\ln(1 + \frac{\sigma}{d}\min(2[\underline{\mathcal{N}}]_1 + \frac{\sigma}{d}||\underline{\mathcal{N}}||^2, 0))\right] \ . \tag{13.24}$$

The following equation holds almost surely

$$\lim_{\mathbb{N}\to\infty} d \times \frac{1}{2}\ln(1 + \sigma/d\min(2[\underline{\mathcal{N}}]_1 + \frac{\sigma}{d}||\underline{\mathcal{N}}||^2, 0))$$
$$\xrightarrow[\mathbb{N}\to\infty]{} \frac{1}{2}\sigma\min(2[\underline{\mathcal{N}}]_1 + \sigma, 0) \ . \tag{13.25}$$

Assuming the uniform integrability of

$$d \times \frac{1}{2}\ln(1 + \sigma/d\min(2[\underline{\mathcal{N}}]_1 + \frac{\sigma}{d}||\underline{\mathcal{N}}||^2, 0)) \ ,$$

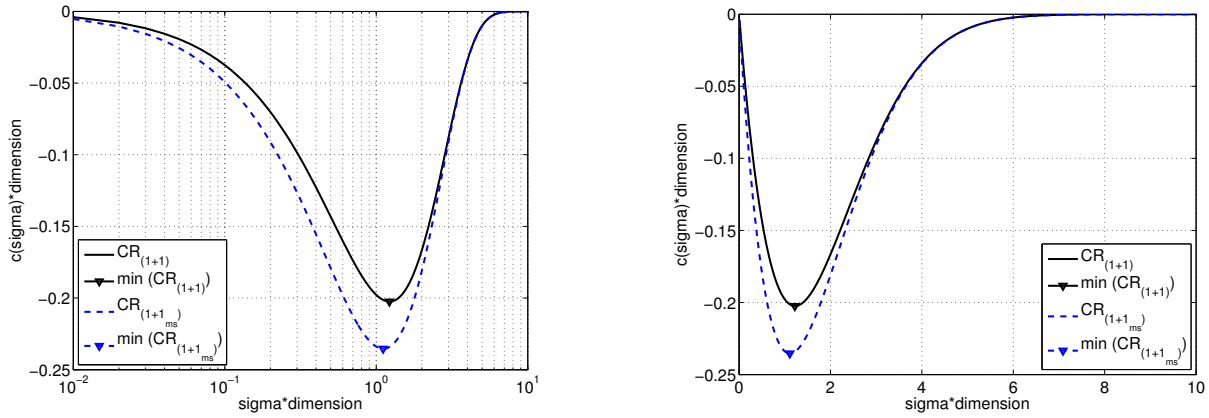



Figure 13.3: Theoretical limit results of the convergence rate for the (1+1)-ES (solid line) and for the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES (any $\lambda \geq 1$, dashed line) if $d$ goes to infinity. Left: versus $\sigma \cdot d$ in log scale; right: versus $\sigma \cdot d$ in linear scale.

we deduce that

$$\lim_{\mathbb{N}\to\infty} d \times \mathrm{CR}_{(1+1)}(\sigma/d) = \frac{\sigma}{2} E[\min(2[\underline{\mathcal{N}}]_1 + \sigma, 0)] \ .$$

Moreover,

$$\begin{aligned} E[\min(2[\underline{\mathcal{N}}]_1 + \sigma, 0)] &= E[(2[\underline{\mathcal{N}}]_1 + \sigma) 1_{\{2[\underline{\mathcal{N}}]_1 + \sigma \leq 0\}}] \\ &= 2E[([\underline{\mathcal{N}}]_1) 1_{\{2[\underline{\mathcal{N}}]_1 + \sigma \leq 0\}}] \\ &\quad + \sigma \Pr(2[\underline{\mathcal{N}}]_1 + \sigma \leq 0) \\ &= 2E[([\underline{\mathcal{N}}]_1) 1_{\{[\underline{\mathcal{N}}]_1 \leq -\sigma/2\}}] \\ &\quad + \sigma \Pr([\underline{\mathcal{N}}]_1 \leq -\sigma/2) \ . \end{aligned}$$

Thus,

$$\lim_{\mathbb{N}\to\infty} d \times \mathrm{CR}_{(1+1)}\left(\frac{\sigma}{d}\right) = \sigma E[[\underline{\mathcal{N}}]_1 1_{\{[\underline{\mathcal{N}}]_1 \leq -\frac{\sigma}{2}\}}] + \frac{\sigma^2}{2} \Phi\left(\frac{-\sigma}{2}\right) \ .$$

Using now Lemma 13.5.2, we obtain the result. □

This limit of the normalized convergence rate of the (1+1)-ES found in the previous theorem equals to the negated progress rate of the (1+1)-ES on the sphere function [315].

### 13.5.3 Deriving the Asymptotic Convergence Rate of the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES

We can now combine Lemma 13.5.1 and Theorem 13.5.3 to derive the asymptotic convergence rate of the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES with scale-invariant step-size. Note again that the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES is here the same as the $(1+1_{\mathrm{ms}})$-ES.

#### 13.5.4 Theorem

*Let $\sigma > 0$, the convergence rate of the $(1+1_{\mathrm{ms}})$-ES with scale-invariant step-size on spherical functions*

*satisfies*

$$\lim_{\mathbb{N}\to\infty} d \times \mathrm{CR}_{(1+1_{\mathrm{ms}})}\left(\frac{\sigma}{d}\right) = \frac{2}{2-\Phi(-\sigma/2)} \times \left(\frac{-\sigma}{\sqrt{2\pi}}\exp(-\frac{\sigma^2}{8}) + \frac{\sigma^2}{2}\Phi\left(-\frac{\sigma}{2}\right)\right) \quad . \quad (13.26)$$

*Proof.* Since the convergence rate of the $(1+1_{\mathrm{ms}})$-ES equals the convergence rate of the $(1+1)$-ES times $2/(2-p_s)$ we have that the limit for $\mathbb{N}$ to infinity satisfies

$$\lim_{\mathbb{N}\to\infty} d \times \mathrm{CR}_{(1+1_{\mathrm{ms}})}\left(\frac{\sigma}{d}\right) = \lim_{\mathbb{N}\to\infty}\left(\frac{2}{2-p_s}\right) \times \lim_{\mathbb{N}\to\infty}\left(d \times \mathrm{CR}_{(1+1)}\left(\frac{\sigma}{d}\right)\right) \quad .$$

Using Lemma 13.5.1 for the limit of $2/(2-p_s)$ and Proposition 13.5.3, we obtain the result. □

Figure 13.3 represents the limit of the normalized convergence rates of the $(1+1)$-ES and the $(1+1_{\mathrm{ms}})$-ES. The minimal value of the convergence rate of the $(1+1)$-ES and $(1+1_{\mathrm{ms}})$-ES respectively equal approximately $-0.202$ and $-0.235$. Mirrored sampling and sequential selection speed up the fastest single-parent evolution strategy asymptotically by a factor of about 1.16.

## 13.6 Numerical simulation of convergence rates

To conclude on the improvements that can be brought by mirrored samples and sequential selection, we now compare the different convergence rates. However, those convergence rates are expressed only implicitly as the expectation of some random variables. We therefore simulate the convergence rate with a Monte-Carlo technique. For each convergence rate expression, we have simulated $10^7$ times the random variables inside the expectation and averaged to obtain an estimate of the expectation and therefore of the convergence rate for different $\sigma$. Here, $\sigma$ has been chosen such that $0.01 \leq \sigma \cdot \mathbb{N} \leq 10$ and with steps of 0.01 in $\sigma \cdot \mathbb{N}$. The minimum of the measured convergence rates over $\sigma \cdot \mathbb{N}$ is used as estimate of the *best* convergence rate for each algorithm and dimension—resulting in a slightly (systematically) smaller value than the true one, due to taking the minimal value from several random estimates.

The plots of **Fig. 13.4** show the resulting convergence rate estimates versus $\sigma$ in several dimensions. Overall, mirroring and/or sequential selection do not essentially change the picture. The $(1+1)$-ES realizes the largest optimal step-size of all variants, also compared with comma selection (not shown). **Figure 13.5 shows the relative improvement. For small step-sizes the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES is up to about 33% faster (compare** (13.17)**). For large step-sizes, both,** $(1+1)$**- and $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES, show very similar convergence rates. For close to optimal step-sizes (somewhat above one), the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES is about 15% to 20% faster.**

**Figure 13.6** presents the estimated best convergence rates for several algorithms versus dimension. Here, the $(1, 4^{\mathrm{s}}_{\mathrm{ms}})$-ES is shown additionally.[4] The convergence rate of the $(1, \lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES is monotonically increasing in $\lambda$ (not shown) and in the limit for $\lambda \to \infty$, the $(1, \lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES coincides with the $(1+1_{\mathrm{ms}})$-ES. In small dimension, already for $\lambda = 4$ the convergence rate is very close to the convergence rate of the $(1+1_{\mathrm{ms}})$-ES. In all cases, the convergence rate of the $(1, 4^{\mathrm{s}}_{\mathrm{ms}})$-ES is closer to the $(1+1_{\mathrm{ms}})$-ES than to the $(1+1)$-ES. The difference between the original $(1+1)$-ES and the $(1+1_{\mathrm{ms}})$-ES is roughly between 15 and 20%. In dimension 320, the values are very close to the limit value.

[4] Note that in previous publications such as [27, 28, 85], the slightly different notation $(1, 4^{\mathrm{s}}_{\mathrm{m}})$-ES was used for the same algorithm.

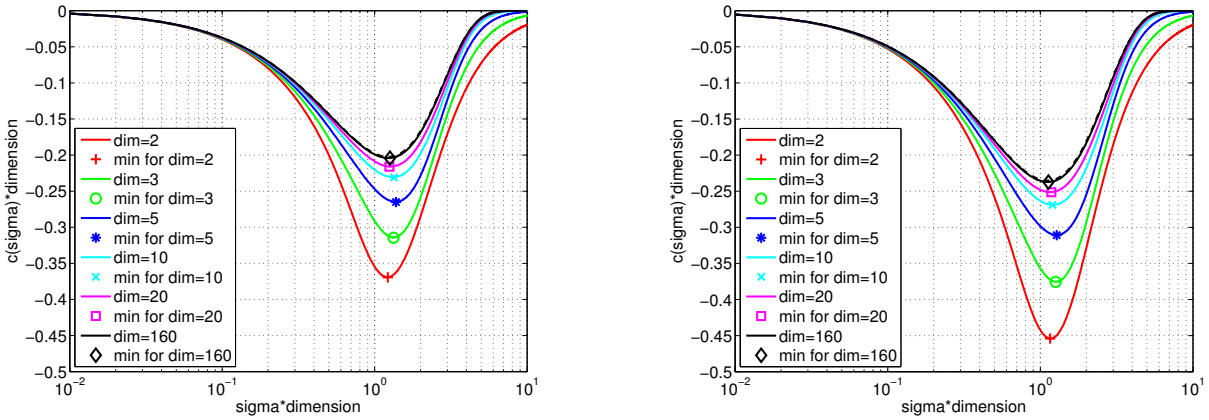



Figure 13.4: Convergence rate $c(\sigma)$ for different dimensions $d$ and the (1+1)-ES and (1+2$_\mathrm{m}$)-ES (both have the same convergence rate, left figure), and the (1+$\lambda^\mathrm{s}_\mathrm{ms}$)-ES (the same for all $\lambda \geq 1$, right figure), all with scale-invariant step-size. The dashed (uppermost) line shows the limit result for $d$ to infinity.

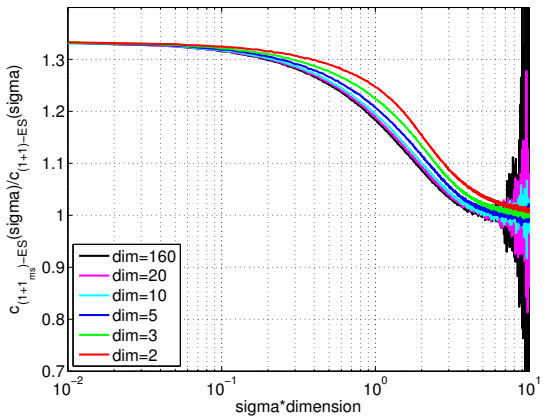



Figure 13.5: Relative improvement in the convergence rates $c(\sigma)$ of the (1+1$_\mathrm{ms}$)-ES over the (1+1)-ES plotted versus $\sigma$ times dimension for scale-invariant step sizes. Smaller dimensions show the (slightly) larger improvements. The huge fluctuations to the right are due to the small success probability for large step-sizes and therefore a large variance when measuring a few events.

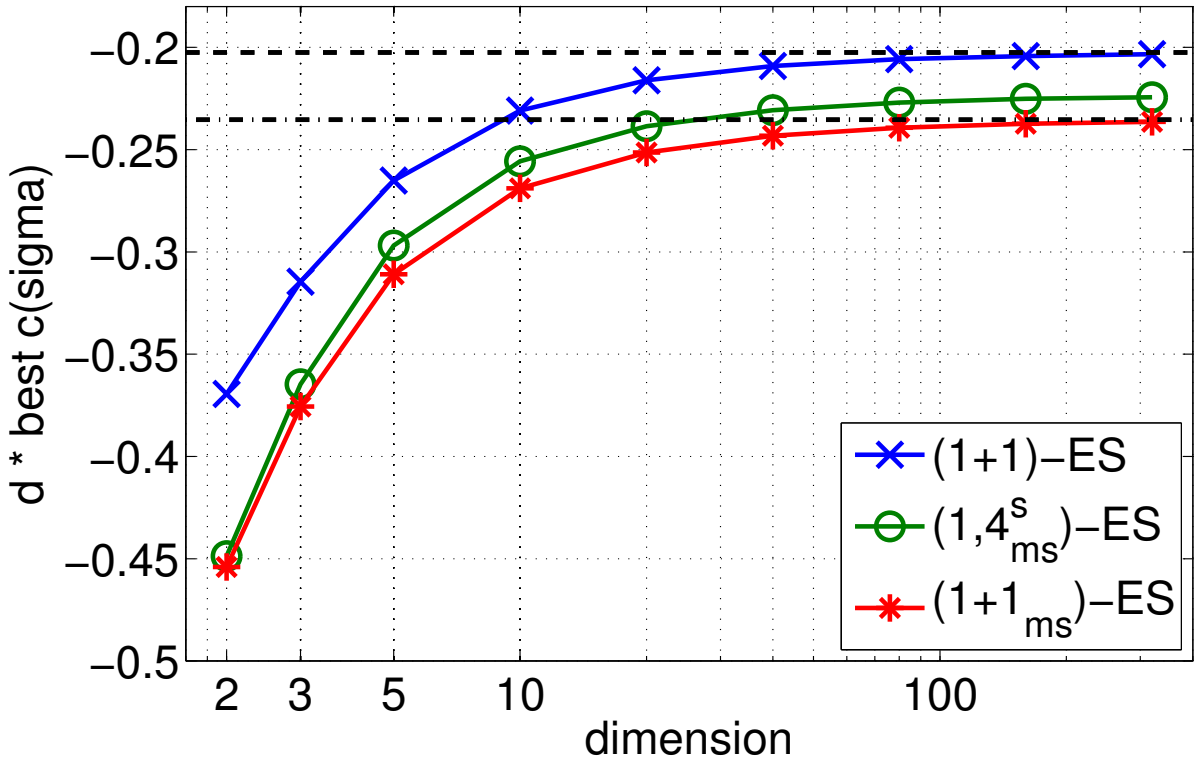



Figure 13.6: Estimated optimal convergence rates, in parts extracted from Fig. 13.4, multiplied by the dimension and plotted versus dimension for the algorithms (1+1)-ES (equivalent to (1+2$_m$)-ES), (1+1$_{ms}$)-ES, and (1, 4$^s_{ms}$)-ES with scale-invariant step-size. In addition, the theoretical limit results for $d$ to infinity are shown for the (1+1)-ES (dashed) and for the (1+1$_{ms}$)-ES (dotted-dashed).

## 13.7 Discussion

In this report we have analyzed the $(1+\lambda)$-ES with mirrored sampling and/or sequential selection. With sequential (plus) selection, the parameter $\lambda$ loses most of its meaning. Given that the step-size (and all other variation parameters) are updated in an identical way, the $(1+\lambda^{\mathrm{s}})$-ES, where $^{\mathrm{s}}$ denotes sequential selection, and also the $(1,\infty^{\mathrm{s}})$-ES depict the same strategy for all $\lambda \geq 1$. The same holds analogously for the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ESs, where the subscript $_{\mathrm{ms}}$ denotes *m*irrored sampling with *s*kip mirror applied (on success).

We have obtained tight lower bounds for the convergence rate of the $(1+2_{\mathrm{m}})$-ES and of the $(1+1_{\mathrm{ms}})$-ES that coincides with the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ESs for any $\lambda \geq 1$. These bounds are also the convergence rate with scale-invariant optimal step-size on the sphere function. The $(1+2_{\mathrm{m}})$-ES has the same convergence rate as the (1+1)-ES, asymptotically with the dimension to $\infty$ being $\geq -0.202\ldots$ The asymptotic convergence rate of the $(1+1_{\mathrm{ms}})$-ES is $\geq -0.235\ldots$ and the relation

$$\begin{aligned}
\mathrm{CR}_{(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})}(\sigma) &= \mathrm{CR}_{(1+1_{\mathrm{ms}})}(\sigma) && (13.27)\\
&= \frac{2}{2-p_s(\sigma)}\,\mathrm{CR}_{(1+2_{\mathrm{m}})}(\sigma) && (13.28)\\
&= \frac{2}{2-p_s(\sigma)}\,\mathrm{CR}_{(1+1)}(\sigma) && (13.29)
\end{aligned}$$

holds, where $p_s(\sigma)$ is the probability that an offspring, sampled isotropically with step-size $\sigma$, is better than its parent. The factor $2/(2-p_s(\sigma)) < 4/3$ is the improvement brought by sequential selection for $\lambda = 2$, with plus as well as comma selection.

As to our knowledge, the $(1+\lambda^{\mathrm{s}}_{\mathrm{ms}})$-ES is now the single-parent evolution strategy with the fastest known convergence rate, more than 15% faster than the (1+1)-ES, but no more than 5% faster than the $(1,4^{\mathrm{s}}_{\mathrm{ms}})$-ES. Only strategies with weighted recombination can exhibit even faster convergence rates (also denoted as serial efficiencies), namely $\geq -0.25$ when positive recombination weights are used [20].

The convergence rates derived assume that the step-size equals a constant times the distance to the optimum. This assumption simplifies the linear convergence derivation as the law of large numbers for independent random variables can then be used. For real adaptation schemes however, the analysis on spherical functions is in general more complicated, as $\sigma_k/\|\underline{X}_k\|$ is not a constant but a Markov chain. Law of large numbers for Markov chains can be used to prove linear convergence, the difficult task being to prove that $\sigma_k/\|\underline{X}_k\|$ is stable enough to satisfy a LLN [23, 42].

# Chapter 14

# Mirrored Sampling in Evolution Strategies With Weighted Recombination

Anne Auger, Dimo Brockhoff, Nikolaus Hansen

*Accepted at Genetic and Evolutionary Computation Conference (GECCO) 2011*

*This report introduces mirrored sampling into evolution strategies with weighted multi-recombination.* Pairwise selection *selects at most one of two mirrored vectors in order to avoid a bias due to recombination.* Selective mirroring *only mirrors the originally worst solutions of the population. Convergence rates on the sphere function are derived also yielding lower bounds. The optimal ratio of mirrored offspring is 1/2 (maximal) for randomly selected mirrors and about 1/6 for selective mirroring, where the convergence rate reaches a value of* 0.390*. This is an improvement of more than* 50% *compared to the best known convergence rate of* 0.25 *with positive recombination weights. Selective mirroring is combined with CMA-ES and benchmarked on unimodal functions and on the COCO/BBOB-2010 testbed.*

## 14.1 Introduction

Derandomization of random numbers is a general technique where independent samples are replaced by dependent ones. Recent studies showed, for example, how derandomization via *mirrored sampling* can improve $(1, \lambda)$- and $(1 + \lambda)$-ES [85, 39]. Instead of generating $\lambda$ independent and identically distributed (i.i.d.) search points in iteration $k$ as $\underline{X}_k + \sigma_k \underline{\boldsymbol{\mathcal{N}}}^i$ where $\underline{X}_k$ is the current search point, $\sigma_k$ the current step size, and $\underline{\boldsymbol{\mathcal{N}}}^i$ a random sample from a multivariate normal distribution, the $(1 \overset{+}{,} \lambda)$-ES with mirrored sampling always pairs samples one by one and produces $\lambda/2$ independent search points $\underline{X}_k + \sigma_k \underline{\boldsymbol{\mathcal{N}}}^i$ and $\lambda/2$ dependent ones as $\underline{X}_k - \sigma_k \underline{\boldsymbol{\mathcal{N}}}^i$ $(1 \leq i \leq \lambda/2)$. In the end, the best out of these $\lambda$ search points is used as next search point $\underline{X}_{k+1}$ in the $(1, \lambda)$-ES (and in the $(1+\lambda)$-ES the best out of the $\lambda$ new *and* the old $\underline{X}_k$). Several ES variants using mirrored mutations showed noticeable improvements over their unmirrored counterparts—not only in theoretical investigations on simple test functions such as the sphere function, but also in exhaustive experiments within the COCO/BBOB framework [85, 39]. Up to now, the results were restricted to single-parent $(1 \overset{+}{,} \lambda)$-ESs though the idea is, in principle, applicable in a straight-forward manner to population-based ESs such as the $(\mu/\mu_w, \lambda)$-ES where the $\mu$ best out of the $\lambda$ offspring are used to compute the new search point $\underline{X}_{k+1}$ via weighted recombination. However,

the direct application of mirrored mutations in population-based ES, as for example proposed in a more general way by Teytaud et al. [344], results in an undesired bias on the step size, as was argued already in [85].

The purpose of this report is to introduce mirrored mutations into ESs with weighted recombination *without* introducing a bias on the length of the recombined step. The main idea hereby is *pairwise selection* that allows only the better solution of a mirrored/unmirrored solution pair to possibly contribute to the weighted recombination. In detail, the contributions of this report are

- the introduction of several ES variants that combine mirrored mutations and weighted recombination without a bias on the recombined step,
- a theoretical investigation of the algorithms' convergence rates (in finite and infinite dimension) on spherical functions,
- the computation and estimation of optimal recombination weights,
- experimental results comparing estimated convergence rates with only positive recombination weights and in particular evaluating the impact of mirrored mutations, and
- numerical results on the performance of the $(\mu/\mu_\mathrm{w}, \lambda)$-CMA-ES [180] combined with mirrored mutations on unimodal functions and on the COCO/BBOB-2010 test bed.

The paper is organized as follows. After introducing the baseline $(\mu/\mu_\mathrm{w}, \lambda)$-ES, Sec. 14.2 explains in detail how mirrored mutations can be introduced in this algorithm. Section 14.3 theoretically investigates the convergence rate of three variants in finite and infinite dimension. Section 14.4 presents a comparison of the algorithms based on the numerical estimation of their convergence rates on the sphere function. Section 14.5 applies the mirroring idea to the $(\mu/\mu_\mathrm{w}, \lambda)$-CMA-ES and shows performance results on unimodal functions and the COCO/BBOB-2010 test function suite. Section 14.6 summarizes and concludes the paper.

**Notations** For a (random) vector $\underline{x} \in \mathbb{R}^n$, $[\underline{x}]_1$ will denote its first coordinate. The vector $(1, 0, \ldots, 0)$ will be denoted $\underline{e}_1$. A random variable following a multivariate normal distribution with mean vector zero and covariance matrix identity will be called *standard* multivariate normal distribution.

## 14.2 Mirroring and Weighted Recombination

### 14.2.1 The Standard $(\mu/\mu_\mathrm{w}, \lambda)$-ES

As our baseline algorithm, we briefly recapitulate the standard $(\mu/\mu_\mathrm{w}, \lambda)$-ES with weighted recombination and show its pseudocode in Algorithm 8. Given a starting point $\underline{X}_0 \in \mathbb{R}^\mathbb{N}$, an initial step size $\sigma_0 > 0$, a population size $\lambda \in \mathbb{N}^+$, and weights $\underline{w} \in \mathbb{R}^\mu$ with $\sum_{i=1}^{\mu} |w_i| = 1$ for a chosen $1 \leq \mu \leq \lambda$, the $(\mu/\mu_\mathrm{w}, \lambda)$-ES generates at iteration $k$ $\lambda$ independent search points from a multivariate normal distribution with mean $\underline{X}_k$ and variance $\sigma_k^2$ and recombines the best $\mu$ of them in terms of a weighted sum to become the new mean $\underline{X}_{k+1}$ of the next iteration.

Typically, $\mu$ is chosen as $\lfloor \lambda/2 \rfloor$ and $w_i = \ln(\frac{\lambda+1}{2}) - \ln(i) > 0$ in the scope of the CMA-ES [181]. As update rule for the step size $\sigma_k$ in the $(\mu/\mu_\mathrm{w}, \lambda)$-ES, several techniques such as self-adaptation [326] or cumulative step size adaptation [298] are available. Of theoretical interest is the so-called scale-invariant step size $\sigma_k = \sigma \|\underline{X}_k\|$ which depends on the distance to the origin and which allows to prove bounds on the convergence rate of evolution strategies with any adaptive step size update, see Sec. 14.3.

**Algorithm 8**: $(\mu/\mu_w, \lambda)$-ES

1: **given**: $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$, $\underline{X}_0 \in \mathbb{R}^{\mathbb{N}}$, $\sigma_0 > 0$, $\lambda \in \mathbb{N}^+$, $(\underline{\mathcal{N}}^r)_{r\in\mathbb{N}}$, weights $\underline{w} \in \mathbb{R}^\mu$, $\sum_{i=1}^{\mu} |w_i| = 1$ and $1 \le \mu \le \lambda$
2: $r \leftarrow 0$ number of random samples used
3: $k \leftarrow 0$ iteration counter for notational consistency
4: **while** stopping criterion not fulfilled **do**
5: /* *offspring generation* */
6: $i \leftarrow 0$ offspring counter
7: **while** $i < \lambda$ **do**
8: $i \leftarrow i + 1$
9: $r \leftarrow r + 1$
10: $\underline{X}_k^i = \underline{X}_k + \sigma_k \underline{\mathcal{N}}^r$
11: **end while**
12: $\underline{X}_{1:\lambda}, \ldots, \underline{X}_{\lambda:\lambda} = \text{argsort}(f(\underline{X}_k^1), \ldots, f(\underline{X}_k^\mu))$
13: /* *weighted recombination* */
14: $\underline{X}_{k+1} = \sum_{i=1}^{\mu} w_i \underline{X}_{i:\lambda}$
15: $\sigma_{k+1} = \text{update}(\sigma_k, f(\underline{X}_k^1), \ldots, f(\underline{X}_k^\mu))$
16: $k \leftarrow k + 1$ iteration counter
17: **end while**

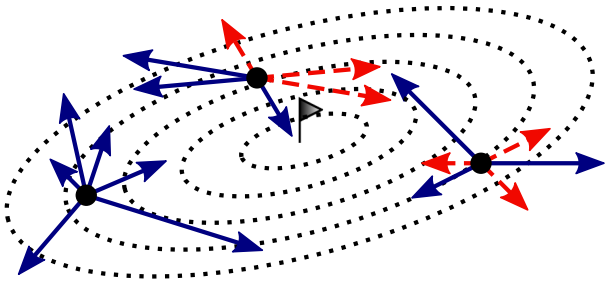


Figure 14.1: Illustration of i.i.d. mutations (left) and mirrored mutations with dependent (middle) and i.i.d. lengths (right). Independent samples are shown as solid and mirrored samples as dashed lines. Dotted lines connect search points with equal function value.

### 14.2.2 The Mirroring Idea

Derandomized mutations [344] and more recently mirrored mutations [85, 39] have been proposed to replace the independent mutations in evolution strategies by dependent ones in order to reduce the probability of "unlucky events"—resulting in an increase in the convergence speed of the algorithms. Instead of sampling the $\lambda$ offspring i.i.d. as in line 10 of Algorithm 8, an algorithm with mirrored mutations samples only $\lceil \lambda/2 \rceil$ i.i.d. offspring as $\underline{X}_k^i = \underline{X}_k + \sigma_k \underline{\mathcal{N}}^i$ $(1 \le i \le \lceil \lambda/2 \rceil)$ and the remaining $\lfloor \lambda/2 \rfloor$ offspring depending on the already drawn samples as $\underline{X}_k^i = \underline{X}_k - \sigma_k \underline{\mathcal{N}}^{i-\lceil \lambda/2 \rceil}$ (for $\lceil \lambda/2 \rceil + 1 \le i \le \lambda$), see Fig. 14.1, left and middle.

When using mirrored mutations in evolution strategies with weighted recombination and cumulative step-size adaption, the mirrored mutations cause a bias towards smaller step sizes [85], see Fig 14.2. The bias can cause premature convergence of the algorithm. The reason for the bias is that if both samples $\underline{X}_k + \sigma_k \underline{\mathcal{N}}^i$ and $\underline{X}_k - \sigma_k \underline{\mathcal{N}}^i$ are considered within weighted recombination, they partly cancel each other out and the realized shift of $\underline{X}_k$ will be smaller than with independent mutations. Consequently, derandomized step-size control like cumulative step size adaptation (CSA, [298]) will cause the step size to shrink.

In this report, we therefore introduce *pairwise selection* which prevents this bias: unmirrored and

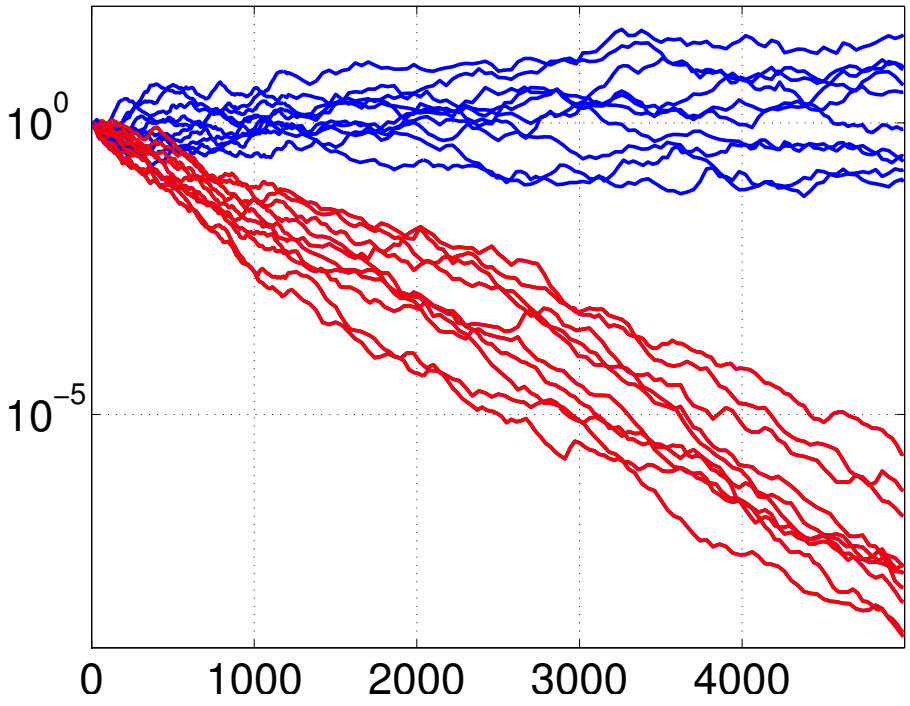



Figure 14.2: Illustration of the bias towards smaller step sizes under random selection introduced by recombination of mirrored vectors in the CMA-ES. Shown is the step-size $\sigma$ versus the number of function evaluations of 20 runs on a purely random fitness function in dimension 10. The upper ten graphs show the $(5/5_{\mathrm{w}}, 10)$-CMA-ES revealing a random walk on $\log(\sigma)$. The lower ten graphs show the $(5/5_{\mathrm{w}}, 5 + 5_{\mathrm{m}})$-CMA-ES without pairwise selection of mirrored samples.

mirrored offspring are paired two-by-two and only the better among the unmirrored sample $\underline{X}_k + \sigma_k \underline{\mathcal{N}}^i$ and its mirrored counterpart $\underline{X}_k - \sigma_k \underline{\mathcal{N}}^i$ is used within the weighted recombination but never both. Here, we introduce the following notations: the number of independent offspring per iteration is denoted by $\lambda_{\text{iid}}$ and the number of mirrored offspring per iteration is denoted by $\lambda_{\text{m}}$, where each iteration $\lambda = \lambda_{\text{iid}} + \lambda_{\text{m}}$ solutions are evaluated on $f$. As a result $0 \leq \lambda_{\text{m}} \leq \lambda_{\text{iid}}$ which results in the standard $(\mu/\mu_{\text{w}}, \lambda)$-ES in case $\lambda_{\text{m}} = 0$. We call the new algorithm $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}})$-ES.

Note that the idea of sequential mirroring of [85, 39], i.e., stopping the generation of new offspring as soon as a better solution than the parent is found, is not applied here. With recombination, the meaning of a comparison with "the parent" is not unique and additional algorithm design decisions were necessary[1]. Instead, *selective mirroring* is introduced.

We consider two variants of the $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}}+\lambda_{\text{m}})$-ES[2] that differ in the choice of mirrored offspring. The $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{rand}})$-ES, where $\lambda_{\text{m}}$ randomly chosen offspring are mirrored, and the $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{sel}})$-ES with *selective mirroring*, where only the $\lambda_{\text{m}}$ worst offspring are selected for mirroring.

### 14.2.3 Random and Selective Mirroring

The reason behind the latter variant of selecting the worst offspring for mirroring is the following: in particular on objective functions with convex sublevel sets[3] we do not expect the best of $\lambda_{\text{iid}}$ offspring to improve by mirroring. For an offspring that is better than the current search point $\underline{X}_k$, the mirroring would always result in a worse solution since never both an independently drawn solution and its mirrored counterpart can be better than the parent in case of convex sublevel sets [39, Proposition 2][4].

Regarding the comparison of random and selective mirroring, two questions arise: (i) how much faster than with random mirroring can an ES become if selective mirroring is used and (ii) what is the optimal choice for the number $\lambda_{\text{m}}$ of mirrored offspring. Both questions will be answered in the following by theoretical investigations of the algorithms' convergence rates.

### 14.2.4 Random Lengths of Mirrored Offspring

Within the $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{sel}})$-ES, solutions might be bad and selected for mirroring for two reasons: on the one hand, the solution could be bad because of a wrong direction and mirroring the point makes perfectly sense. On the other hand, a comparatively large distance of the new solution to the current mean might cause a bad function value. In this case, also mirroring will more likely produce a bad solution. Hence, we consider a variant of mirroring where the lengths of the mirrored vectors $\sigma_k \underline{\mathcal{N}}^r$ are i.i.d. resampled, i.e., where $\underline{X}_{k+1} = \underline{X}_k - \sigma_k \underline{\mathcal{N}}^r$ is replaced by $\underline{X}_{k+1} = \underline{X}_k - \sigma_k \frac{\|\widetilde{\mathcal{N}^r}\|}{\|\underline{\mathcal{N}}^r\|} \underline{\mathcal{N}}^r$ with $\|\widetilde{\mathcal{N}^r}\|$ the newly sampled length of the mirrored vector, cp. Fig. 14.1.

We refer to this last variant as $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{sel}})$-ES with i.i.d. mirrored mutation lengths. Algorithm 9 shows the pseudo code of all variants with random/selective mirroring and with/without independent lengths of the mirrored offspring. Theoretical results in the next section will not only show how much improvement in the convergence rate can be gained by the i.i.d. lengths but also that the variants with and without i.i.d. lengths are the same if the dimension goes to infinity. Later, we will apply independent lengths only to selective mirroring.

[1] The super-parent and distribution mean $\underline{X}_k$, resulting from the weighted recombination, is not directly comparable to the offspring, because depending on $\mathbb{N}$, $\mu$ and $\lambda$ with a large probability all i.i.d. sampled offspring might be worse. However, a feasible heuristic could be to compare with the best offspring from the last iteration.

[2] Adaptive variants with a variable number of mirrored offspring that depends on the observed fitness function values have also been considered but are not included here.

[3] The sublevel set $S_l$ contains all search points with an objective function value of at most $l$: $S_l = \{x \in \mathbb{R}^{\mathbb{N}} \mid f(x) \leq l\}$.

[4] In the case of random selection, even half of the $\lambda_{\text{iid}}$ independently generated offspring are in expectation better than the old mean such that only the mirroring of the $\lambda_{\text{iid}}/2$ worst of them will result in a further improvement.

### 14.2.5 Algorithm Parameters

The algorithms we have described involve several parameters, the number $\lambda_{\text{iid}}$ of independent samples, the number $\lambda_{\text{m}}$ of mirrored offspring, the number of offspring to be recombined, $\mu$, and the weights $\underline{w}$ for the recombination. The convergence rates depend on the choice of these parameters. In the sequel, we are going to investigate the optimal choice of the different parameters by investigating lower bounds for the convergence rate on spherical functions $g(\|\underline{x}\|)$, $g \in \mathcal{M}$ where $\mathcal{M}$ denotes the set of functions $g : \mathbb{R} \mapsto \mathbb{R}$ that are strictly increasing.

## 14.3 Convergence Rate Lower Bounds on Spherical Functions

In order to find optimal settings for the different parameters, we investigate convergence rates on spherical functions having WLOG the optimum in zero, i.e., $g(\|x\|)$, $g \in \mathcal{M}$. Convergence rates depend on the step-size adaptation chosen. We study the case of the scale-invariant step-size where $\sigma_k = \sigma\|\underline{X}_k\|$, that leads, for an optimal choice of the constant $\sigma$, to lower bounds for convergence rates achievable by any strategy with step-size adaptation on spherical functions (see below). For the different algorithm variants with scale-invariant step-size, we prove linear convergence in expectation in the following sense: there exists a $\text{CR} \in \mathbb{R}$ such that for all $k, k_0 \in \mathbb{N}$ with $k > k_0$

$$\frac{1}{\Lambda}\frac{1}{k-k_0} E\left[\ln \frac{\|\underline{X}_k\|}{\|\underline{X}_{k_0}\|}\right] = \text{CR} \ , \tag{14.1}$$

where $\Lambda$ is the number of evaluations per iteration introduced to define the convergence rate per function evaluation. The constant CR[5] is the convergence rate of the algorithm and depends on the different parameters $(\sigma, \lambda, \mu, \ldots)$. The previous equation defining the convergence rate is compatible with the almost sure convergence that we will be proving as well [41]. Almost surely we will prove that with scale-invariant step-size

$$\frac{1}{\Lambda}\frac{1}{k} \ln \frac{\|\underline{X}_k\|}{\|\underline{X}_0\|} \xrightarrow[k\to\infty]{} \text{CR} \ . \tag{14.2}$$

### 14.3.1 The $(\mu/\mu_{\mathbf{w}}, \lambda)$-ES

To serve as a baseline algorithm for a later comparison with algorithms using mirrored mutations, we first investigate the convergence rate of the scale-invariant version of the standard $(\mu/\mu_{\text{w}}, \lambda)$-ES (see Algorithm 8 for the pseudo code).

#### Finite Dimension Results

At each step, $\lambda$ independent vectors following a standard multivariate normal distribution $\underline{\mathcal{N}}^i$ are sampled to create the offspring $\underline{X}_k^i = \underline{X}_k + \sigma\|\underline{X}_k\|\underline{\mathcal{N}}^i$. The offspring are ranked according to their fitness function value. We denote $(\mathcal{Z}[1:\lambda], \ldots, \mathcal{Z}[\lambda:\lambda])$ the sorted multivariate normal distribution, namely the best offspring equals $\underline{X}_k + \sigma\|\underline{X}_k\|\mathcal{Z}[1:\lambda]$, the second best $\underline{X}_k + \sigma\|\underline{X}_k\|\mathcal{Z}[2:\lambda]$, etc. The distribution of $(\mathcal{Z}[1:\lambda], \ldots, \mathcal{Z}[\lambda:\lambda])$ depends *a priori* on $\underline{X}_k$. However, in the scale-invariant step-size case on spherical functions the distribution is independent of $\underline{X}_k$ and is determined by ranking of $\|\underline{e}_1 + \sigma\underline{\mathcal{N}}^i\|$ for $i = 1 \ldots \lambda$, that can be simplified by ranking $2[\underline{\mathcal{N}}^i]_1 + \sigma\|\underline{\mathcal{N}}^i\|^2$. Those results are stated in the following lemma.

[5] Convergence takes place if and only if $\text{CR} < 0$, however in general we are not able to investigate *theoretically* the sign of CR.

**14.3.1 Lemma**
*For the scale-invariant $(\mu/\mu_w, \lambda)$-ES minimizing spherical functions, the probability distribution of the vector $(\mathbb{Z}[1:\lambda], \dots \mathbb{Z}[\lambda:\lambda])$ is independent of $\underline{X}_k$ and equals*

$$(\mathbb{Z}[1:\lambda], \dots, \mathbb{Z}[\lambda:\lambda]) = \text{argsort}\{h_\sigma(\underline{\mathcal{N}}^1), \dots, h_\sigma(\underline{\mathcal{N}}^\lambda)\} \tag{14.3}$$

*where $h_\sigma(\underline{x}) = 2[\underline{x}]_1 + \sigma\|\underline{x}\|^2$ and $(\underline{\mathcal{N}}^i)_{1\le i\le\lambda}$ are $\lambda$ independent standard multivariate normal distribution.*

*Proof.* All proofs can be found in the appendix. □

In the $(\mu/\mu_w, \lambda)$-ES, the $\mu$ best offspring $\underline{X}_k + \sigma\|\underline{X}_k\|\mathbb{Z}[i:\lambda]$ for $i = 1, \dots, \mu$ are recombined into the new parent $\underline{X}_{k+1} = \underline{X}_k + \sigma\|\underline{X}_k\| \sum_{i=1}^{\mu} w_i \mathbb{Z}[i:\lambda]$ where $(w_1, \dots, w_\mu) \in \mathbb{R}^\mu$ and $\sum_{i=1}^{\mu} |w_i| = 1$. The next theorem gives the expression of the convergence rate associated to the $(\mu/\mu_w, \lambda)$-ES with scale-invariant step-size as a function of $\sigma$ and $\underline{w} = (w_1, \dots, w_\mu)$.

**14.3.2 Theorem**
*For the $(\mu/\mu_w, \lambda)$-ES with scale-invariant step-size on $g(\|x\|)$, $g \in \mathcal{M}$,* (14.1) *and* (14.2) *hold and the convergence rate equals*

$$\text{CR}(\sigma, \underline{w}) = \frac{E \ln\left[1 + 2\sum_{i=1}^{\mu} \sigma w_i [\mathbb{Z}[i:\lambda]]_1 + \left\|\sum_{i=1}^{\mu} \sigma w_i \mathbb{Z}[i:\lambda]\right\|^2\right]}{2\lambda} ,$$

*where $w_i \in \mathbb{R}$ and $\sum_{i=1}^{\mu} |w_i| = 1$.*

The convergence rate of the $(\mu/\mu_w, \lambda)$-ES is a function of $\sigma$ and the weights. However, defining the function

$$\mathcal{D}_{(\mu/\mu_w,\lambda)}(\underline{y}) = \frac{1}{2\lambda} E \ln\left[1 + 2\sum_{i=1}^{\mu} y_i [\mathbb{Z}[i:\lambda]]_1 + \left\|\sum_{i=1}^{\mu} y_i \mathbb{Z}[i:\lambda]\right\|^2\right] ,$$

where $\underline{y} = (y_1, \dots, y_\mu)$ and the distribution of $\mathbb{Z}[i:\lambda]$ is determined with (14.3) and $\sigma = \sum_{i=1}^{\mu} |y_i|$, we have that

$$\text{CR}_{(\mu/\mu_w,\lambda)}(\sigma, \underline{w}) = \mathcal{D}_{(\mu/\mu_w,\lambda)}(\sigma\underline{w}) .$$

The optimal convergence rate of the scale-invariant $(\mu/\mu_w, \lambda)$-ES is defined as the solution of the minimization problem

$$\text{CR}^{\text{opt}}_{(\mu/\mu_w,\lambda)} = \min_{\underline{y}\in\mathbb{R}^\mu} \mathcal{D}_{(\mu/\mu_w,\lambda)}(\underline{y}) . \tag{14.4}$$

Optimal weights will then be obtained as

$$w_i^{\text{opt}} = \frac{y_i^{\text{opt}}}{\sum_{i=1}^{\mu} |y_i^{\text{opt}}|} \text{ with } \underline{y}^{\text{opt}} = \text{argmin}\, \mathcal{D}_{(\mu/\mu_w,\lambda)}(\underline{y})$$

and the optimal step-size equals $\sum_{i=1}^{\mu} |y_i^{\text{opt}}|$. The convergence rate of the scale-invariant step-size $(\mu/\mu_w, \lambda)$-ES corresponds to a lower bound for the convergence rate of any step-size adaptive $(\mu/\mu_w, \lambda)$-ES with isotropic distribution on spherical functions [205, Theorem 1].

**Asymptotic Results**

In the following, we investigate the limit of the convergence rate when the dimension of the search space goes to infinity.

**14.3.3 Theorem**
*The convergence rate of the $(\mu/\mu_{\mathrm{w}}, \lambda)$-ES with scale-invariant step size and weights $\underline{w} \in \mathbb{R}^\mu$ with $\sum_{i=1}^{\mu} |w_i| = 1$ on the class of spherical functions $g(\|x\|), g \in \mathcal{M}$ satisfies*

$$\lim_{d\to\infty} d\,\mathrm{CR}\left(\frac{\sigma}{d}, w\right) = \frac{1}{\lambda}\left(\frac{\sigma^2}{2}\sum_{i=1}^{\mu} w_i^2 + \sigma\sum_{i=1}^{\mu} w_i E(\mathcal{N}_{i:\lambda})\right)$$

*where $\mathcal{N}_{i:\lambda}$ is the $i$th order statistic of $\lambda$ independent normal distributions with mean 0 and variance 1, i.e., the $i$th smallest of $\lambda$ independent variables $\mathcal{N}_i \sim N(0,1)$.*

We define the limit in the previous theorem, multiplied by $(-1)$, as the function

$$\varphi_{(\mu/\mu_{\mathrm{w}},\lambda)}(\sigma, \underline{w}) := -\frac{1}{\lambda}\left(\frac{\sigma^2}{2}\sum_{i=1}^{\mu} w_i^2 + \sigma\sum_{i=1}^{\mu} w_i E(\mathcal{N}_{i:\lambda})\right)$$

corresponding to the asymptotic progress rate of the $(\mu/\mu_{\mathrm{w}}, \lambda)$-ES [21]. As for the finite dimension case, we consider the variable $y = \sigma \cdot \underline{w} \in \mathbb{R}^\mu$ and introduce the function

$$\mathcal{G}(\underline{y}) = \frac{1}{\lambda}\left(\frac{1}{2}\sum_{i=1}^{\mu} y_i^2 + \sum_{i=1}^{\mu} y_i E(\mathcal{N}_{i:\lambda})\right) \tag{14.5}$$

that satisfies $\mathcal{G}(\sigma \cdot \underline{w}) = -\varphi_{(\mu/\mu_{\mathrm{w}},\lambda)}(\sigma, \underline{w})$. Optimal asymptotic convergence rates realize the minimum of (14.5) that is reached for $y_i^{\mathrm{opt}} = -E(\mathcal{N}_{i:\lambda})$ and equals

$$\mathrm{CR}^{\mathrm{opt},\infty}_{(\mu/\mu_{\mathrm{w}},\lambda)} := \min_{\underline{y}\in\mathbb{R}^\mu} \mathcal{G}(\underline{y}) = -\frac{1}{2\lambda}\sum_{i=1}^{\mu} E(\mathcal{N}_{i:\lambda})^2 \ , \tag{14.6}$$

as already found in [21]. Optimal weights are proportional to $y_i^{\mathrm{opt}}$ and their absolute value sum to one. Thus $w_i^{\mathrm{opt},\infty} = -E(\mathcal{N}_{i:\lambda})/\sum_{i=1}^{\mu} |E(\mathcal{N}_{i:\lambda})|$. Whether or not *negative* weights are allowed does not effect the optimal *positive* weight values, besides from a different normalization factor.

### 14.3.2 The $(\mu/\mu_{\mathrm{w}}, \lambda)$-ES With Random and Selective Mirroring

Following the same approach than in the previous section, we analyze the convergence rate of the different mirroring variants first for finite dimension and then asymptotically in the dimension. We define as $(\mathbb{Z}[1], \ldots, \mathbb{Z}[\lambda_{\mathrm{iid}}])$, the vector of ordered steps to be recombined, namely for a given variant, the best point to be recombined (for which the highest weight will be given) is $\underline{X}_k + \sigma\|\underline{X}_k\|\mathbb{Z}[1]$, the second best $\underline{X}_k + \sigma\|\underline{X}_k\|\mathbb{Z}[2], \ldots$. Among the different variants, the distribution of the vector of ordered steps is changing. We express in the sequel this distribution for the variants with random mirroring, selective mirroring and selective mirroring with independent length.

**Selected vector for random mirroring** In random mirroring, we randomly mirror $\lambda_{\mathrm{m}}$ vectors among the $\lambda_{\mathrm{iid}}$ independent ones. Without loss of generality, we can mirror the $\lambda_{\mathrm{m}}$ last vectors. For the mirrored pairs, only the best of the two vectors is recombined. The distribution of the resulting vector of ordered steps is expressed in the following lemma:

**14.3.4 Lemma**
*In the $(\mu/\mu_{\mathrm{w}}, \lambda_{\mathrm{iid}} + \lambda_{\mathrm{m}}^{\mathrm{rand}})$-ES with scale-invariant step-size on spherical functions, the distribution of the vector of ordered steps to be recombined is given by*

$$\begin{aligned}(\mathbb{Z}[1], \ldots, \mathbb{Z}[\lambda_{\mathrm{iid}}]) = \mathrm{argsort}\{&h_\sigma(\underline{\mathcal{N}}^1), \ldots, h_\sigma(\underline{\mathcal{N}}^{\lambda_{\mathrm{iid}}-\lambda_{\mathrm{m}}}),\\ &\min\{h_\sigma(\underline{\mathcal{N}}^{\lambda_{\mathrm{iid}}-\lambda_{\mathrm{m}}+1}), h_\sigma(-\underline{\mathcal{N}}^{\lambda_{\mathrm{iid}}-\lambda_{\mathrm{m}}+1})\}, \ldots,\\ &\min\{h_\sigma(\underline{\mathcal{N}}^{\lambda_{\mathrm{iid}}}), h_\sigma(-\underline{\mathcal{N}}^{\lambda_{\mathrm{iid}}})\}\}\end{aligned} \tag{14.7}$$

*where* $h_\sigma(\underline{x}) = 2[x]_1 + \sigma\|x\|^2$.

**Selected vector for selective mirroring** In selective mirroring, we mirror the worse offspring, the first step is then to sort the $\lambda_{\text{iid}}$ offspring to determine which offspring to mirror. Let $\mathcal{Y}$ be defined as

$$\mathcal{Y} := (\underline{Y}_1, \ldots, \underline{Y}_{\lambda_{\text{iid}}}) := \text{argsort}\{h_\sigma(\underline{\mathcal{N}}^1), \ldots, h_\sigma(\underline{\mathcal{N}}^{\lambda_{\text{iid}}})\}$$

where $h_\sigma(\underline{x}) = 2[x]_1 + \sigma\|x\|^2$. Then for the worst $\lambda_{\text{m}}$ vectors of $\mathcal{Y}$, we select the pair-wise best among offspring and mirrored one and we keep the other vectors unchanged:

$$\underline{Y}_i^{\text{sel}} = \underline{Y}_i, i = 1, \ldots, \lambda_{\text{iid}} - \lambda_{\text{m}} \tag{14.8}$$

$$\underline{Y}_i^{\text{sel}} = \text{argmin}\{h_\sigma(\underline{Y}_i), h_\sigma(-\underline{Y}_i)\}, \lambda_{\text{iid}} - \lambda_{\text{m}} + 1 \leq i \leq \lambda_{\text{iid}} \tag{14.9}$$

Finally, as expressed in the following lemma, the distribution of the $\lambda_{\text{iid}}$ ordered steps to be recombined is the result of the sorting of the $\underline{Y}_i^{\text{sel}}$ vectors:

**14.3.5 Lemma**
*In the* $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{sel}})$*-ES with scale-invariant step-size on spherical functions, the distribution of the vector of ordered steps to be recombined is given by*

$$(\mathbb{Z}[1], \ldots, \mathbb{Z}[\lambda_{\text{iid}}]) \quad = \quad \text{argsort}\{h_\sigma(\underline{Y}_1^{\text{sel}}), \ldots, h_\sigma(\underline{Y}_{\lambda_{\text{iid}}}^{\text{sel}})\} \ , \tag{14.10}$$

*where* $\underline{Y}_i^{\text{sel}}$ *is defined in* (14.8) *and* (14.9).

**Selected vector for selective mirroring with independent length** The selective mirroring with random length algorithm differs from the previous one for the mirroring step where only the direction is kept and the length is changed by sampling an independent length (distributed according to a $\chi$-distribution with d degrees of freedom). Assuming that sorting of the $\lambda_{\text{iid}}$ offspring has been made according to $\mathcal{Y}$ as described above, the $\underline{Y}^{\text{sel}}$ vector is given by

$$\underline{Y}_i^{\text{sel}} = \underline{Y}_i \text{ for } i = 1, \ldots, \lambda_{\text{iid}} - \lambda_{\text{m}} \tag{14.11}$$

and for $i = \lambda_{\text{iid}} - \lambda_{\text{m}} + 1, \ldots, \lambda_{\text{iid}}$,

$$\underline{Y}_i^{\text{sel}} = \text{argmin}\left\{h_\sigma(\underline{Y}_i), h_\sigma\left(-\|\widetilde{\underline{\mathcal{N}}^i}\|\frac{\underline{Y}_i}{\|\underline{Y}_i\|}\right)\right\} \tag{14.12}$$

where $\widetilde{\underline{\mathcal{N}}^i}$ are independent vectors following a standard multivariate normal distribution. As for the previous algorithm, the distribution of the $\lambda_{\text{iid}}$ ordered steps to be recombined is the result of the sorting of the $\underline{Y}_i^{\text{sel}}$ vectors:

**14.3.6 Lemma**
*In the* $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{sel}})$*-ES with scale-invariant step-size on spherical functions, the distribution of the vector of ordered steps to be recombined is given by*

$$(\mathbb{Z}[1], \ldots, \mathbb{Z}[\lambda_{\text{iid}}]) \quad = \quad \text{argsort}\{h_\sigma(\underline{Y}_1^{\text{sel}}), \ldots, h_\sigma(\underline{Y}_{\lambda_{\text{iid}}}^{\text{sel}})\} \ , \tag{14.13}$$

*where* $\underline{Y}_i^{\text{sel}}$ *is defined in* (14.11) *and* (14.12).

Similarly as for the $(\mu/\mu_{\text{w}}, \lambda)$-ES, we find that the convergence rate of the $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{rand}})$-ES and the $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{sel}})$-ES with and without i.i.d. mirrored mutation lengths can be expressed in the following way

**14.3.7 Theorem**
*The convergence rate of the $(\mu/\mu_w, \lambda_{iid} + \lambda_m^{rand})$-ES and the $(\mu/\mu_w, \lambda_{iid} + \lambda_m^{sel})$-ES with and without i.i.d. mirrored mutation lengths equals*

$$\mathrm{CR}(\sigma, \underline{w}) = \frac{E \ln\left[1 + 2\sum_{i=1}^{\mu} \sigma w_i [\mathcal{Z}[i]]_1 + \left\|\sum_{i=1}^{\mu} \sigma w_i \mathcal{Z}[i]\right\|^2\right]}{2(\lambda_{iid} + \lambda_m)} ,$$

*where $w_i \in \mathbb{R}$ and $\sum_{i=1}^{\mu} |w_i| = 1$ and the distribution of the random vector $(\mathcal{Z}[1], \ldots, \mathcal{Z}[\lambda_{iid}])$ are defined in Lemma 14.3.4, 14.3.5 and 14.3.6 respectively.*

As for the $(\mu/\mu_w, \lambda)$-ES, optimal convergence rates are solutions of the minimization problem $\min_{\underline{y} \in \mathbb{R}^\mu} \mathcal{D}(\underline{y})$ where

$$\mathcal{D}(\underline{y}) = \frac{1}{\lambda_{iid} + \lambda_m} \frac{1}{2} E \ln\left[1 + 2\sum_{i=1}^{\mu} y_i [\mathcal{Z}[i]]_1 + \left\|\sum_{i=1}^{\mu} y_i \mathcal{Z}[i]\right\|^2\right]$$

and $\sigma$—needed to obtain the distribution of $\mathcal{Z}[i]$—equals $\sum |y_i|$. Optimal weights are then obtained by normalizing the solution vector $\underline{y}^{opt}$.

**Asymptotic Results**

We investigate the limit of the convergence rate given in Theorem 14.3.7 when the dimension goes to infinity. For the $(\mu/\mu_w, \lambda_{iid} + \lambda_m^{rand})$, we define the random vector $(Z_1, \ldots, Z_{\lambda_{iid}}) \in \mathbb{R}^{\lambda_{iid}}$ as

$$(Z_1, \ldots, Z_{\lambda_{iid}}) = \mathrm{argsort}\{\mathcal{N}^1, \ldots, \mathcal{N}^{\lambda_{iid}-\lambda_m}, - |\mathcal{N}^{\lambda_{iid}-\lambda_m+1}|, \ldots, -|\mathcal{N}^{\lambda_{iid}}|\} \quad (14.14)$$

where $\mathcal{N}^i$ are $\lambda_{iid}$ independent standard normal distribution.

For the $(\mu/\mu_w, \lambda_{iid} + \lambda_m^{sel})$ with or without independent lengths, we define the vector

$$(Z_1, \ldots, Z_{\lambda_{iid}}) = \mathrm{argsort}\{Y_1, \ldots, Y_{\lambda_{iid}-\lambda_m}, - |Y_{\lambda_{iid}-\lambda_m+1}|, \ldots, -|Y_{\lambda_{iid}}|\} \quad (14.15)$$

where $(Y_1, \ldots, Y_{\lambda_{iid}}) = \mathrm{argsort}\{\mathcal{N}^1, \ldots, \mathcal{N}^{\lambda_{iid}}\}$. The asymptotic convergence rate for different variants is given in the following theorem.

**14.3.8 Theorem**
*The convergence rate of the $(\mu/\mu_w, \lambda_{iid} + \lambda_m^{rand})$-ES and the $(\mu/\mu_w, \lambda_{iid} + \lambda_m^{sel})$-ES (with or without independent lengths) with scale-invariant step size and weights $\underline{w} \in \mathbb{R}^\mu$ with $\sum_{i=1}^{\mu} |w_i| = 1$ on the class of spherical functions $g(\|x\|)$, $g \in \mathcal{M}$ satisfies*

$$\lim_{d \to \infty} d\,\mathrm{CR}\left(\frac{\sigma}{d}, w\right) = \frac{1}{\lambda_{iid} + \lambda_m} \left(\frac{\sigma^2}{2} \sum_{i=1}^{\mu} w_i^2 + \sigma \sum_{i=1}^{\mu} w_i E(Z_i)\right)$$

*where the distribution of $Z_i$ is given in* (14.14) *for the $(\mu/\mu_w, \lambda_{iid} + \lambda_m^{rand})$-ES and in* (14.15) *for the $(\mu/\mu_w, \lambda_{iid} + \lambda_m^{sel})$-ES with or without independent length for the mirroring.*

Similarly to the $(\mu/\mu_w, \lambda)$-ES case, we find that the optimal convergence rate is given by

$$\mathrm{CR}^{opt,\infty} := -\frac{1}{2(\lambda_{iid} + \lambda_m)} \sum_{i=1}^{\mu} E(Z_i)^2 , \quad (14.16)$$

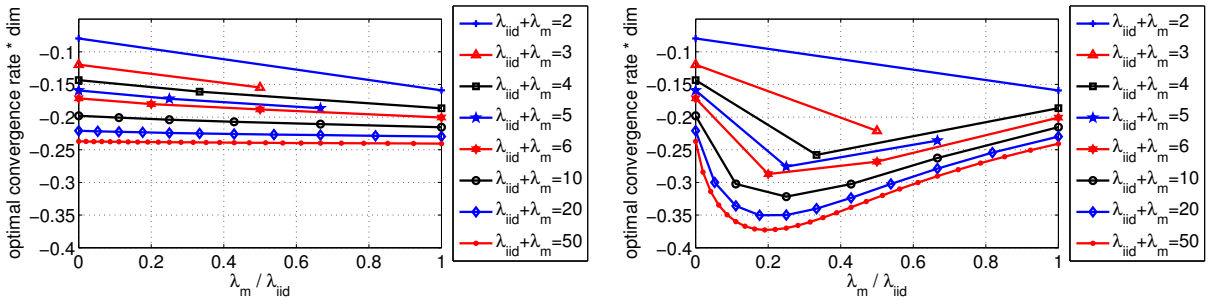



Figure 14.3: Estimated optimal convergence rates for the $(\mu/\mu_w, \lambda_{iid} + \lambda_m)$-ES depending on the ratio of mirrored and independent offspring comparing choices with constant $\lambda_{iid} + \lambda_m$: on the left, the random variant $(\mu/\mu_w, \lambda_{iid} + \lambda_m^{rand})$-ES and on the right, the $(\mu/\mu_w, \lambda_{iid} + \lambda_m^{sel})$-ES, mirroring the worst $\lambda_m$ independent offspring.

and the optimal weights equal $w_i^{opt} = -E(Z_i)/\sum_{i=1}^{\mu} |E(Z_i)|$. We remark that the asymptotic convergence rate for the selective mirroring is the same with or without sampling independent lengths for the mirroring vectors. Thus the independent length sampling can only affect *finite dimension* results. In the remainder, all convergence rates are estimated with only positive recombination weights.

## 14.4 Simulation of Convergence Rates

Due to their implicit nature, the above derived optimal convergence rates are difficult to compare directly. However, we can estimate the rates by means of Monte Carlo sampling easily which allows us to compare the performance of the proposed algorithms in infinite and finite dimension. Moreover, we are able to decide what is the optimal ratio between the number of mirrored and unmirrored offspring and which is, theoretically and on spherical functions, the best of the algorithms. If not mentioned differently, $10^6$ samples are used for each combination of $\lambda_{iid}$ and $\lambda_m$, each dimension, and for each algorithm. The MATLAB code will be made available online after the double-blind reviewing process.

**Random and Selective Mirroring in Infinite Dimension** Let us first have a look at the scale-invariant random mirroring strategy in infinite dimension. The lefthand plot in Fig. 14.3 show the estimated convergence rates versus the number of independent offspring and versus the total number of offspring respectively. We see that in all cases, the convergence rate monotonically improves, the more offspring are mirrored. The optimal ratio between mirrored and independent offspring is therefore 1. Moreover, we see that with an increasing number of offspring, the convergence rate approaches $-0.25$.

The result looks quite different with *selective* mirroring, see the righthand plot in Fig. 14.3: with increasing number of mirrored solutions per iteration, the optimal convergence rate first improves from the convergence rate of the standard $(\mu/\mu_w, \lambda)$-ES ($\lambda_m = 0$) to its minimum at around $\lambda_m \approx \frac{\lambda_{iid}}{5}$ and with further increased $\lambda_m$, the convergence rate worsens again. Similar than with random mirroring, the optimal convergence rate improves with increased number of offspring, but reaches far better optimal values with selective mirroring. The limit values for $\lambda \to \infty$ are 0.25 for random mirroring and above 0.386 for selective mirroring. Figure 14.4 shows the best convergence rates among all selective mirroring strategies with constant $\lambda_{iid} + \lambda_m$ from Fig. 14.3 together with the corresponding ratio $(\lambda_m/\lambda_{iid})$. The optimal convergence rate approaches about 0.3875 when the number of offspring increases and the optimal ratio indeed goes to a value of (slightly below) 1/5. Note that the unsmoothness of the ratio stems from discretization: not all values for the ratio of $\lambda_m/\lambda_{iid}$ are possible which in particular has an effect for small $\lambda_{iid}$.

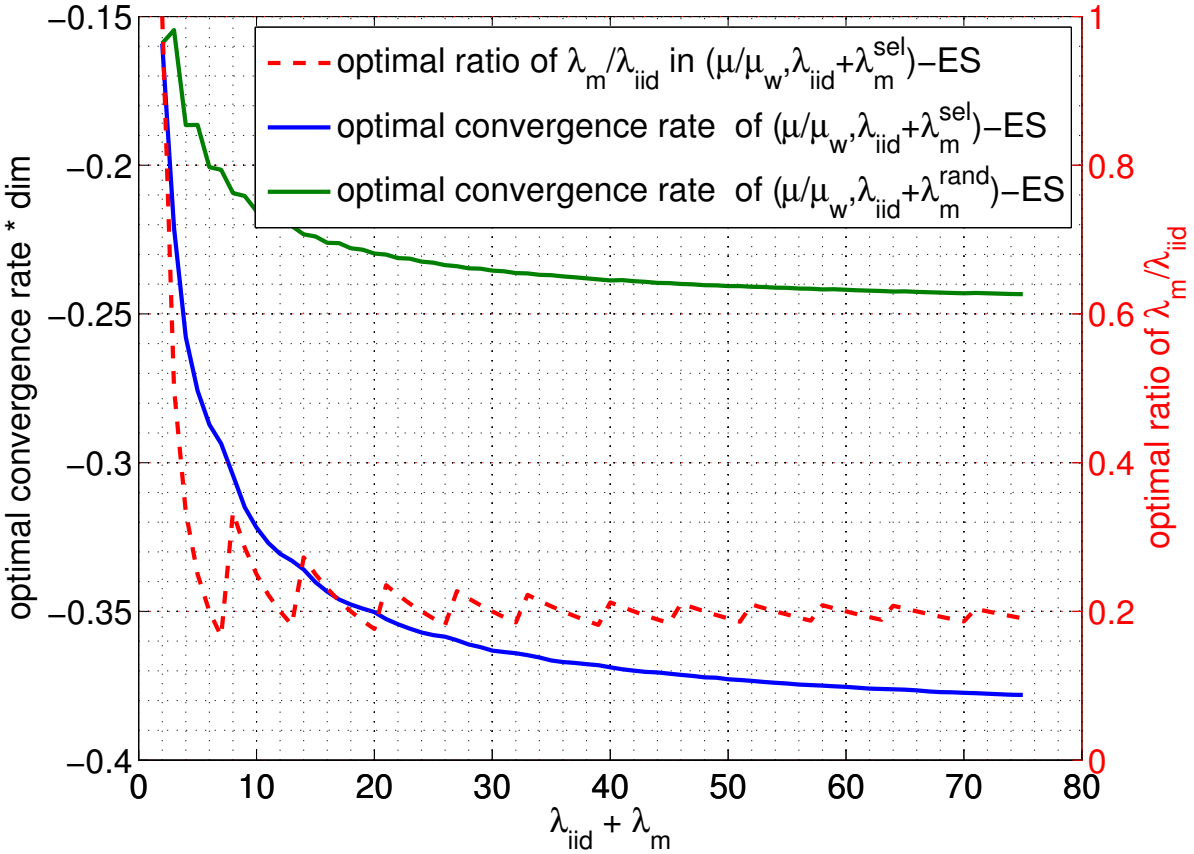



Figure 14.4: Extracted normalized optimal convergence rates (solid lines) for the $(\mu/\mu_w, \lambda_{iid} + \lambda_m^{rand})$-ES (top) and $(\mu/\mu_w, \lambda_{iid} + \lambda_m^{sel})$-ES (bottom) of Fig. 14.3 for different numbers of offspring $(\lambda_{iid} + \lambda_m)$ together with the corresponding optimal ratio of mirrored and unmirrored offspring for the selective mirroring variant (dashed).

**Selective Mirroring in Finite Dimension** When estimating the convergence rates of the proposed strategies in finite dimension, one faces first another problem: the derivation of optimal weights. Whereas in the infinite case, the optimal weights are known theoretically, this is not the case in finite dimension. We therefore estimate the optimal weights by optimizing the convergence rates of Sec. 14.3 numerically. The convergence rate can be estimated by a Monte Carlo simulation, resulting thus in a noisy optimization problem. We use UH-CMA-ES (CMA-ES with uncertainty handling, [182])[6] to solve this problem. The number of Monte Carlo estimates to compute the convergence rate per offspring is adapted with an initial value of 100 and a maximal value of $10^4$. The shown weights are extracted by averaging the sample distribution mean from CMA-ES over the last half of the overall iterations, typically the last 150 iterations. For each setting two experiments are done.

The results with independently re-sampled vector lengths are shown in Fig. 14.5 for $\lambda = 10$ in dimension 3 and 30 and compared to the optimal weights for infinite dimension (right). For small or moderate dimension, when $\lambda \not\gg \mathbb{N}$, the spread of the weights is much more pronounced than with $\mathbb{N}$-$\infty$-optimal weights. Also the negative weights get less important with increasing $\mathbb{N}/\lambda$.

For the optimal $\lambda_m = 2$, the default recombination weights $w_i \propto \log((\lambda+1)/2) - \log(i)$ as used in the CMA-ES resemble the optimal weights much better than the $\mathbb{N}$-infinity optimal weights. Further experiments have been conducted with mirrored lengths and also with $\lambda = 20$ leading to the same qualitative observations. We conclude that the default recombination weights are well applicable with mirroring.

[6] We have used the Matlab code cmaes.m from `http://www.lri.fr/~hansen/cmaes_inmatlab.html`, Version 3.54.beta.

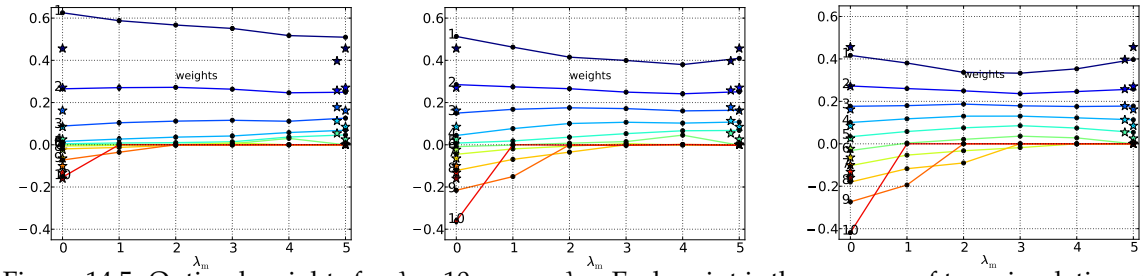



Figure 14.5: Optimal weights for $\lambda = 10$ versus $\lambda_m$. Each point is the average of two simulations (small black dots). All weights are normalized such that the sum of positive weights is one. Stars depict the default weights in CMA-ES and the $\mathbb{N}$-$\infty$-optimal weights for $\lambda_m = 5$. Dimension 3, 30 (both estimated with UH-CMA-ES) and $\infty$ from left to right.

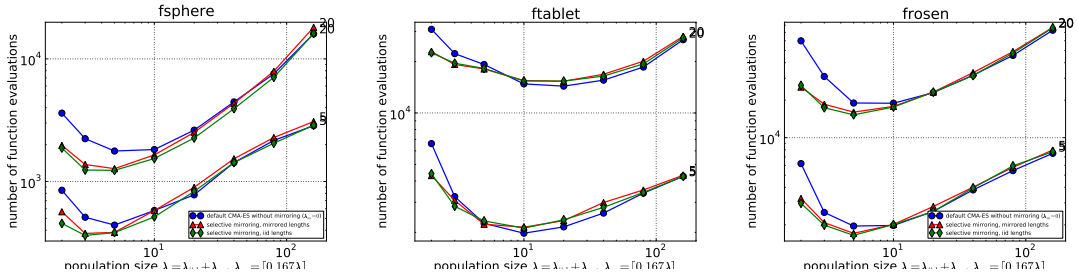



Figure 14.6: Number of function evaluations to reach function value $10^{-8}$ with CMA-ES on the sphere, tablet and Rosenbrock function in 5- and 20-D versus population size $\lambda$.

## 14.5 $(\mu/\mu_w, \lambda)$-CMA-ES With Mirroring

We apply selective mirroring as outlined in Algorithm 9 to the CMA-ES. Default recombination weights and parameter settings as in the downloaded MATLAB code are used with the modifications from [85]. The number of mirrored solutions is set to $\lceil 0.167\lambda \rceil$.

Results on the sphere, $\|x\|^2$, the tablet, $10^6 x_1^2 + \sum_{i=2}^{\mathbb{N}} x_i^2$, and Rosenbrock function, $\sum_{i=2}^{\mathbb{N}} 100(x_i^2 - x_{i-1})^2 + (x_i - 1)^2$, are shown in Fig. 14.6. Initial values are coordinate-wise 0.1 for the mean and 0.2 for the step-size. Shown are the number of function evaluations to reach function value $10^{-8}$.

On all but one unimodal functions that we have tested, the effect of mirroring resembles the effect on the sphere function. Mirroring is only effective with a small population size ($\lambda < \mathbb{N}$) and i.i.d. lengths are slightly favorable. The only exception is the tablet function where we also see a slight disadvantage for larger values of $\lambda$ in Fig. 14.6.

**Testing on the COCO/BBOB Function Suite** The IPOP-CMA-ES (CMA-ES restarted with Increasing POPulation size [40]) with mirroring using i.i.d. lengths and without mirroring has been run on the BBOB-2010 benchmark [178, 177]. The empirical run length distribution over all functions with and without mirroring for up to $3 \times 10^4 \mathbb{N}$ function evaluations is shown in Fig. 14.7.

The overall effect is almost negligible. Also on most single functions, the performance difference is small. The one exception is the Schwefel function f20 shown in Fig. 14.7, right. Mirroring has a strong adversarial effect on the performance on this function.

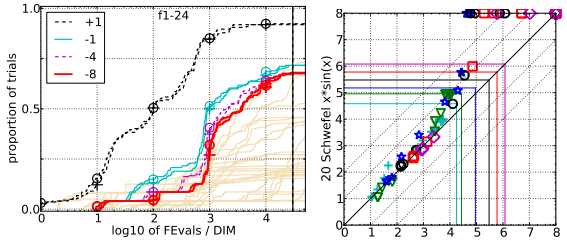



Figure 14.7: Left: run length distribution (in number of function evaluations) of the IPOP-CMA-ES with ∘ and without + mirroring over all BBOB-2010 functions in 20-D using target values $10^{-8,-4,-1,1}$. Right: average number of function evaluations (in log10) to reach respective target values, $x$-value without mirroring, $y$-value with mirroring. Markers on the edge indicate the target value was not reached. Markers represent dimension: 2:+, 3:▽, 5:★, 10:∘, 20:□, 40:◇.

## 14.6 Summary and Conclusion

We have introduced mirrored sampling in ESs with multi-recombination. Two important tricks are used. Selective mirroring: only the worst offspring are mirrored; pairwise selection: only at most one from a mirrored/unmirrored pair is selected for recombination. Our theoretical results support the effectiveness in particular of selective mirroring: the new algorithm improves the known convergence rate record for ESs with positive recombination weights by 56% from 0.25 to 0.390. This is a huge improvement and the new $(\mu/\mu_w, \lambda_{iid} + \lambda_m)$-ES is also more than 60% faster than the fastest single-parent mirroring $(1+1_{ms})$-ES and almost twice as fast as the regular (1+1)-ES in the asymptotic limit.

Applied to CMA-ES, mirrored sampling improves the convergence speed in small populations. However, on a single multimodal function in the BBOB benchmark it leads to a considerable performance decline. Nevertheless, we believe that mirrored sampling has yet a great future potential, in particular when also cleverly exploited for the covariance matrix update.

**Algorithm 9**: $(\mu/\mu_\mathrm{w}, \lambda_\mathrm{iid}+\lambda_\mathrm{m})$-ES with random or selective mirroring with or without i.i.d. lengths of the mirrored vectors

**given**: $f : \mathbb{R}^\mathbb{N} \to \mathbb{R}$, $\underline{X}_0 \in \mathbb{R}^\mathbb{N}$, $\sigma_0 > 0$, $\lambda_\mathrm{iid} \in \mathbb{N}^+$, $\lambda_\mathrm{m} \in \{0, \ldots, \lambda_\mathrm{iid}\}$, $\mu \le \lambda_\mathrm{iid}$, $(\underline{\mathcal{N}}^r)_{r\in\mathbb{N}}$, weights $\underline{w} \in \mathbb{R}^\mu$ with $\sum_{i=1}^{\mu} |w_i| = 1$ and $|\{w_i \ge 0\}| \ge \lambda_\mathrm{m}$
$r \leftarrow 0$ *(number of random samples used)*
$k \leftarrow 0$ *(iteration counter for notational consistency)*
**while** stopping criterion not fulfilled **do**
  */* use/rename samples */*
  $\mathbf{N}^1 \ldots, \mathbf{N}^{\lambda_\mathrm{iid}} = \underline{\mathcal{N}}^{r+1}, \ldots, \underline{\mathcal{N}}^{r+\lambda_\mathrm{iid}}$
  $r = r + \lambda_\mathrm{iid}$
  */* generate $\lambda_\mathrm{iid}$ offspring independently */*
  $i \leftarrow 0$ *(offspring counter)*
  **while** $i < \lambda_\mathrm{iid}$ **do**
    $i \leftarrow i + 1$
    $\underline{X}_k^i = \underline{X}_k + \sigma_k \mathbf{N}^i$
  **end while**
  **if** selective **then**
    $(\underline{X}_k^1, \ldots, \underline{X}_k^{\lambda_\mathrm{iid}}), \pi = \mathrm{argsort}(f(\underline{X}_k^1), \ldots, f(\underline{X}_k^{\lambda_\mathrm{iid}}))$
  **else**
    $\pi = id$
  **end if**
  */* mirror $\lambda_\mathrm{m}$ offspring */*
  **while** $i < \lambda_\mathrm{iid} + \lambda_\mathrm{m}$ **do**
    $i \leftarrow i + 1$
    */* dependent sample with(out) new length $\|\underline{\mathcal{N}}^r\|$ */*
    **if** i.i.d. lengths **then**
      r = r+1;
      $\underline{X}_k^i = \underline{X}_k - \sigma_k \frac{\|\underline{\mathcal{N}}^r\|}{\|\mathbf{N}^{\pi^{-1}(2\lambda_\mathrm{iid}+1-i)}\|} \mathbf{N}^{\pi^{-1}(2\lambda_\mathrm{iid}+1-i)}$
    **else**
      $\underline{X}_k^i = \underline{X}_k - \sigma_k \mathbf{N}^{\pi^{-1}(2\lambda_\mathrm{iid}+1-i)}$
    **end if**
  **end while**
  */* weighted recombination */*
  $\underline{X}_k^{1:\lambda_\mathrm{iid}}, \ldots, \underline{X}_k^{\lambda_\mathrm{iid}:\lambda_\mathrm{iid}} =$
    $\mathrm{argsort}(f(\underline{X}_k^1), \ldots, f(\underline{X}_k^{\lambda_\mathrm{iid}-\lambda_\mathrm{m}}),$
      $\min\{f(\underline{X}_k^{\lambda_\mathrm{iid}-\lambda_\mathrm{m}+1}), f(\underline{X}_k^{\lambda_\mathrm{iid}+\lambda_\mathrm{m}})\}, \ldots,$
      $\min\{f(\underline{X}_k^{\lambda_\mathrm{iid}}), \ldots, f(\underline{X}_k^{\lambda_\mathrm{iid}+1})\})$
  $\underline{X}_{k+1} = \sum_{i=1}^{\mu} w_i \underline{X}_k^{i:\lambda_\mathrm{iid}}$
  $\sigma_{k+1} = \mathrm{update}(\sigma_k, f(\underline{X}_k^1), \ldots, f(\underline{X}_k^{\lambda_\mathrm{iid}}))$
  $k \leftarrow k + 1$ iteration counter
**end while**

## 14.7 Appendix (proofs)

**14.7.1 Lemma**
*For the scale-invariant $(\mu/\mu_w, \lambda)$-ES minimizing spherical functions, the probability distribution of the vector*
$(\mathcal{Z}[1:\lambda], \dots \mathcal{Z}[\lambda:\lambda])$ *is independent of* $\underline{X}_k$ *and equals*

$$(\mathcal{Z}[1:\lambda], \dots, \mathcal{Z}[\lambda:\lambda]) = \text{argsort}\{h_\sigma(\underline{\mathcal{N}}^1), \dots, h_\sigma(\underline{\mathcal{N}}^\lambda)\} \tag{14.17}$$

*where* $h_\sigma(\underline{x}) = 2[\underline{x}]_1 + \sigma\|\underline{x}\|^2$, *where* $(\underline{\mathcal{N}}^i)_{1\le i\le\lambda}$ *are* $\lambda$ *independent standard multivariate normal distribution.*

*Proof.* At iteration $k$, starting from $\underline{X}_k$, the distribution of the selected steps is determined by the ranking of

$$(\|\underline{X}_k + \sigma\|\underline{X}_k\|\underline{\mathcal{N}}^i\|)_{1\le i\le\lambda} \quad .$$

Normalizing by $\|\underline{X}_k\|$ will not change the ranking such that the distribution is determined by the ranking of

$$\left\| \frac{\underline{X}_k}{\|\underline{X}_k\|} + \sigma\underline{\mathcal{N}}^i \right\| \text{ for } 1 \le i \le \lambda \quad .$$

However, since the distribution of $\underline{\mathcal{N}}^i$ is spherical, the distribution of the selected steps will be the same if we start from any vector with unit norm (like $\frac{\underline{X}_k}{\|\underline{X}_k\|}$), so WLOG the distribution will be determined by ranking $\left\|\underline{e}_1 + \sigma\underline{\mathcal{N}}^i\right\|$ for $1 \le i \le \lambda$ or since composing by $g(x) = x^2$ will not change the ranking

$$\left\|\underline{e}_1 + \sigma\underline{\mathcal{N}}^i\right\|^2 \text{ for } 1 \le i \le \lambda$$

We develop $\left\|\underline{e}_1 + \sigma\underline{\mathcal{N}}^i\right\|^2$ and obtain $1 + 2\sigma[\underline{\mathcal{N}}^i]_1 + \sigma^2\|\underline{\mathcal{N}}^i\|^2$. Ranking will not be affected if we subtract 1 and divide by $\sigma$ such that the distribution of $\underline{\mathcal{N}}^i$ is determined by the ranking with respect to $h_\sigma(\underline{\mathcal{N}}^i)$. □

**14.7.2 Theorem**
*For the $(\mu/\mu_w, \lambda)$-ES with scale-invariant step-size on $g(\|x\|)$, $g \in \mathcal{M}$,* (14.1) *and* (14.2) *hold and the convergence rate equals*

$$\text{CR}(\sigma, \underline{w}) = \frac{E\ln\left[1 + 2\sum_{i=1}^{\mu}\sigma w_i\,[\mathcal{Z}[i:\lambda]]_1 + \left\|\sum_{i=1}^{\mu}\sigma w_i\mathcal{Z}[i:\lambda]\right\|^2\right]}{2\lambda} \quad ,$$

*where* $w_i \in \mathbb{R}$ *and* $\sum_{i=1}^{\mu}|w_i| = 1$.

*Proof.* We start from

$$\|\underline{X}_{k+1}\| = \|\underline{X}_k + \sigma\|\underline{X}_k\|\sum_{i=1}^{\mu} w_i\mathcal{Z}[i:\lambda]\|$$

that we normalize by $\|\underline{X}_k\|$ and take the logarithm

$$\ln\frac{\|\underline{X}_{k+1}\|}{\|\underline{X}_k\|} = \ln\|\underline{X}_k/\|\underline{X}_k\| + \sigma\sum_{i=1}^{\mu} w_i\mathcal{Z}[i:\lambda]\| \tag{14.18}$$

Using the isotropy of the sphere function and of the multivariate normal distribution, together with the previous lemma, we find that the random variables in the RHS of the previous equation are i.i.d. and distributed as $\ln\|\underline{e}_1 + \sigma\sum_{i=1}^{\mu} w_i\mathcal{Z}[i:\lambda]\|$. Applying the Law of Large Numbers to

$$\frac{1}{\lambda k}\ln\frac{\|\underline{X}_k\|}{\|\underline{X}_0\|} = \frac{1}{\lambda k}\sum_{i=0}^{k-1}\ln\frac{\|\underline{X}_{i+1}\|}{\|\underline{X}_i\|}$$

we find thus that

$$\frac{1}{\lambda k} \ln \frac{\|\underline{X}_k\|}{\|\underline{X}_0\|} = \frac{1}{\lambda} E[\ln \|\underline{e}_1 + \sum_{i=1}^{\mu} \sigma w_i \mathbb{Z}[i:\lambda]\|]$$

We develop the convergence in the RHS of the previous equation using the identity

$$\ln \|\underline{e}_1 + u\| = \frac{1}{2} \ln \left[1 + 2u_1 + \|u\|^2\right], \text{ for } u \in \mathbb{R}^n \tag{14.19}$$

that can be obtained in a straightforward way by writing $\ln \|\underline{e}_1 + u\|$ as $\frac{1}{2} \ln \|\underline{e}_1 + u\|^2$ and developing the norm. We then obtain that (14.2) holds with the convergence rate expression given in the theorem. To obtain the convergence in expectation as defined in (14.1), we take the expectation in (14.18). For a more formal argument that the expectation exists and of the independence of the random variables $\ln \|\underline{X}_k\|/\|\underline{X}_k\|$, we refer to [205]. □

**14.7.3 Theorem**
*The convergence rate of the $(\mu/\mu_w, \lambda)$-ES with scale-invariant step size and weights $\underline{w} \in \mathbb{R}^\mu$ with $\sum_{i=1}^{\mu} |w_i| = 1$ on the class of spherical functions $g(\|x\|)$, $g \in \mathcal{M}$ satisfies*

$$\lim_{d\to\infty} d\,\mathrm{CR}\left(\frac{\sigma}{d}, w\right) = \frac{1}{\lambda}\left(\frac{\sigma^2}{2}\sum_{i=1}^{\mu} w_i^2 + \sigma \sum_{i=1}^{\mu} w_i E(\mathcal{N}_{i:\lambda})\right)$$

*where $\mathcal{N}_{i:\lambda}$ is the ith order statistic of $\lambda$ independent normal distributions with mean 0 and variance 1, i.e., the ith smallest of $\lambda$ independent variables $\mathcal{N}_i \sim N(0,1)$.*

*Proof.* Let $\underline{\mathcal{N}}^i$ be the $\lambda$ independent standard multivariate normal distributions. Let $\mathcal{P}(\lambda)$ denote the permutation of $\{1,\ldots,\lambda\}$, the distribution of log-progress can be expressed by summing over all permutations in $\mathcal{P}(\lambda)$ in the following way

$$\frac{1}{2\lambda} \ln \left[1 + 2\frac{\sigma}{d}\sum_{i=1}^{\mu} w_i [\mathbb{Z}[1:i]]_1 + \frac{\sigma^2}{d}\frac{\|\sum_{i=1}^{\mu} w_i \mathbb{Z}[1:i]\|^2}{d}\right] =$$
$$\frac{1}{2\lambda} \sum_{\pi\in\mathcal{P}(\lambda)} \ln \left[1 + 2\frac{\sigma}{d}\sum_{i=1}^{\mu} w_i [\underline{\mathcal{N}}^{\pi(i)}]_1 + \frac{\sigma^2}{d}\frac{\|\sum_{i=1}^{\mu} w_i \underline{\mathcal{N}}^{\pi(i)}\|^2}{d}\right] \cdot$$
$$1_{\{h_{\sigma/d}(\underline{\mathcal{N}}^{\pi(1)}) \le \ldots \le h_{\sigma/d}(\underline{\mathcal{N}}^{\pi(\lambda)})\}} \tag{14.20}$$

For any permutation $\pi$ and any $i$

$$h_{\sigma/d}(\underline{\mathcal{N}}^{\pi(i)}) = 2[\underline{\mathcal{N}}^{\pi(i)}]_1 + \frac{\sigma}{d}\|\underline{\mathcal{N}}^{\pi(i)}\|^2$$

such that $\lim_{d\to\infty} h_{\sigma/d}(\underline{\mathcal{N}}^{\pi(i)}) = 2[\underline{\mathcal{N}}^{\pi(i)}]_1$. Therefore,

$$1_{\{h_{\sigma/d}(\underline{\mathcal{N}}^{\pi(1)}) \le \ldots \le h_{\sigma/d}(\underline{\mathcal{N}}^{\pi(\lambda)})\}} \xrightarrow[d\to\infty]{} 1_{\{[\underline{\mathcal{N}}^{\pi(1)}]_1 \le \ldots [\underline{\mathcal{N}}^{\pi(\lambda)}]_1\}} \cdot \tag{14.21}$$

In addition, since every component of the vector $\sum_{i=1}^{\mu} w_i \underline{\mathcal{N}}^{\pi(i)}$ follows a standard normal distribution with mean zero and variance $\sum_{i=1}^{\mu} w_i^2$, we have that by the Law of Large Numbers that $\|\sum_{i=1}^{\mu} w_i \underline{\mathcal{N}}^{\pi(i)}\|^2/d$ converges to $\sum_{i=1}^{\mu} w_i^2$ and thus

$$\frac{d}{2} \ln \left[1 + 2\frac{\sigma}{d}\sum_{i=1}^{\mu} w_i [\underline{\mathcal{N}}^{\pi(i)}]_1 + \frac{\sigma^2}{d}\frac{\|\sum_{i=1}^{\mu} w_i \underline{\mathcal{N}}^{\pi(i)}\|^2}{d}\right]$$
$$\xrightarrow[d\to\infty]{} \sigma \sum_{i=1}^{\mu} w_i [\underline{\mathcal{N}}^{\pi(i)}]_1 + \frac{\sigma^2}{2}\sum_{i=1}^{\mu} w_i^2 \tag{14.22}$$

Thus injecting the limits from (14.21) and (14.22) into (14.20), we obtain

$$\frac{d}{2\lambda}\ln\left[1+2\frac{\sigma}{d}\sum_{i=1}^{\mu}w_i[\mathcal{Z}[1:i]]_1+\frac{\sigma^2}{d}\frac{\|\sum_{i=1}^{\mu}w_i\mathcal{Z}[1:i]\|^2}{d}\right]\xrightarrow[d\to\infty]{} \sum_{\pi\in\mathcal{P}(\lambda)}\frac{1}{\lambda}\left(\sigma\sum_{i=1}^{\mu}w_i[\underline{\mathcal{N}}^{\pi(i)}]_1+\frac{\sigma^2}{2}\sum_{i=1}^{\mu}w_i^2\right)1_{\{[\underline{\mathcal{N}}^{\pi(1)}]_1\le\ldots[\underline{\mathcal{N}}^{\pi(\lambda)}]_1\}} \quad (14.23)$$

In the RHS of the previous equation, we recognize the distribution of order statistics of standard normal distributions. Thus

$$\frac{d}{2\lambda}\ln\left[1+2\frac{\sigma}{d}\sum_{i=1}^{\mu}w_i[\mathcal{Z}[1:i]]_1+\frac{\sigma^2}{d}\frac{\|\sum_{i=1}^{\mu}w_i\mathcal{Z}[1:i]\|^2}{d}\right]\xrightarrow[d\to\infty]{} \frac{1}{\lambda}\left(\sigma\sum_{i=1}^{\mu}w_i\mathcal{N}^{i:\lambda}+\frac{\sigma^2}{2}\sum_{i=1}^{\mu}w_i^2\right) \quad (14.24)$$

To find the announced result, we need to obtain the limit in expectation. To do som we need to verify that the random variables are uniformly integrable. This step is quite technical and we refer to [205] for the details. □

**14.7.4 Lemma**
*In the $(\mu/\mu_w, \lambda_{iid}+\lambda_m^{rand})$-ES with scale-invariant step-size on spherical functions, the distribution of the vector of ordered steps to be recombined is given by*

$$(\mathcal{Z}[1],\ldots,\mathcal{Z}[\lambda_{iid}]) = \text{argsort}\{h_\sigma(\underline{\mathcal{N}}^1),\ldots,h_\sigma(\underline{\mathcal{N}}^{\lambda_{iid}-\lambda_m}), \min\{h_\sigma(\underline{\mathcal{N}}^{\lambda_{iid}-\lambda_m+1}),h_\sigma(-\underline{\mathcal{N}}^{\lambda_{iid}-\lambda_m+1})\},\ldots, \min\{h_\sigma(\underline{\mathcal{N}}^{\lambda_{iid}}),h_\sigma(-\underline{\mathcal{N}}^{\lambda_{iid}})\}\} \quad (14.25)$$

*where $h_\sigma(\underline{x}) = 2[x]_1+\sigma\|x\|^2$.*

*Proof.* Let $(\underline{\mathcal{N}}^i)_{1\le i\le\lambda}$ be $\lambda$ independent standard multivariate normal distributions. At iteration $k$ starting from $\underline{X}_k$, we rank the individuals $\underline{X}_k+\sigma\|\underline{X}_k\|\underline{\mathcal{N}}^i$ for $1\le i\le\lambda_{iid}-\lambda_m$ and the best of the mirrored, unmirrored pairs for the $\lambda_m$ last individuals, i.e., we rank $\|\underline{X}_k+\sigma\|\underline{X}_k\|\underline{\mathcal{N}}^i\|$ for $1\le i\le\lambda_{iid}-\lambda_m$ with $\min\{\|\underline{X}_k+\sigma\|\underline{X}_k\|\underline{\mathcal{N}}^i\|,\|\underline{X}_k-\sigma\|\underline{X}_k\|\underline{\mathcal{N}}^i\|\}$ for $i=\lambda_{iid}-\lambda_m+1,\ldots,\lambda_{iid}$. Using the same arguments as in Lemma 14.3.1, we find that the ranking does not change if we normalize by $\|\underline{X}_k\|$ and if we start from $\underline{e}_1$ such that the distribution is determined by the ranking of $\|\underline{e}_1+\sigma\underline{\mathcal{N}}^i\|$ for $1\le i\le\lambda_{iid}-\lambda_m$ and $\min\{\|\underline{e}_1+\sigma\underline{\mathcal{N}}^i\|,\|\underline{e}_1-\sigma\underline{\mathcal{N}}^i\|\}$ for $i=\lambda_{iid}-\lambda_m+1,\ldots,\lambda_{iid}$. As in Lemma 14.3.1, we square the terms and develop them to find that the distribution is determined by the ranking according to $h_\sigma$ as given in (14.7). □

**14.7.5 Lemma**
*In the $(\mu/\mu_w, \lambda_{iid}+\lambda_m^{sel})$-ES with scale-invariant step-size on spherical functions, the distribution of the vector of ordered steps to be recombined is given by*

$$(\mathcal{Z}[1],\ldots,\mathcal{Z}[\lambda_{iid}]) = \text{argsort}\{h_\sigma(\underline{Y}_1^{sel}),\ldots,h_\sigma(\underline{Y}_{\lambda_{iid}}^{sel})\}\ , \quad (14.26)$$

*where $\underline{Y}_i^{sel}$ is defined in* (14.8) *and* (14.9).

*Proof.* As in Lemma 14.3.1 and Lemma 14.3.4, the ranking can be done normalizing by $\underline{X}_k$ and starting from $\underline{e}_1$. Thus, it follows from the way we have defined $\underline{Y}_i^{\text{sel}}$ that the distribution of the vector of ordered steps is determined by (14.10). □

**14.7.6 Lemma**
*In the $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{sel}})$-ES with scale-invariant step-size on spherical functions, the distribution of the vector of ordered steps to be recombined is given by*

$$(\mathbb{Z}[1], \ldots, \mathbb{Z}[\lambda_{\text{iid}}]) \quad = \quad \text{argsort}\{h_\sigma(\underline{Y}_1^{\text{sel}}), \ldots, h_\sigma(\underline{Y}_{\lambda_{\text{iid}}}^{\text{sel}})\} \ , \quad (14.27)$$

*where $\underline{Y}_i^{\text{sel}}$ is defined in* (14.11) *and* (14.12).

*Proof.* As in Lemma 14.3.1 and Lemma 14.3.4, the ranking can be done normalizing by $\underline{X}_k$ and starting from $\underline{e}_1$. Thus, it follows from the way we have defined $\underline{Y}_i^{\text{sel}}$ that the distribution of the vector of ordered steps is determined by (14.10). □

**14.7.7 Theorem**
*The convergence rate of the $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{sel}})$-ES and $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{rand}})$-ES and $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{rand}})$-ES with independent length equals*

$$\text{CR}(\sigma, \underline{w}) = \frac{E \ln \left[ 1 + 2 \sum_{i=1}^{\mu} \sigma w_i \left[\mathbb{Z}[i]\right]_1 + \left\| \sum_{i=1}^{\mu} \sigma w_i \mathbb{Z}[i] \right\|^2 \right]}{2(\lambda_{\text{iid}} + \lambda_{\text{m}})} \ ,$$

*where $w_i \in \mathbb{R}$ and $\sum_{i=1}^{\mu} |w_i| = 1$ and the distribution of the random vector $(\mathbb{Z}[1], \ldots, \mathbb{Z}[\lambda_{\text{iid}}])$ are defined in Lemma 14.3.4, 14.3.5 and 14.3.6 respectively.*

*Proof.* The proof is similar to the proof of Theorem 14.3.2 injecting the distribution of the random vectors $(\mathbb{Z}[1], \ldots, \mathbb{Z}[\lambda_{\text{iid}}])$ for the different algorithms. □

**14.7.8 Theorem**
*The convergence rate of the $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{rand}})$, $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{sel}})$ (with or without independent length) with scale-invariant step size and weights $\underline{w} \in \mathbb{R}^{\mu}$ with $\sum_{i=1}^{\mu} |w_i| = 1$ on the class of spherical functions $g(\|x\|), g \in \mathcal{M}$ satisfies*

$$\lim_{d \to \infty} d \, \text{CR}\left(\frac{\sigma}{d}, w\right) = \frac{1}{\lambda_{\text{iid}} + \lambda_{\text{m}}} \left( \frac{\sigma^2}{2} \sum_{i=1}^{\mu} w_i^2 + \sigma \sum_{i=1}^{\mu} w_i E(Z_i) \right)$$

*where the distribution of $Z_i$ is given in* (14.14) *for the $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{rand}})$ and in* (14.15) *for the $(\mu/\mu_{\text{w}}, \lambda_{\text{iid}} + \lambda_{\text{m}}^{\text{sel}})$ with or without independent length for the mirroring.*

*Proof.* The proof follows the same lines as the proof for the $(\mu/\mu_{\text{w}}, \lambda)$-ES (Theorem 14.3.3). □

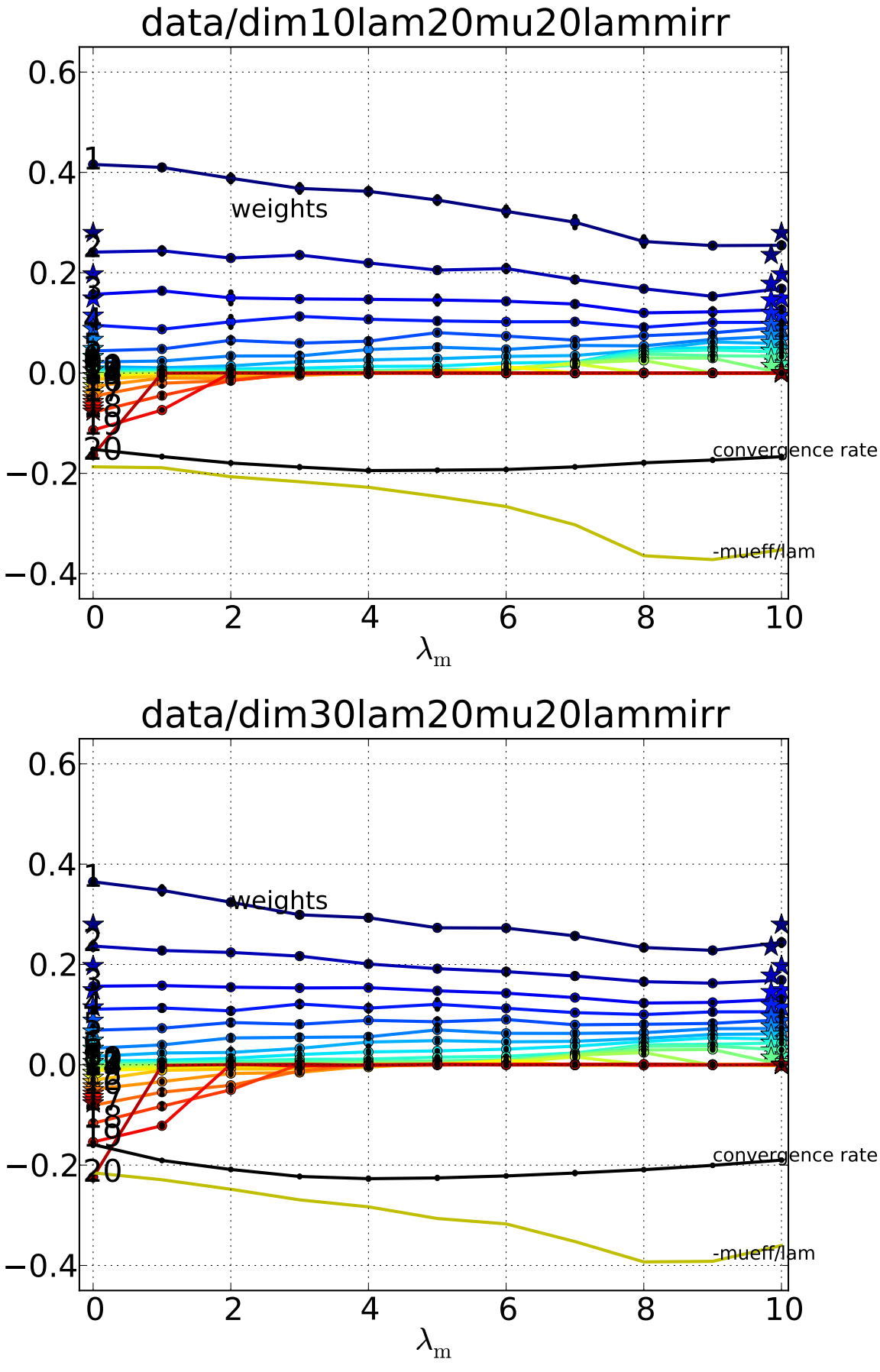



Figure 14.8: Optimal weights found by UH-CMA-ES

# Chapter 15

# Optimal $\mu$-Distributions for the Hypervolume Indicator for Problems With Linear Bi-Objective Fronts: Exact and Exhaustive Results

Dimo Brockhoff

*Published in Simulated Evolution and Learning, LNCS(6457):24-34, Springer 2010*

*To simultaneously optimize multiple objective functions, several evolutionary multiobjective optimization (EMO) algorithms have been proposed. Nowadays, often set quality indicators are used when comparing the performance of those algorithms or when selecting "good" solutions during the algorithm run. Hence, characterizing the solution sets that maximize a certain indicator is crucial—complying with the optimization goal of many indicator-based EMO algorithms. If these optimal solution sets are upper bounded in size, e.g., by the population size $\mu$, we call them* optimal $\mu$-distributions*. Recently, optimal $\mu$-distributions for the well-known hypervolume indicator have been theoretically analyzed, in particular, for bi-objective problems with a linear Pareto front. Although the exact optimal $\mu$-distributions have been characterized in this case, not all possible choices of the hypervolume's reference point have been investigated. Moreover, some of the results rely on a lower bound for the reference point in order to ensure the extremes of the front in the optimal $\mu$-distributions. In this report, we revisit the previous results and rigorously characterize the optimal $\mu$-distributions also for all other reference point choices. In this sense, our characterization is now exhaustive as the result holds for any linear Pareto front and for any choice of the reference point and the optimal $\mu$-distributions turn out to be always unique in those cases. We also prove a tight lower bound (depending on $\mu$) such that choosing the reference point above this bound ensures the extremes of the Pareto front to be always included in optimal $\mu$-distributions.*

## 15.1 Introduction

Many evolutionary multiobjective optimization (EMO) algorithms have been proposed to tackle optimization problems with multiple objectives. The most recent ones employ quality indicators within their selection in order to (i) directly incorporate user preferences into the search [379, 24] and/or to (ii) avoid cyclic behavior of the current population [353, 383]. In particular the hypervolume indicator

[381] is of interest here and due to its refinement property [383] employed in several EMO algorithms [61, 198, 43]. The hypervolume indicator assigns a set of solutions the "size of the objective value space which is covered" and at the same time is bounded by the indicator's reference point [381]. Although maximizing the hypervolume indicator, according to its refinement property, results in finding Pareto-optimal solutions only [150], the question arises which of these points are favored by hypervolume-based algorithms. In other words, we are interested in the optimization goal of hypervolume-based algorithms with a fixed population size $\mu$, i.e., in finding a set of $\mu$ solutions with the highest hypervolume indicator value among all sets with $\mu$ solutions. Also in performance assessment, the hypervolume is used quite frequently [384]. Here, knowing the set of points maximizing the hypervolume is crucial as well. On the one hand, it allows to evaluate whether hypervolume-based algorithms really converge towards their optimization goal on certain test functions. On the other hand, only the knowledge of the best hypervolume value achievable with $\mu$ solutions allows to compare algorithms in an absolute manner similar to the state-of-the-art approach of benchmarking single-objective continuous optimization algorithms in the horizontal-cut view scenario, see [176, appendix] for details.

Theoretical investigations of the sets of $\mu$ points maximizing the hypervolume indicator—also known under the term of *optimal $\mu$-distributions* [25]—have been started only recently. Although quite strong, i.e., very general, results on optimal $\mu$-distributions are known [25, 84], most of them are approximation or limit results in order to study a wide range of problem classes. The only exact results consider problems with very specific Pareto fronts, namely linear fronts that can be described by a function $f : x \in [x_{\min}, x_{\max}] \mapsto \alpha x + \beta$ where $\alpha < 0$ and $\beta \in \mathbb{R}$ in the bi-objective case [60, 137, 25] or fronts that can be expressed as $f : x \in [1, c] \mapsto c/x$ with $c > 1$ [156].

The main aim of this report is to revisit the results on optimal $\mu$-distributions for bi-objective problems with linear Pareto fronts and to consider all conditions under which the exact optimal $\mu$-distributions have not been characterized yet. The result is both exact and exhaustive, in the sense that a single formula is proven that characterizes the unique optimal $\mu$-distribution for any choice of the hypervolume indicator's reference point and for any $\mu \geq 2$, covering also the previously known cases. It turns out that the specific case of $\mu = 2$ complies with a previous results of [25] and that for all linear front shapes, the optimal $\mu$-distributions are always unique.

Before we present our results in Sec. 15.5–15.7, we introduce basic notations and definitions in Sec. 15.2, define and discuss the problem of finding optimal $\mu$-distributions in Sec. 15.3 in more detail, and give an extensive overview of the known results in Sec. 15.4.

## 15.2 Preliminaries

Without loss of generality (w.l.o.g.), we consider bi-objective minimization problems where a vector-valued function $\mathcal{F} : X \to \mathbb{R}^2$ has to be minimized with respect to the weak Pareto dominance relation $\preceq$. We say a solution $x \in X$ is weakly dominating another solution $y \in X$ ($x \preceq y$) iff $\mathcal{F}_1(x) \leq \mathcal{F}_1(y)$ and $\mathcal{F}_2(x) \leq \mathcal{F}_2(y)$ where $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2)$. We also say $x \in X$ is dominating $y \in X$ ($x \prec y$) if $x \preceq y$ but $y \not\preceq x$. The set of nondominated solutions is the so-called Pareto set $\mathcal{P}_s = \{x \in X \,|\, \nexists y \in X : y \prec x\}$ and its image $\mathcal{F}(\mathcal{P}_s)$ in objective space is called Pareto front. Note that, to keep things simple, we make an abuse of terminology throughout the paper and use the term *solution* both for a point $x$ in the decision space $X$ and for its corresponding objective vector $\mathcal{F}(x) \in \mathbb{R}^2$. Moreover, we also define the orders $\preceq$ and $\prec$ on objective vectors.

In order to optimize multiobjective optimization problems like the bi-objective ones considered here, several recent EMO algorithms aim at optimizing the *hypervolume indicator* [381], a set quality indicator $I_H(A, r)$ that assigns a set $A$ the Lebesgue measure $\lambda$ of the set of solutions that are weakly dominated by solutions in $A$ but that at the same time weakly dominate a given reference point $r \in \mathbb{R}^2$, see Fig. 15.1:

$$I_H(A, r) = \lambda\left(\{z \in \mathbb{R}^2 \,|\, \exists a \in A : f(a) \preceq z \preceq r\}\right) \tag{15.1}$$

The hypervolume indicator has the nice property of being a refinement of the Pareto dominance relation [383]. This means that maximizing the hypervolume indicator is equivalent to obtaining solutions in the Pareto set only [150]. However, it is more interesting to know *where* the solutions maximizing the hypervolume lie on the Pareto front if we restrict the size of the sets $A$ to let us say, the population size $\mu$. This set of $\mu$ points maximizing the hypervolume indicator among all sets of $\mu$ points is known under the term *optimal $\mu$-distribution* [25] and finding an optimal $\mu$-distribution coincides with the optimization goal of hypervolume-based algorithms with fixed population size.

To investigate optimal $\mu$-distributions in this report, we assume the Pareto front to be given by a function $f : \mathbb{R} \to \mathbb{R}$ and two values $x_{\min}, x_{\max} \in \mathbb{R}$ such that all points on the Pareto front have the form $(x, f(x))$ with $x \in [x_{\min}, x_{\max}]$. In case of a linear Pareto front, $f(x) = \alpha x + \beta$ for $\alpha, \beta \in \mathbb{R}$, see Fig. 15.1 for an example. W.l.o.g, we assume that $x_{\min} = 0$ and $\beta > 0$ in the remainder of the paper—otherwise, a simple linear transformation brings us back to this case. Moreover, $\alpha < 0$ follows from minimization. Note also that under not too strong assumptions on the Pareto front, and in particular for linear fronts, optimal $\mu$-distributions always exist, see [25].

## 15.3 Problem Statement

In case of a linear Pareto front described by the function $f(x) = \alpha x + \beta$ ($\alpha < 0, \beta \in \mathbb{R}$), finding the optimal $\mu$-distribution for the hypervolume indicator with reference point $r = (r_1, r_2)$ can be written as finding the minimum of the function

$$\begin{aligned} I_H(x_1, \ldots, x_\mu) &= \sum_{i=1}^{\mu} (x_{i+1} - x_i)\,(f(x_0) - f(x_i)) = \sum_{i=1}^{\mu} (x_{i+1} - x_i)\,(\alpha x_0 - \alpha x_i) \\ &= \alpha \sum_{i=1}^{\mu} \left[ (x_i)^2 + x_0 x_{i+1} - x_0 x_i - x_i x_{i+1} \right] \qquad (15.2) \\ &\text{with } x_{\min} \le x_i \le x_{\max} \text{ for all } 1 \le i \le \mu \end{aligned}$$

where we define $x_{\mu+1} = r_1$ and $x_0 = f^{-1}(r_2)$ [25], Fig. 15.1. According to [25], we denote the $x$-values of the optimal $\mu$-distribution, maximizing (15.2), as $x_1^\mu \ldots x_\mu^\mu$. Although the term in (15.2) is quadratic in the variables $x_0, \ldots, x_{\mu+1}$, and therefore, in principle, solvable analytically, the restrictions of the variables to the interval $[x_{\min}, x_{\max}]$ makes it difficult to solve the problem. In the following, we therefore investigate the minima of (15.2) depending on the choice of $r_1$ and $r_2$ with another approach: we use the necessary condition for optimal $\mu$-distributions of [25, Proposition 1] and apply it to linear fronts while the restriction of the variables to $[x_{\min}, x_{\max}]$ are handled "by hand".

## 15.4 Overview of Recent and New Results

Characterizing optimal $\mu$-distributions for the hypervolume indicator has been started only recently but the number of results is already quite extensive, see for example [150, 60, 137, 25, 24, 156, 26, 84, 43]. Here, we restate, to the best of our knowledge, all previous results that relate to linear Pareto fronts and point out which problems are still open.

Besides the proof that maximizing the hypervolume indicator yields Pareto-optimal solutions [150], the authors of [60] and [137] were the first to investigate optimal $\mu$-distributions for linear fronts. Under the assumption that the extreme points $(0, \beta)$ and $(x_{\max}, 0)$ are included in the optimal $\mu$-distribution, it was shown for linear fronts with $\alpha = -1$ that neighbored points within a set maximizing the hypervolume are equally spaced. However, the result does not state where the leftmost and rightmost point of the optimal $\mu$-distribution have to be placed in order to maximize the hypervolume and it has been shown later [25] that the assumption about the extreme points does not always hold.

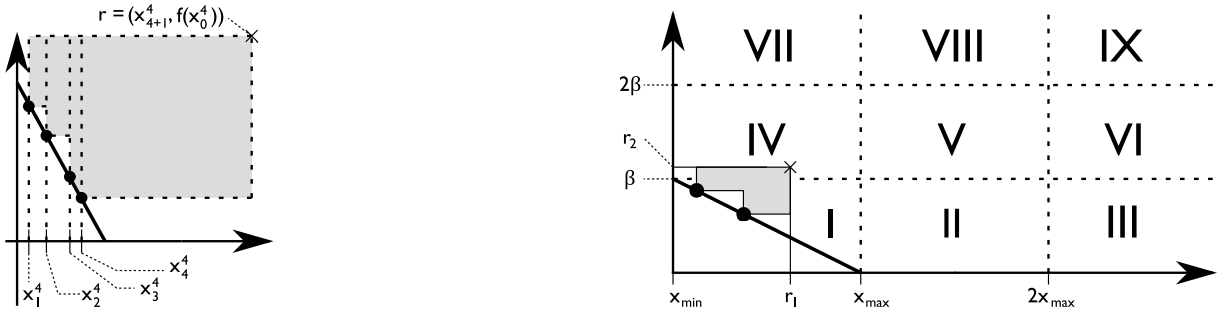



Figure 15.1: **Left:** Illustration of the hypervolume indicator $I_H(A, r)$ (gray area). **Right:** Optimal $\mu$-distributions and the choice of the reference point for linear fronts of shape $y = \alpha x + \beta$. Up-to-now, theoretical results are only known if the reference point is chosen within the regions I and IX [25]. Exemplary, the optimal 2-distribution (circles) is shown when choosing the reference point (cross) within region IV.

The first results without assuming the positions of the leftmost and rightmost point have been proven in [25] where the result is based on a more general necessary condition about optimal $\mu$-distributions for the hypervolume indicator. In particular, [25] presents the exact distribution of $\mu$ points maximizing the hypervolume indicator when the reference point is chosen close to the Pareto front (region I in Fig. 15.1, cp. [25, Theorem 5]) or far away from the front (region IX in Fig. 15.1, cp. [25, Theorem 6]). In the former case, both extreme points of the front do not dominate the reference point and the (in this case unique) optimal $\mu$-distribution reads

$$x_i^\mu = f^{-1}(r_2) + \frac{i}{\mu + 1} \cdot (r_1 - f^{-1}(r_2)) \ . \tag{15.3}$$

In the latter case, the reference point is chosen far enough such that—independent of the reference point and $\mu$—both extreme points are included in an optimal $\mu$-distribution[1] and the (again unique) optimal $\mu$-distributions can be expressed as

$$x_i^\mu = x_{\min} + \frac{i-1}{\mu - 1} (x_{\max} - x_{\min}) \ . \tag{15.4}$$

Note that the region IX, corresponding to choices of the reference point within Theorem 6 of [25] does not depend on $\mu$ but on a lower bound on the reference point to ensure that both extremes are included in the optimal $\mu$-distribution. Recently, a limit result has been proven [26] which shows that the lower bound of [25, Theorem 6] converges to the nadir point[2] if $\mu$ goes to infinity but the result does not state how fast (in $\mu$) the nadir point is approached. Clearly, choosing the reference point within the other regions II–VIII in Fig. 15.1 is possible and the question arises how the reference point influences the optimal $\mu$-distributions in these uninvestigated cases as well. The answer to this question is the main focus of this report.

## 15.5 If the Reference Point is Dominated by Only the Right Extreme

As a first new result, we consider choosing the reference point within the regions II or III of Fig. 15.1. Here, the left extreme cannot be included in an optimal $\mu$-distribution as it is never dominating the reference point and thereby always has a zero hypervolume contribution. Thus, the proof of the optimal $\mu$-distribution has to consider only the restrictions of the $\mu$ points at the right extreme. Moreover, the uniqueness of the optimal $\mu$-distribution in the cases II and III follows directly from case I.

[1] Which is proven to be true for $r_1 > 2x_{\max}$ and $r_2 > 2\beta$ in another general theorem [25].
[2] In case of a linear front as defined above, the nadir point equals $n = (x_{\max}, f(x_{\min}))$.

**15.5.1 Theorem**
*Given $\mu \in \mathbb{N}_{\geq 2}$, $\alpha \in \mathbb{R}_{<0}$, $\beta \in \mathbb{R}_{>0}$, and a linear Pareto front $f(x) = \alpha x + \beta$ within $[0, x_{\max} = -\frac{\beta}{\alpha}]$. If $r_2 \leq \beta$ and $r_1 \geq x_{\max}$ (cases II and III), the unique optimal $\mu$-distribution $(x_1^\mu, \ldots, x_\mu^\mu)$ for the hypervolume indicator $I_H$ with reference point $(r_1, r_2)$ can be described by*

$$x_i^\mu = f^{-1}(r_2) + \frac{i}{\mu+1}\left(\min\left\{r_1, \frac{\mu+1}{\mu}x_{\max} - \frac{f^{-1}(r_2)}{\mu}\right\} - f^{-1}(r_2)\right) \quad . \tag{15.5}$$

*Proof.* According to (15.3) and assuming no restrictions of the solutions on the linear front $\alpha x + \beta$ with $x \in \mathbb{R}$, the optimal $\mu$-distribution would be given by $x_i^\mu = f^{-1}(r_2) + \frac{i}{\mu+1} \cdot (r_1 - f^{-1}(r_2))$ where the $x_i^\mu$ are possibly lying outside the interval $[0, x_{\max}]$. However, as long as $r_1$ is chosen such that $x_\mu^\mu \leq x_{\max}$, we can use (15.3) for describing the optimal $\mu$-distributions, i.e., in the case that

$$\begin{aligned} x_\mu^\mu = f^{-1}(r_2) + \frac{\mu}{\mu+1} \cdot (r_1 - f^{-1}(r_2)) \leq x_{\max} &\Leftrightarrow \frac{f^{-1}(r_2)}{\mu+1} + \frac{\mu}{\mu+1}r_1 \leq x_{\max} \\ &\Leftrightarrow r_1 \leq \frac{\mu+1}{\mu}x_{\max} - \frac{f^{-1}(r_2)}{\mu}\left(= \frac{-r_2 - \beta\mu}{\alpha\mu}\right) \quad . \end{aligned} \tag{15.6}$$

With larger $r_1$, the optimal $\mu$-distribution does not change any further (only the hypervolume contribution of $x_\mu^\mu$ increases linearly with $r_1$), i.e., we can rewrite (15.3) as (15.5). □ □

The previous theorem allows us also a more precise statement of when the right extreme is included in optimal $\mu$-distributions than the statement in [25].

**15.5.2 Corollary**
*In case that $r_2 \leq \beta$ and $r_1 \geq \frac{\mu+1}{\mu}x_{\max} - \frac{f^{-1}(r_2)}{\mu}$, the right extreme point $(x_{\max}, 0)$ is included in all optimal $\mu$-distributions for the front $\alpha x + \beta$.* □

Note that the choice of $r_1$ to guarantee the right extreme in optimal $\mu$-distributions depends both on $\mu$ and $r_2$ here whereas the (not so tight) bound for $r_1$ to ensure the right extreme proven in [25] equals $2x_{\max}$. This is independent of $\mu$ and coincides with the new (tighter) result if $\mu = 2$ and $r_2 = \beta$. Figure 15.2 illustrates the region for which, if the reference point is chosen within, the right extreme is always included in an optimal $\mu$-distribution. Compare also to the old result of [25] which states this inclusion of the right extreme only in case the reference point is chosen in region IX of Fig. 15.1. The description of the line $y = -\alpha\mu x - \mu\beta$ where choosing the reference point to the right of it ensures the right extreme in the optimal $\mu$-distribution results from writing $r_2$ within $r_1 = \frac{\mu+1}{\mu}x_{\max} - \frac{f^{-1}(r_2)}{\mu}$ as a function of $r_1$.

## 15.6 If the Reference Point is Dominated by Only the Left Extreme

Obviously, the two cases IV and VII of Fig. 15.1 are symmetrical to the cases II and III where mainly the left extreme and the reference point's coordinate $r_2$ take the roles of the right extreme and the coordinate $r_1$ respectively from the previous proof.

**15.6.1 Theorem**
*Given $\mu \in \mathbb{N}_{\geq 2}$, $\alpha \in \mathbb{R}_{<0}$, $\beta \in \mathbb{R}_{>0}$, and a linear Pareto front $f(x) = \alpha x + \beta$ within $[0, x_{\max} = -\frac{\beta}{\alpha}]$. If $r_1 \leq x_{\max}$ and $r_2 \geq \beta$ (cases IV and VII), the unique optimal $\mu$-distribution $(x_1^\mu, \ldots, x_\mu^\mu)$ for the hypervolume indicator $I_H$ with reference point $(r_1, r_2)$ can be described by*

$$x_i^\mu = f^{-1}\left(\min\left\{r_2, \tfrac{\mu+1}{\mu}\beta - \tfrac{f(r_1)}{\mu}\right\}\right) + \frac{i}{\mu+1}\left(r_1 - f^{-1}\left(\min\left\{r_2, \tfrac{\mu+1}{\mu}\beta - \tfrac{f(r_1)}{\mu}\right\}\right)\right). \tag{15.7}$$

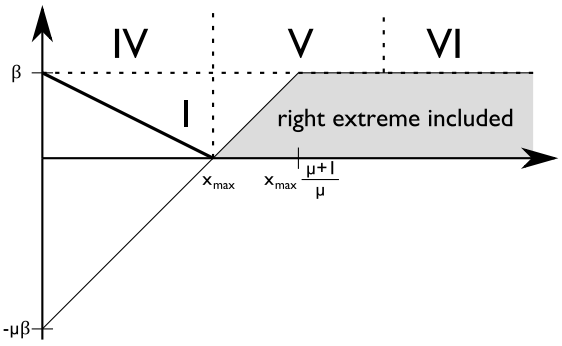



Figure 15.2: When choosing the reference point within regions II and III, we prove that the right extreme is included in optimal $\mu$-distributions if the reference point is chosen within the gray shaded area right of the line $y = -\alpha\mu x - \mu\beta$, see Corollary 15.5.2. The picture corresponds to $\mu = 2$.

*Proof.* The proof is similar to the one of Theorem 15.5.1: As in case I, we can write the optimal $\mu$-distribution according to (15.3) except that we have to ensure that $x_1^\mu \geq x_{\min} = 0$. This is equivalent to $f^{-1}(r_2) + \frac{1}{\mu+1}\left(r_1 - f^{-1}(r_2)\right) \geq 0$ or $\frac{r_2-\beta}{\alpha} + \frac{1}{\mu+1}\left(r_1 - \frac{r_2-\beta}{\alpha}\right) \geq 0$ or $\frac{r_2-\beta}{\alpha} + \frac{\alpha r_1 - r_2 + \beta}{(\mu+1)\alpha} \geq 0$. With $\alpha < 0$, this gives $(\mu+1)r_2 - (\mu+1)\beta + \alpha r_1 - r_2 + \beta \leq 0$ and finally $r_2 \leq \frac{(\mu+1)\beta - (\alpha r_1 + \beta)}{\mu} = \frac{\mu+1}{\mu}\beta - \frac{f(r_1)}{\mu}$ such that (15.3) becomes (15.7). □ □

## 15.7 General Result for All Cases I–IX

By combining the above results, we can now characterize the optimal $\mu$-distributions also for the other cases V, VI, VII, and IX and give a general description of optimal $\mu$-distributions for problems with bi-objective linear fronts, given any $\mu \geq 2$ and any meaningful choice of the reference point[3].

**15.7.1 Theorem**
*Given $\mu \in \mathbb{N}_{\geq 2}$, $\alpha \in \mathbb{R}_{<0}$, $\beta \in \mathbb{R}_{>0}$, and a linear Pareto front $f(x) = \alpha x + \beta$ within $[0, x_{\max} = -\frac{\beta}{\alpha}]$, the unique optimal $\mu$-distribution $(x_1^\mu, \ldots, x_\mu^\mu)$ for the hypervolume indicator $I_H$ with reference point $(r_1, r_2) \in \mathbb{R}^2_{>0}$ can be described by*

$$x_i^\mu = f^{-1}(F_l) + \frac{i}{\mu+1}\left(F_r - f^{-1}(F_l)\right) \tag{15.8}$$

*for all $1 \leq i \leq \mu$ where*

$$F_l = \min\{r_2, \frac{\mu+1}{\mu}\beta - \frac{1}{\mu}f(r_1), \frac{\mu}{\mu-1}\beta\} \text{ and}$$
$$F_r = \min\{r_1, \frac{\mu+1}{\mu}x_{\max} - \frac{1}{\mu}f^{-1}(r_2), \frac{\mu}{\mu-1}x_{\max}\} \ .$$

*Proof.* Again, the optimal $\mu$-distribution would be given by (15.3) if we prolongate the front linearly outside the interval $[x_{\min}, x_{\max}]$ and therefore, no restrictions on the $x_i^\mu$ would hold. However, the points $x_i^\mu$ are restricted to $[x_{\min}, x_{\max}]$ and therefore (since we assume $x_i^\mu < x_{i+1}^\mu$) we have to ensure that both

[3]Choosing the reference point such that it weakly dominates a Pareto-optimal point does not make sense as no feasible solution would have a positive hypervolume.

$x_1^\mu \geq x_{\min} = 0$ and $x_\mu^\mu \leq x_{\max} = -\beta/\alpha$ hold. According to the above proofs, the former is equivalent to

$$r_2 \leq \frac{\mu+1}{\mu}\beta - \frac{f(r_1)}{\mu} \tag{15.9}$$

and the latter is equivalent to

$$r_1 \leq \frac{\mu+1}{\mu}x_{\max} - \frac{f^{-1}(r_2)}{\mu} \tag{15.10}$$

however, with restrictions on $r_1$ ($r_1 \leq x_{\max}$) and $r_2 \leq \beta$ respectively which we do not have here. As long as both (15.9) and (15.10) hold as in the white area in Fig. 15.3, i.e., no constraint is violated, (15.3) can be used directly to describe the optimal $\mu$-distribution as in region I. To cover all other cases, we could, at first sight, simply combine the results for the cases II, III, IV, and VII from above and use

$$F_l^* = \min\left\{r_2, \frac{\mu+1}{\mu}\beta - \frac{f(r_1)}{\mu}\right\} \text{ and } F_r^* = \min\left\{r_1\frac{\mu+1}{\mu}x_{\max} - \frac{f^{-1}(r_2)}{\mu}\right\}$$

as the extremes influencing the set $x_i^{\mu,*} = F_l^* + \frac{i}{\mu+1}\left(F_r^* - F_l^*\right)$. However, $r_1$ and $r_2$ are unrestricted and thus, $F_l^*$ and $F_r^*$ can become too large such that the points $x_i^{\mu,*}$ lie outside the feasible front part $[x_{\min}, x_{\max}]$. To this end, we compute where the two constraints (15.9) and (15.10) meet, i.e., what is the smallest possible reference point that results in having both extremes in the optimal $\mu$-distribution. This point is depicted as the lower left point of the dark gray area in Fig. 15.3.

By combining the equalities in (15.9) and (15.10) which is equivalent to $r_2 = -\alpha\mu r_1 - \beta\mu$ (see end of Sec. 15.5), we obtain

$$r_2 = \tfrac{\mu+1}{\mu}\beta - \tfrac{f(r_1)}{\mu} = \tfrac{\mu+1}{\mu}\beta - \tfrac{\alpha r_1+\beta}{\mu} = -\alpha\mu r_1 - \beta\mu \text{ or } r_1 = -\tfrac{\beta}{\alpha}\tfrac{\mu}{\mu-1} = \tfrac{\mu}{\mu-1}x_{\max}$$

and thus $r_2 = \frac{\mu+1}{\mu}\beta - \frac{f(\frac{\mu}{\mu-1}x_{\max})}{\mu} = \frac{\mu}{\mu-1}\beta$. Hence, if we choose the reference point $r = (r_1, r_2)$ such that $r_1 \geq \frac{\mu}{\mu-1}x_{\max}$ and $r_2 \geq \frac{\mu}{\mu-1}\beta$, both extremes will be included in the optimal $\mu$-distribution $x_i^\mu = F_l^{\text{extr}} + \frac{i}{\mu+1}\left(F_r^{\text{extr}} - F_l^{\text{extr}}\right)$ with $F_l^{\text{extr}} = \frac{\mu}{\mu-1}\beta$ and $F_r^{\text{extr}} = \frac{\mu}{\mu-1}x_{\max}$. With this result, we know that, independent of $r_2$, the right extreme is included if $r_1 \geq \frac{\mu}{\mu-1}x_{\max}$ (if the leftmost extreme is not included, $r_2$ must be smaller than $\frac{\mu}{\mu-1}\beta$ and in this case $r_1 \geq \frac{\mu+1}{\mu}x_{\max}$ ensures that it is also greater or equal to $\frac{\mu+1}{\mu}x_{\max} - \frac{f^{-1}(r_2)}{\mu}$). The same can be said for the left extreme, which is included in an optimal $\mu$-distribution whenever $r_2 \geq \frac{\mu}{\mu-1}\beta$. The optimal $\mu$-distribution for those cases are the same than the optimal $\mu$-distributions if we restrict $r_1$ and $r_2$ to be at most $\min\{\frac{\mu+1}{\mu}x_{\max} - \frac{1}{\mu}f^{-1}(r_2), \frac{\mu}{\mu-1}x_{\max}\}$, and $\min\{\frac{\mu+1}{\mu}\beta - \frac{1}{\mu}f(r_1), \frac{\mu}{\mu-1}\beta\}\}$ respectively, i.e., to the cases where the reference point is lying on the boundary of the white region of Fig. 15.3 and having one or even both extremes included in the optimal $\mu$-distributions. In those cases, (15.3) can be used again for characterizing the optimal $\mu$-distribution as the constraints on the $x_i^\mu$ are fulfilled. Using the mentioned restrictions on $r_1$ and $r_2$ results in the theorem. □ □

Note that the previous proof gives a tighter bound for how to choose the reference point $r = (r_1, r_2)$ in order to obtain the extremes in comparison to the old result in [25]: The former result states that whenever $r_1$ is chosen strictly larger than $2x_{\max}$ and $r_2$ is chosen strictly larger than $2\beta$, both extremes are included in an optimal $\mu$-distribution in the case of a linear Pareto front. This bound holds for every $\mu \geq 2$ but the previous theorem precises this bound to $r_1 \geq \frac{\mu+1}{\mu}x_{\max}$ and $r_2 \geq \frac{\mu+1}{\mu}\beta$ for a given $\mu$ which coincides with the old bound for $\mu = 2$ but is the closer to the nadir point $(x_{\max}, \beta)$, the larger $\mu$ gets—a result that has been previously shown as a limit result for arbitrary Pareto fronts [26].

Last, we want to note that, though the two equations (15.8) and (15.4) do not look the same at first sight, Theorem 15.7.1 complies with the characterization of optimal $\mu$-distributions given in (15.4) [25, Theorem 6] for the case IX which can be shown by simple algebra.


Figure 15.3: How to choose the reference point to obtain the extremes in optimal $\mu$-distributions: $\mu = 2$ (left) and $\mu = 4$ (right) for one and the same front $y = -x/2 + 1$.

## 15.8 Conclusions

Finding optimal $\mu$-distributions, i.e., sets of $\mu$ points that have the highest quality indicator value among all sets of $\mu$ solutions coincides with the optimization goal of indicator-based multiobjective optimization algorithms and it is therefore important to characterize them. Here, we rigorously analyze optimal $\mu$-distributions for the often used hypervolume indicator and for problems with linear Pareto fronts. The results are exhaustive in a sense that a single formula covers all possible choices of the hypervolume's reference point, including two previously proven cases. In addition to the newly covered cases, the new results show also how the choice of $\mu$ influences the fact that the extremes of the Pareto front are included in optimal $\mu$-distributions for the case of linear fronts—a fact that has been only shown before by a lower bound result of choosing the reference point and not exact as here. The proofs also show that the optimal $\mu$-distributions for problems with linear Pareto fronts are, given a $\mu \geq 2$ and a certain choice of the reference point, always unique.

Besides being the first exhaustive theoretical investigation of optimal $\mu$-distributions for a specific front shape, the presented results are expected to have an impact in practical performance assessment as well. For the first time, it is now possible to use the exact optimal $\mu$-distribution and its corresponding hypervolume when comparing algorithms on test problems with linear fronts such as DTLZ1 [119] or WFG3 [193] for any choice of the reference point[4]. It remains future work to theoretically characterize the optimal $\mu$-distributions for test problems with other front shapes for which the optimal $\mu$-distributions can only be approximated numerically at the moment [25].

[4]Theorem 15.7.1 can be applied directly with $\alpha = -1$ and $\beta = 0.5$ (DTLZ1) or $\beta = 1$ (WFG3).

# Chapter 16

# Hypervolume-based Multiobjective Optimization: Theoretical Foundations and Practical Implications

Anne Auger, Johannes Bader, Dimo Brockhoff, Eckart Zitzler

*To appear in Theoretical Computer Science*

*In recent years, indicator-based evolutionary algorithms, allowing to implicitly incorporate user preferences into the search, have become widely used in practice to solve multiobjective optimization problems. When using this type of methods, the optimization goal changes from optimizing a set of objective functions simultaneously to the single-objective optimization goal of finding a set of $\mu$ points that maximizes the underlying indicator. Understanding the difference between these two optimization goals is fundamental when applying indicator-based algorithms in practice. On the one hand, a characterization of the inherent optimization goal of different indicators allows the user to choose the indicator that meets her preferences. On the other hand, knowledge about the sets of $\mu$ points with optimal indicator values—so-called* optimal $\mu$-distributions*—can be used in performance assessment whenever the indicator is used as a performance criterion. However, theoretical studies on indicator-based optimization are sparse. One of the most popular indicators is the weighted hypervolume indicator. It allows to guide the search towards user-defined objective space regions and at the same time has the property of being a refinement of the Pareto dominance relation with the result that maximizing the indicator results in Pareto-optimal solutions only. In previous work, we theoretically investigated the* unweighted *hypervolume indicator in terms of a characterization of optimal $\mu$-distributions and the influence of the hypervolume's reference point for general bi-objective optimization problems. In this report, we generalize those results to the case of the* weighted *hypervolume indicator. In particular, we present general investigations for finite $\mu$, derive a limit result for $\mu$ going to infinity in terms of a density of points and derive lower bounds (possibly infinite) for placing the reference point to guarantee the Pareto front's extreme points in an optimal $\mu$-distribution. Furthermore, we state conditions about the slope of the front at the extremes such that there is no finite reference point that allows to include the extremes in an optimal $\mu$-distribution—contradicting previous belief that a reference point chosen just above the nadir point or the objective space boundary is sufficient for obtaining the extremes. However, for fronts where there exists a finite reference point allowing to obtain the extremes, we show that for $\mu$ to infinity, a reference point that is slightly worse in all objectives than the nadir point is a sufficient choice. Last, we apply the theoretical results to problems of the ZDT, DTLZ, and WFG test problem suites.*

## 16.1 Introduction

Multiobjective optimization aims at optimizing several criteria simultaneously. In the last decades, evolutionary algorithms have been shown to be well-suited for those problems in practice [115, 103]. A recent trend is to use quality indicators to turn a multiobjective optimization problem into a single-objective one by optimizing the quality indicator itself. An *indicator-based algorithm* uses a specific quality indicator to assign every individual a single-objective fitness—most of the time proportional to the *indicator loss*, a measure of how much the quality indicator decreases if the corresponding individual is removed from the population. Instead of optimizing the objective functions directly, indicator-based algorithms therefore aim at finding a set of solutions that maximizes the underlying quality indicator and a fundamental question is whether these two optimization goals coincide or how they differ. In practice, the population size of indicator-based algorithms is usually finite, i.e., equal to $\mu \in \mathbb{N}$, and the optimization goal changes to finding a set of $\mu$ solutions optimizing the quality indicator[1]. We call such a set an *optimal $\mu$-distribution for the given indicator* generalizing the definition given by [25]. In this case, the additional questions arise how the number of points $\mu$ influences the optimization goal and to which set of $\mu$ objective vectors the optimal $\mu$-distribution is mapped, i.e., which search bias is introduced by changing the optimization goal. Ideally, the optimal $\mu$-distribution for an indicator only contains Pareto-optimal points and an increase in $\mu$ covers more and more points until the entire Pareto front is covered if $\mu$ approaches infinity. It is clear that in general, two different quality indicators yield a priori two different optimal $\mu$-distributions, or in other words, introduce a different search bias. This has for instance been shown experimentally by [156] for the multiplicative $\varepsilon$-indicator and the hypervolume indicator.

The hypervolume indicator and its weighted version [376] are particularly interesting indicators since they are refinements of the Pareto dominance relation [383][2]. Thus, an optimal $\mu$-distribution for these indicators contains only Pareto-optimal solutions and the set (probably unbounded in size) that maximizes the (weighted) hypervolume indicator covers the entire Pareto front [150]. Many other quality indicators do not have this fundamental property. It explains the success of the hypervolume indicator as quality indicator applied to environmental selection of indicator-based evolutionary algorithms such as ESP [194], SMS-EMOA [61], MO-CMA-ES [198], or HypE [44]. Nevertheless, it has been argued that using the (weighted) hypervolume indicator to guide the search introduces a certain bias. Interestingly, several contradicting beliefs about this bias have been reported in the literature which we will discuss later on in more detail (see Sec. 16.3). They range from stating that *convex regions may be preferred to concave regions* to the argumentation that *the hypervolume is biased towards boundary solutions*. In the light of those contradicting beliefs, a thorough investigation of the effect of the hypervolume indicator on optimal $\mu$-distributions is necessary.

Another important issue when dealing with the hypervolume indicator is the choice of the reference point, a parameter, both the unweighted and the weighted hypervolume indicator depend on. The influence of this reference point on optimal $\mu$-distributions has not been fully understood, especially for the weighted hypervolume indicator, and only rules-of-thumb exist on how to choose the reference point in practice. In particular, it could not be observed from practical investigations how the reference point has to be set to ensure to find the extremes of the Pareto front. Several authors recommend to use the corner of a space that is a little bit larger than the actual objective space as the reference point [219, 61]. For performance assessment, others recommend to use the estimated nadir point as the reference point [310, 309, 195]. Also here, theoretical investigations are highly needed to assist in practical applications.

First theoretical studies on optimal $\mu$-distributions for the (unweighted) hypervolume indicator and the choice of its reference point have been published in an earlier work by the authors [25]. The theoretical analyses resulted in a better understanding of the search bias the hypervolume indicator introduces and in theoretically founded recommendations on where to place the reference point in the

[1] Sometimes, the population size might not be fixed, e.g., when deleting all dominated solutions, but the maximum number of simultaneously considered solutions is typically upper bounded by a constant $\mu$.

[2] Other studies introduced the equivalent terms of being *compatible* or *compliant* with the Pareto dominance relation [220, 384].

case of two objectives. In particular, some beliefs about the indicator's search bias could be disproved and others confirmed, the optimal $\mu$-distributions for linear Pareto fronts were characterized exactly, and lower bounds on the reference point's objective values that allow to include the extremes of the Pareto front in certain cases have been given. Recently, a specific result of [25] has been already generalized to the weighted hypervolume indicator [24] and another exact result for specific Pareto fronts have been provided [156].

In this report, we extend *all* results by [25] to the weighted case and provide a general theory of the weighted hypervolume indicator in terms of both the inherently introduced search bias and the choice of the reference point. In particular, we

- characterize the sets of $\mu$ points that maximize the (weighted) hypervolume indicator; besides general investigations for finite $\mu$, we derive a limit result for $\mu$ going to infinity in terms of a density of points. The presented results for the weighted hypervolume indicator comply with the results for the unweighted case [25]. Furthermore, we
- investigate the influence of the reference point on optimal $\mu$-distributions, i.e., we derive lower bounds for the objective values of the reference point (possibly infinite) for guaranteeing the Pareto front's extreme points in an optimal $\mu$-distribution and investigate cases where the extremes are never contained in such a set; these results generalize the work by [25] to the weighted hypervolume indicator. In addition, we
- prove, in case the extremes can be obtained, that for any reference point dominated by the nadir point—with any small but positive distance between the two points—there is a finite number of points $\mu_0$ (possibly large in practice) such that for all $\mu > \mu_0$, the extremes are included in optimal $\mu$-distributions. Last, we
- apply the theoretical results to linear Pareto fronts [25] and to benchmark problems of the ZDT [378], DTLZ [119], and WFG [193] test problem suites resulting in recommended choices of the reference point including numerical and sometimes analytical expressions for the resulting density of points on the front.

The paper is structured as follows. First, we recapitulate the basics of the (weighted) hypervolume indicator and introduce the notations and definitions needed in the remainder of the paper (Sec. 16.2). Then, we consider the bias of the weighted hypervolume indicator in terms of optimal $\mu$-distributions. After characterizing optimal $\mu$-distributions for a finite number of solutions (Sec. 16.3.1), we derive results on the density of points if the number of points goes to infinity (Sec. 16.3.2). Section 16.4 investigates the influence of the reference point on optimal $\mu$-distributions especially on the extremes. The application of the results to test problems is presented in Sec. 16.5, and Sec. 16.6 concludes the paper.

## 16.2 The Hypervolume Indicator: General Aspects and Notations

Throughout this study we consider, without loss of generality, minimization problems where $k$ objective functions $\mathcal{F}_i : X \to Z$, $1 \leq i \leq k$ have to be minimized simultaneously. The vector function $\mathcal{F} := (\mathcal{F}_1, \ldots, \mathcal{F}_k)$ thereby maps each solution $x$ in the decision space $X$ to its corresponding objective vector $\mathcal{F}(x)$ in the objective space $\mathcal{F}(X) = Z \subseteq \mathbb{R}^k$. Furthermore, we assume that the underlying dominance structure is given by the weak Pareto dominance relation $\preceq$ which is defined between arbitrary solution pairs. We say $x \in X$ *weakly dominates* $y \in X$ if for all $1 \leq i \leq k$, $\mathcal{F}_i(x) \leq \mathcal{F}_i(y)$ and write $x \preceq y$. This weak Pareto dominance relation is generalized to sets of solutions in the following straightforward manner: we say a set $A$ of solutions weakly dominates another solution set $B$ if for all $b \in B$ there exists an $a \in A$ such that $a \preceq b$. The *Pareto(-optimal) set* $P_s$ consists of all solutions $x^* \in X$, such that there is no $x \in X$ that satisfies $x \preceq x^*$ and $x^* \not\preceq x$. The image of $P_s$ under $\mathcal{F}$ is called *Pareto(-optimal) front* or *front* for short. We

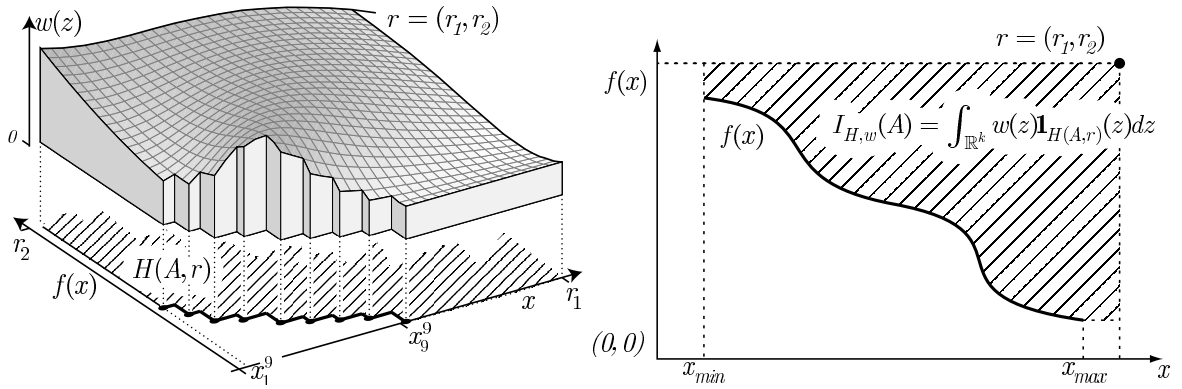



Figure 16.1: The hypervolume indicator $I_{H,w}(A)$ corresponds to the integral of a weight function $w(z)$ over the set of objective vectors that are weakly dominated by a solution set $A$ and in addition weakly dominate the reference point $r$ (hatched areas). On the left, the set $A$ consists of nine objective vectors whereas on the right, the infinite set $A$ can be described by a function $f : [x_{\min}, x_{\max}] \to \mathbb{R}$. The left-hand plot shows an example of a weight function $w(z)$, where for all objective vectors $z$ that are not dominated by $A$ or not enclosed by $r$ the function $w$ is not plotted, such that the weighted hypervolume indicator corresponds to the volume of the gray shape.

also use the weak Pareto dominance relation notation $\preceq$ among objective vectors, i.e., for two objective vectors $x = (x_1, \ldots, x_k), y = (y_1, \ldots, y_k) \in \mathbb{R}^k$ we define $x \preceq y$ if and only if for all $1 \leq i \leq k : x_i \leq y_i$.

In the following, in order to simplify notations[3], we define the indicators for *sets of objective vectors* $A \subseteq \mathbb{R}^k$ instead for *solution sets* $A' \subseteq X$ as it was already done before [376, 25]. The weighted hypervolume indicator $I_{H,w}(A, r)$ for a set of objective vectors $A \subseteq Z$ is then the weighted Lebesgue measure of the set of objective vectors weakly dominated by the solutions in $A$ that at the same time weakly dominate a so-called reference point $r \in Z$ [44][4]:

$$I_{H,w}(A, r) = \int_{\mathbb{R}^k} w(z)\mathbf{1}_{H(A,r)}(z)dz \tag{16.1}$$

where $H(A, r) := \{z \in Z \mid \exists a \in A : a \preceq z \preceq r\}$, $\mathbf{1}_{H(A,r)}(z)$ is the characteristic function of $H(A, r)$ that equals 1 iff $z \in H(A, r)$ and 0 otherwise, and $w : \mathbb{R}^k \to \mathbb{R}_{>0}$ is a strictly positive weight function integrable on any bounded set, i.e., $\int_{B(0,\gamma)} w(z)dz < \infty$ for any $\gamma > 0$, where $B(0, \gamma)$ is the open ball centered in 0 and of radius $\gamma$. In other words, we assume that the measure associated to $w$ is $\sigma$-finite[5]. Throughout the paper, the notation $I_H$ refers to the non-weighted hypervolume where the weight is 1 everywhere, and we will explicitly use the term non-weighted hypervolume for $I_H$ while the weighted hypervolume indicator $I_{H,w}$ is, for simplicity, referred to as hypervolume.

The left-hand plot of Fig. 16.1 illustrates the hypervolume $I_{H,w}$ for a bi-objective problem. The three-objective plot shows the objective values of nine points on the first two axes and the weight function $w$ on the third axis. The hypervolume indicator $I_{H,w}(A)$ for the set $A$ of nine points equals the integral of the weight function over the objective space that is weakly dominated by the set $A$ and which weakly dominates the reference point $r = (r_1, r_2)$.

In what follows, we consider bi-objective problems. The Pareto front can thus be described by a one-dimensional function $f$ mapping the image of the Pareto set under the first objective $\mathcal{F}_1$ onto the

[3]Considering an indicator on solution sets introduces the possibility of solutions that map to the same objective vector. Adding such a so-called *indifferent solution* to a solution set does not affect the set's hypervolume indicator value but the consideration of such solutions makes the text less readable if we want to state the results formally correct.

[4]Instead of a reference set as by [44], we consider one reference point only as in earlier publications [376].

[5]Several results presented in this report also hold if the weight is strictly positive almost everywhere, i.e., it can be 0 for null sets. However, we decided to consider only strictly positive weights to keep the proofs simple.

image of the Pareto set under the second objective $\mathcal{F}_2$,

$$f : x \in D \mapsto f(x) \ ,$$

where $D$ denotes the image of the Pareto set under the first objective. $D$ can be, for the moment, either a finite or an infinite set. An illustration is given in the right-hand plot of Fig. 16.1 where the function $f$ describing the front has a domain of $D = [x_{\min}, x_{\max}]$.

**16.2.1 Example**
*Consider the bi-objective problem DTLZ2 [119] which is defined as*

$$\begin{aligned} \textit{minimize} \quad & \mathcal{F}_1(d) = \Big(1 + g(d_M)\Big)\cos(d_1\pi/2) \\ \textit{minimize} \quad & \mathcal{F}_2(d) = \Big(1 + g(d_M)\Big)\sin(d_1\pi/2) \\ & g(d_M) = \sum_{d_i \in d_M} (d_i - 0.5)^2 \\ \textit{subject to} \quad & 0 \le d_i \le 1 \textit{ for } i = 1, \dots n \end{aligned} \tag{16.2}$$

*where $d_M$ denotes a subset of the decision variables $d = (d_1, \dots, d_n) \in [0,1]^n$ with $g(d_M) \ge 0$. The Pareto front is reached for $g(d_M) = 0$. Hence, the Pareto-optimal points have objective vectors* $(\cos(d_1\pi/2), \sin(d_1\pi/2))$ *with $0 \le d_1 \le 1$ which can be rewritten as points $(x, f(x))$ with $f(x) = \sqrt{1-x^2}$ and $x \in D = [0,1]$, see Fig. 16.9(f).*

Since $f$ represents the shape of the trade-off surface, we can conclude that, for minimization problems, $f$ is strictly monotonically decreasing in $D$[6]. The coordinates of a point belonging to the Pareto front are given as a pair $(x, f(x))$ with $x \in D$ and therefore, a point is entirely determined by the function $f$ and the first coordinate $x \in D$. For $\mu$ points on the Pareto front, we denote their first coordinates as $(x_1, \dots, x_\mu)$. Without loss of generality, it is assumed that $x_i \le x_{i+1}$, for $i = 1, \dots, \mu - 1$ and for notation convenience, we set $x_{\mu+1} := r_1$ and $f(x_0) := r_2$ where $r_1$ and $r_2$ are the first and second coordinate of the reference point (see Figure 16.2). The weighted hypervolume enclosed by these points can be decomposed into $\mu$ components, each corresponding to the integral of the weight function $w$ over a rectangular area (see Figure 16.2). The resulting weighted hypervolume writes:

$$I_{H,w}((x_1, \dots, x_\mu)) := \sum_{i=1}^{\mu} \int_{x_i}^{x_{i+1}} \left( \int_{f(x_i)}^{f(x_0)} w(x,y)dy \right) dx \ . \tag{16.3}$$

When the weight function equals one everywhere, one retrieves the expression for the (non-weighted) hypervolume [25]

$$I_H((x_1, \dots, x_\mu)) := \sum_{i=1}^{\mu} (x_{i+1} - x_i)(f(x_0) - f(x_i)) \ . \tag{16.4}$$

Indicator-based evolutionary algorithms that aim at optimizing a unary indicator $I : 2^X \to \mathbb{R}$ such as the hypervolume transform a multiobjective problem into the single-objective one consisting of finding a set of points maximizing the respective indicator $I$. In practice, the cardinality of these sets of points is usually upper bounded by a constant $\mu$, typically the population size. Generalizing the definition by [25], we define an *optimal $\mu$-distribution* as a set of $\mu$ points maximizing $I$.

**16.2.2 Definition (Optimal $\mu$-distribution)**
*For $\mu \in \mathbb{N}$ and a unary indicator $I$, a set of $\mu$ points maximizing $I$ is called an optimal $\mu$-distribution for $I$.*

[6] If $f$ is not strictly monotonically decreasing, we can find Pareto-optimal points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ with $x_1, x_2 \in D$ such that, without loss of generality, $x_1 < x_2$ and $f(x_1) \le f(x_2)$, i.e., $(x_1, f(x_1))$ dominates $(x_2, f(x_2))$.

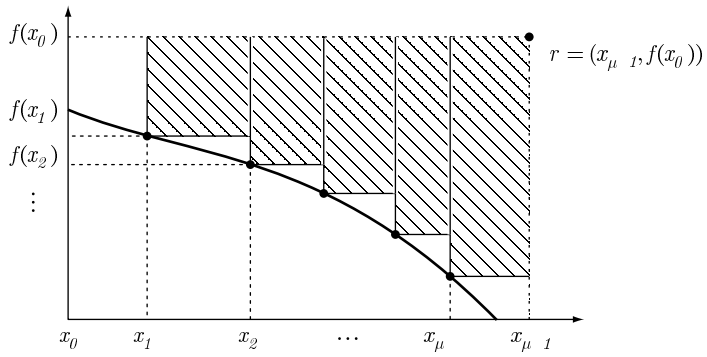



Figure 16.2: Computation of the hypervolume indicator for $\mu$ solutions $(x_1, f(x_1)), \ldots, (x_\mu, f(x_\mu))$ and the reference point $r = (r_1, r_2)$ in the bi-objective case as defined in Eq. 16.3 and Eq. 16.4 respectively.

The rest of the paper is devoted to understand optimal $\mu$-distributions for the hypervolume indicator in the bi-objective case. The $x$-coordinates of an optimal $\mu$-distribution for the hypervolume $I_{H,w}$ will be denoted $(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu)$ and will thus satisfy

$$I_{H,w}((\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu)) \geq I_{H,w}((x_1, \ldots, x_\mu)) \text{ for all } (x_1, \ldots, x_\mu) \in D \times \ldots \times D \ .$$

Note, that the optimal $\mu$-distribution might not be unique, and $(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu)$ therefore refers to *one* optimal $\mu$-distribution. The corresponding value of the hypervolume will be denoted $\overline{I_{H,w}^\mu}$, i.e., $\overline{I_{H,w}^\mu} = I_{H,w}((\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu))$.

**16.2.3 Remark**
*Looking at Eq. 16.3 and Eq. 16.4, we see that for a fixed $f$, a fixed weight $w$, and a fixed reference point, the problem of finding a set of $\mu$ points maximizing the weighted hypervolume amounts to finding the solution of a $\mu$-dimensional single-objective maximization problem, i.e., optimal $\mu$-distributions are the solution of a single objective problem of $\mu$ variables.*

## 16.3 Characterization of Optimal $\mu$-Distributions for Hypervolume Indicators

Several contradicting beliefs about the bias introduced by the hypervolume indicator have been reported in the literature. For example, [381] stated that, when optimizing the hypervolume in maximization problems, "convex regions may be preferred to concave regions", which has been also stated by [262] later on, whereas [116] argued that "[...] the hyper-volume measure is biased towards the boundary solutions". [221] observed that a local optimum of the hypervolume indicator "seems to be 'well-distributed'" which was also confirmed empirically [222, 136]. [61], in addition, state several properties of the hypervolume's bias: (i) optimizing the hypervolume indicator focuses on knee points; (ii) the distribution of points on the extremes is less dense than on knee points; (iii) only linear front shapes allow for equally spread solutions; and (iv) extremal solutions are maintained. In the light of these contradicting statements, a thorough characterization of optimal $\mu$-distributions for the hypervolume indicator is necessary. Especially for the weighted hypervolume indicator, the bias of the indicator and the influence of the weight function $w$ on optimal $\mu$-distributions in particular has not been fully understood.

In this section, we first prove the existence of optimal $\mu$-distributions for lower semi-continuous fronts, we show the monotonicity in $\mu$ of the hypervolume associated with optimal $\mu$-distributions, and derive necessary conditions satisfied by optimal $\mu$-distributions. In a second part, we derive the density associated with optimal $\mu$-distributions when $\mu$ grows to infinity.

### 16.3.1 Finite Number of Points

#### Existence of Optimal $\mu$-Distributions

Before to further investigate optimal $\mu$-distributions for $I_{H,w}$, we establish a setting ensuring their existence. We will from now on assume that $D$ is a closed interval that we denote $[x_{\min}, x_{\max}]$ such that $f$ writes:

$$x \in [x_{\min}, x_{\max}] \mapsto f(x).$$

A function is lower semi-continuous if for all $x_0$, $\liminf_{x \to x_0} f(x) \geq f(x_0)$. If $f$ is decreasing (which is the case when $f$ describes a Pareto front), lower semi-continuous is equivalent to continuity to the right. As shown in the following theorem, a sufficient setting for the existence of optimal distributions is the lower semi-continuity of $f$.

**16.3.1 Theorem (Existence of optimal $\mu$-distributions)**
*Let $\mu \in \mathbb{N}$, if the function $f$ describing the Pareto front is lower semi-continuous, there exists (at least) one set of $\mu$ points maximizing the hypervolume.*

*Proof.* We are going to prove that $I_{H,w}$ is upper semi-continuous if $f$ is lower semi-continuous, and then apply the Extreme Value Theorem. Since $I_{H,w}$ is the sum of $\mu$ functions $g(x_i, x_{i+1})$ where $g(\alpha, \beta) = \int_\alpha^\beta \left( \int_{f(\alpha)}^{f(x_0)} w(x,y)dy \right) dx$, we will prove the upper semi-continuity of $g(x_i, x_{i+1})$ for $(x_i, x_{i+1}) \in [x_{\min}, x_{\max}]$. This will imply the upper semi-continuity of $I_{H,w}$. Let $(x_i, x_{i+1}) \in [x_{\min}, x_{\max}]$ and let $(x_i^n, x_{i+1}^n)_{n \in \mathbb{N}}$ converging to $(x_i, x_{i+1})$. We will now prove that $\limsup g(x_i^n, x_{i+1}^n) \leq g(x_i, x_{i+1})$. Since

$$\limsup_{n \to \infty} g(x_i^n, x_{i+1}^n) = \limsup_{n \to \infty} \int \int \mathbf{1}_{[x_i^n, x_{i+1}^n]}(x) \mathbf{1}_{[f(x_i^n), f(x_0)]}(y) w(x,y) dy dx \ ,$$

and $\mathbf{1}_{[x_i^n, x_{i+1}^n]}(x) \mathbf{1}_{[f(x_i), f(x_0)]}(x) w(x,y) \leq \mathbf{1}_{[x_{\min}, x_{\max}]}(x) \mathbf{1}_{[f(x_{\max}), f(x_0)]}(x) w(x,y)$ we can use the (Reverse) Fatou Lemma that implies $\limsup g(x_i^n, x_{i+1}^n) \leq \int \int \limsup \mathbf{1}_{[x_i^n, x_{i+1}^n]}(x) \mathbf{1}_{[f(x_i^n), f(x_0)]}(y) w(x,y) dy dx$. Since $f$ is lower semi-continuous, $\liminf f(x_i^n) \geq f(x_i)$ holds which is equivalent to $\limsup(f(x_0) - f(x_i^n)) = f(x_0) - \liminf f(x_i^n) \leq f(x_0) - f(x_i)$. Hence, $\limsup \mathbf{1}_{[f(x_i^n), f(x_0)]}(y) \leq \mathbf{1}_{[f(x_i), f(x_0)]}(y)$ and thus

$$\limsup_{n \to \infty} g(x_i^n, x_{i+1}^n) \leq \int \int \mathbf{1}_{[x_i, x_{i+1}]}(x) \mathbf{1}_{[f(x_i), f(x_0)]}(y) w(x,y) dy dx = g(x_i, x_{i+1}) \ .$$

We have proven the upper semi-continuity of $g$ which implies the upper semi-continuity of $I_{H,w} : [x_{\min}, x_{\max}]^\mu \to \mathbb{R}$. Given that $[x_{\min}, x_{\max}]^\mu$ is compact, we can imply from the Extreme Value Theorem that there exists a set of $\mu$ points maximizing the hypervolume indicator. □

Note that, in case of bi-objective *maximization* problems, the lower semi-continuity of $f$ has to be changed into upper semi-continuity which has been proven recently for the unweighted hypervolume [84]. Note also that the previous theorem states the existence but not the uniqueness, which cannot be guaranteed in general. With this respect, we would like to mention that the question of uniqueness is related loosely to another property of the hypervolume which is not discussed here but has high importance in practice: For indicator-based algorithms and the analysis of their convergence speed, it is highly important whether *local* optima are observed during the search. This property is, however, defined within the decision space $X$ and especially depends on the mapping between the decision space and the objective space which is not taken into account in this study.

Furthermore, if the front is not semi-continuous, optimal $\mu$-distributions might not exist. In the following proposition, we construct an example of a front where this is the case, i.e., where there is *no* optimal $\mu$-distribution for $\mu = 1$.

**16.3.2 Proposition**
*Let $r = (r_1, r_1)$ be a reference point with $r_1 > 1.2$. Consider the front $f_{ce} : [0,1] \to [0,1.2]$ with*

$$f_{ce}(x) = \begin{cases} 1 - x + 0.2 & \text{if } x \leq \frac{1}{2}, \\ 1 - x & \text{if } x \in ]\frac{1}{2}, 1] \end{cases} .$$

*Then $f$ does not admit an optimal 1-distribution for the unweighted hypervolume.*

*Proof.* Consider first the linear front $f : x \in [0,1] \to [0,1], x \mapsto 1 - x$. Here, the optimal 1-distribution is the point $(0.5, 0.5)$ with a corresponding hypervolume value of $\gamma = (r_1 - \frac{1}{2})(r_1 - \frac{1}{2})$ [7]. Consider now $h(x) = f_{ce}(x)$ for all $x \in [0,1]$ except for $x = 0.5$ where $h(x) = 0.5$. Then, $h$ is continuous to the right and thus lower semi-continuous. Hence, according to Theorem 16.3.1 it admits an optimal 1-distribution. In addition, remark that the hypervolume contribution for any $x \in [0, 0.5[$ is strictly smaller for $h$ than for $f$ and equal for $x \in [0.5, 1]$. Thus $(0.5, 0.5)$ is also the optimal 1-distribution of $h$ with hypervolume $\gamma$. However, for $f_{ce}$, the hypervolume contribution is strictly smaller than for $f$ for $x \in [0, 0.5]$ and equal for $x \in ]0.5, 1]$ with a gap at 0.5 such that $\gamma$ cannot be reached for any point in $[0,1]$ though one has values arbitrary close from it for $x$ arbitrary close from 0.5 to the right. □

We have chosen $\mu = 1$ in the previous proposition for the sake of simplicity, however, such a counter-example can be generalized for arbitrary $\mu$ by following the same idea. Let us also note that, lower semi-continuity is not a necessary condition for the existence of optimal $\mu$-distributions: if we simply introduce the discontinuity of the function $f_{ce}$ in the previous proposition somewhere in $]0, 0.5[$ instead of at $x = 0.5$, the optimal 1-distribution would exist (and be located at $x = 0.5$) though the function describing the front is not lower semi-continuous.

### Strict Monotonicity of Hypervolume in $\mu$ for Optimal $\mu$-Distributions

The following proposition establishes that the hypervolume of optimal $(\mu + 1)$-distributions is strictly larger than the hypervolume of optimal $\mu$-distributions. This result is a generalization of [25, Lemma 1].

**16.3.3 Proposition**
*Let $D \subseteq \mathbb{R}$, possibly finite and $f : x \in D \mapsto f(x)$ describe a Pareto front. Let $\mu_1$ and $\mu_2 \in \underline{\mathcal{N}}$ with $\mu_1 < \mu_2$, then*

$$\overline{I_{H,w}^{\mu_1}} < \overline{I_{H,w}^{\mu_2}}$$

*holds if $D$ contains at least $\mu_1 + 1$ elements $x_i$ for which $x_i < r_1$ and $f(x_i) < r_2$ holds.*

*Proof.* To prove the proposition, it suffices to show the inequality for $\mu_2 = \mu_1 + 1$. Assume $D_{\mu_1} = \{\overline{x}_1^{\mu_1}, \ldots, \overline{x}_{\mu_1}^{\mu_1}\}$ with $\overline{x}_i^{\mu} \in \mathbb{R}$ is the set of $x$-values of the objective vectors of the optimal $\mu_1$-distribution for $I_{H,w}$ with a hypervolume value of $\overline{I_{H,w}^{\mu_1}}$ if the Pareto front is described by $f$. Since $D$ contains at least $\mu_1 + 1$ elements, the set $D \backslash D_{\mu_1}$ is not empty and we can pick any $x_{\text{new}} \in D \backslash D_{\mu_1}$ that is not contained in the optimal $\mu_1$-distribution for $I_{H,w}$ and for which $f(x_{\text{new}})$ is defined. Let $x_r := \min\{x | x \in D_{\mu_1} \cup \{r_1\}, x > x_{\text{new}}\}$ be the closest element of $D_{\mu_1}$ to the right of $x_{\text{new}}$ (or $r_1$ if $x_{\text{new}}$ is larger than all elements of $D_{\mu_1}$). Similarly, let $f_l := \min\{r_2, \{f(x) | x \in D_{\mu_1}, x < x_{\text{new}}\}\}$ be the function value of the closest element of $D_{\mu_1}$ to the left of $x_{\text{new}}$ (or $r_2$ if $x_{\text{new}}$ is smaller than all elements of $D_{\mu_1}$). Then, all objective vectors within $H_{\text{new}} := [x_{\text{new}}, x_r[ \times [f(x_{\text{new}}), f_l[$ are weakly dominated by the new point $(x_{\text{new}}, f(x_{\text{new}}))$ but are not dominated by any objective vector given by $D_{\mu_1}$. Furthermore, $H_{\text{new}}$ is not a null set (i.e., has a strictly positive measure) since $x_{\text{new}} > x_r$ and $f_l > f(x_{\text{new}})$ and the weight $w$ is strictly positive which gives $\overline{I_{H,w}^{\mu_1}} < \overline{I_{H,w}^{\mu_2}}$. □

[7] In case $\mu = 1$ and $f(x) = 1 - x$, we can easily compute the maximum of the hypervolume $I_{H,w}(x) = (r_1 - x)(r_1 - (1 - x)) = r_1^2 - r_1 + x - x^2$ of the single point at $x$ by computing the derivative of $I_{H,w}(x)$ and setting it to zero: $I'_{H,w}(x) = 1 - 2x = 0$.

**Characterization of Optimal $\mu$-Distributions for Finite $\mu$**

In this section, we derive a general result to characterize optimal $\mu$-distributions for the hypervolume indicator if $\mu$ is finite. The result holds under the assumption that the front $f$ is differentiable and is a direct application of the fact that solutions of a maximization problem that do not lie on the boundary of the search domain are stationary points, i.e., points where the gradient is zero.

**16.3.4 Theorem (Necessary conditions for optimal $\mu$-distributions for $I_{H,w}$)**
*If $f$ is continuous and differentiable and $(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu)$ are the x-coordinates of an optimal $\mu$-distribution for $I_{H,w}$, then for all $\overline{x}_i^\mu$ with $\overline{x}_i^\mu > x_{\min}$ and $\overline{x}_i^\mu < x_{\max}$*

$$f'(\overline{x}_i^\mu) \int_{\overline{x}_i^\mu}^{\overline{x}_{i+1}^\mu} w(x, f(\overline{x}_i^\mu))dx = \int_{f(\overline{x}_{i-1}^\mu)}^{f(\overline{x}_i^\mu)} w(\overline{x}_i^\mu, y)dy \tag{16.5}$$

*holds where $f'$ denotes the derivative of $f$, $f(\overline{x}_0^\mu) = r_2$ and $\overline{x}_{\mu+1}^\mu = r_1$.*

*Proof.* The proof idea is simple: optimal $\mu$-distributions maximize the $\mu$-dimensional function $I_{H,w}$ defined in Eq. 16.3 and should therefore satisfy necessary conditions for local extrema of a $\mu$-dimensional function stating that the coordinates of local extrema either lie on the boundary of the domain (here $x_{\min}$ or $x_{\max}$) or satisfy that the partial derivative with respect to this coordinate is zero. Hence, we see that the partial derivatives of $I_{H,w}$ have to be computed. This step is quite technical and is presented in Appendix 16.7.1 on page 220 together with the full proof of the theorem. □

The previous theorem proves an implicit relation between the points of an optimal $\mu$-distribution. However, in certain cases of weights, this implicit relation can be made explicit as illustrated first on the example of the weight function $w(x, y) = \exp(-x)$, aiming at favoring points with small values along the first objective.

**16.3.5 Example**
*If $w(x, y) = \exp(-x)$, Eq. 16.5 simplifies into the explicit relation*

$$f'(\overline{x}_i^\mu)(e^{-\overline{x}_i^\mu} - e^{-\overline{x}_{i+1}^\mu}) = e^{-\overline{x}_i^\mu}(f(\overline{x}_i^\mu) - f(\overline{x}_{i-1}^\mu)) \ . \tag{16.6}$$

Another example where the relation is explicit is given for the unweighted hypervolume $I_H$ that we can obtain as a corollary of the previous theorem and which coincides with a previous result [25, Proposition 1].

**16.3.6 Corollary**
*(Necessary condition for optimal $\mu$-distributions for $I_H$) If $f$ is continuous, differentiable and $(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu)$ are the x-coordinates of an optimal $\mu$-distribution for $I_H$, then for all $\overline{x}_i^\mu$ with $\overline{x}_i^\mu > x_{\min}$ and $\overline{x}_i^\mu < x_{\max}$*

$$f'(\overline{x}_i^\mu)(\overline{x}_{i+1}^\mu - \overline{x}_i^\mu) = f(\overline{x}_i^\mu) - f(\overline{x}_{i-1}^\mu) \tag{16.7}$$

*holds where $f'$ denotes the derivative of $f$, $f(\overline{x}_0^\mu) = r_2$ and $\overline{x}_{\mu+1}^\mu = r_1$.*

*Proof.* The proof follows immediately from setting $w = 1$ in Eq. 16.5. □

**16.3.7 Remark**
*Corollary 16.3.6 implies that the points of an optimal $\mu$-distribution for $I_H$ are linked by a second order recurrence relation. Thus, in this case, finding optimal $\mu$-distributions for $I_H$ does not correspond to solving a $\mu$-dimensional optimization problem as stated in Remark 16.2.3 but to a 2-dimensional one. The same remark holds for $I_{H,w}$ and $w(x, y) = \exp(-x)$ as can be seen in Eq. 16.6.*

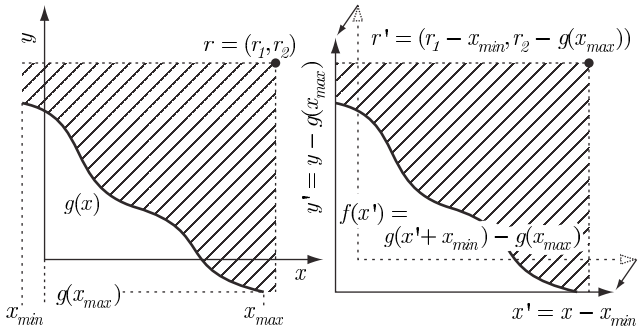



Figure 16.3: Every continuous front $g(x)$ (left) can be described by a function $f : x' \in [0, x'_{max}] \mapsto f(x')$ with $f(x'_{max}) = 0$ (right) by a simple translation.

The previous corollary can also be used to characterize optimal $\mu$-distributions for certain Pareto fronts more generally as the following example shows.

**16.3.8 Example**
*Consider a linear Pareto front, i.e., a front that can be formally defined as* $f : x \in [x_{\min}, x_{\max}] \mapsto \alpha x + \beta$ *where* $\alpha < 0$ *and* $\beta \in \mathbb{R}$. *Then, it follows immediately from Corollary 16.3.6 and Eq. 16.7 that the optimal* $\mu$*-distribution for* $I_H$ *maps to objective vectors with equal distances between two neighbored solutions (see also Theorem 16.5.1 in Sec. 16.5.1):*

$$\alpha\left(\overline{x}^{\mu}_{i+1} - \overline{x}^{\mu}_{i}\right) = f(\overline{x}^{\mu}_{i}) - f(\overline{x}^{\mu}_{i-1}) = \alpha(\overline{x}^{\mu}_{i} - \overline{x}^{\mu}_{i-1})$$

*for* $i = 2, \ldots, \mu - 1$. *Note that this result coincides with earlier results for linear fronts with slope* $\alpha = -1$ *[60] or the even more specific case of a front of shape* $f(x) = 1 - x$ *[137].*

### 16.3.2 Number of Points Going to Infinity

Besides for simple fronts, like the linear one, Eq. 16.5 and Eq. 16.7 cannot be easily exploited to derive optimal $\mu$-distributions explicitly. However, one is interested in knowing how the hypervolume indicator influences the spread of points on the front and in characterizing the bias introduced by the hypervolume. To reply to these questions, we will assume that the number of points $\mu$ grows to infinity and derive the density of points associated with optimal $\mu$-distributions for the hypervolume indicator.

We assume without loss of generality that $x_{\min} = 0$ and that $f : x \in [0, x_{\max}] \mapsto f(x)$ with $f(x_{\max}) = 0$ (Fig. 16.3). We also assume that $f$ is continuous within $[0, x_{\max}]$, differentiable, and that its derivative is a continuous function $f'$ defined in the interval $]0, x_{\max}[$. Instead of maximizing the weighted hypervolume indicator $I_{H,w}$, it is easy to see that, since $r_1 r_2$ is constant, one can equivalently minimize

$$r_1 r_2 - I_{H,w}((x_1, \ldots, x_\mu)) = \sum_{i=0}^{\mu} \int_{x_i}^{x_{i+1}} \int_{0}^{f(x_i)} w(x, y)\, dy\, dx$$

with $x_0 = 0$, $f(x_0) = r_2$, and $x_{\mu+1} = r_1$ (see Fig. 16.4(b)). If we subtract the area below the front curve, i.e., the integral $\int_0^{x_{\max}} \left( \int_0^{f(x)} w(x, y) dy \right) dx$ of constant value (Fig. 16.4(c)), we see that minimizing

$$\sum_{i=0}^{\mu} \int_{x_i}^{x_{i+1}} \int_{0}^{f(x_i)} w(x, y)\, dy\, dx - \int_{0}^{x_{\max}} \int_{0}^{f(x)} w(x, y)\, dy\, dx \tag{16.8}$$

is equivalent to maximizing the weighted hypervolume indicator (Fig. 16.4(d)).

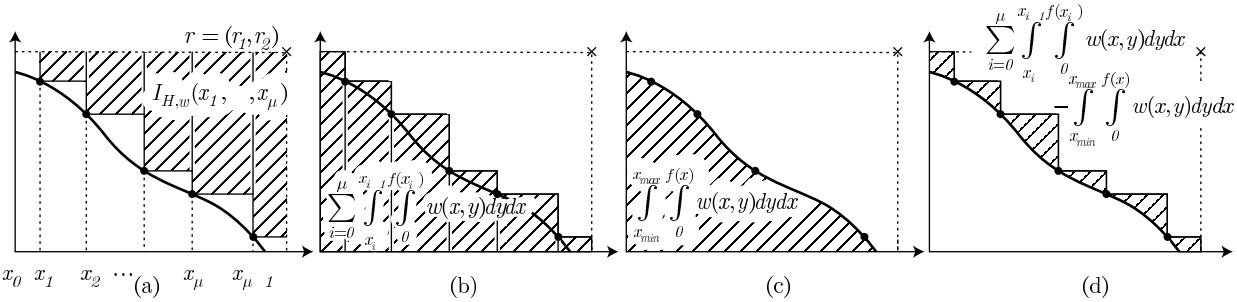



Figure 16.4: Illustration of the idea behind deriving the optimal density: Instead of maximizing the weighted hypervolume indicator $I_{H,w}((x_1, \ldots, x_\mu))$ (a), one can minimize the shaded area in (b) which is equivalent to minimizing the integral between the attainment surface of the solution set and the front itself which can be expressed with the help of the integral of $f$ (d).

For a fixed integer $\mu$, we now consider a sequence of $\mu$ ordered points $\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu$ in $[0, x_{\max}]$ that lie on the Pareto front. We assume that the sequence converges—when $\mu$ goes to $\infty$—to a density $\delta(x)$ that is regular enough. Formally, the density in $x \in [0, x_{\max}]$ is defined as the limit of the number of points contained in a small interval $[x, x+h[$ normalized by the total number of points $\mu$ when both $\mu$ goes to $\infty$ and $h$ to 0, i.e., $\delta(x) = \lim_{\substack{\mu\to\infty \\ h\to 0}} \left(\frac{1}{\mu h} \sum_{i=1}^{\mu} \mathbf{1}_{[x,x+h[}(\overline{x}_i^\mu)\right)$. As explained above, maximizing the weighted hypervolume is equivalent to minimizing Eq. 16.8, which is also equivalent to minimizing

$$E_\mu = \mu \left[ \sum_{i=0}^{\mu} \int_{\overline{x}_i^\mu}^{\overline{x}_{i+1}^\mu} \left( \int_0^{f(\overline{x}_i^\mu)} w(x,y)dy \right) dx - \int_0^{x_{\max}} \left( \int_0^{f(x)} w(x,y)dy \right) dx \right] , \tag{16.9}$$

where we have multiplied Eq. 16.8 by $\mu$ to obtain a quantity that will converge to a limit when $\mu$ goes to $\infty$. Indeed Eq. 16.8 converges to 0 when $\mu$ increases. We now conjecture that the equivalence between minimizing $E_\mu$ and maximizing the hypervolume also holds for $\mu$ going to infinity. Therefore, our proof consists of two steps: (1) compute the limit of $E_\mu$ when $\mu$ goes to $\infty$. This limit is going to be a function of a density $\delta$. (2) Find the density $\delta$ that minimizes $E(\delta) := \lim_{\mu\to\infty} E_\mu$. The first step therefore consists in computing the limit of $E_\mu$.

**16.3.9 Lemma**
*If $f$ is continuous, differentiable with the derivative $f'$ continuous, if $x \mapsto w(x, f(x))$ is continuous, if $\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu$ converge to a continuous density $\delta$, with $\frac{1}{\delta} \in L^2(0, x_{\max})$[8], and $\exists c \in \mathbb{R}^+$ such that*

$$\mu \sup \left( \left( \sup_{0 \le i \le \mu-1} |\overline{x}_{i+1}^\mu - \overline{x}_i^\mu| \right), |x_{\max} - \overline{x}_\mu^\mu| \right) \to c$$

*then $E_\mu$ converges for $\mu \to \infty$ to*

$$E(\delta) := -\frac{1}{2} \int_0^{x_{\max}} \frac{f'(x) w(x, f(x))}{\delta(x)} dx \ . \tag{16.10}$$

*Proof.* For the technical proof, we refer to Appendix 16.7.2 on page 222. □

The limit density of a $\mu$-distribution for $I_{H,w}$, as explained before, minimizes $E(\delta)$. It remains therefore to find the density which minimizes $E(\delta)$. This optimization problem is posed in a functional space and is also a constrained problem since the density $\delta$ has to satisfy the constraint $J(\delta) := \int_0^{x_{\max}} \delta(x)dx = 1$. The

[8] $L^2(0, x_{\max})$ is a functional space (Banach space) defined as the set of all functions whose square is integrable in the sense of the Lebesgue measure.

constraint optimization problem (P) that needs to be solved is summarized in:

$$\begin{aligned} &\text{minimize } E(\delta) \\ &\text{subject to } J(\delta) = 1 \ . \end{aligned} \tag{P}$$

In a similar way than Theorem 7 in [25] where $-f'$ needs to be replaced everywhere by $-f'w$[9], we find that the density solution of the constraint optimization problem (P) equals

$$\delta(x) = \frac{\sqrt{-f'(x)w(x,f(x))}}{\int_0^{x_{\max}} \sqrt{-f'(x)w(x,f(x))}dx} \ .$$

For $x_{\min} \neq 0$, the density reads

$$\delta(x) = \frac{\sqrt{-f'(x)w(x,f(x))}}{\int_{x_{\min}}^{x_{\max}} \sqrt{-f'(x)w(x,f(x))}dx} \ . \tag{16.11}$$

**16.3.10 Remark**
*The previous density corresponds to the density of points of the front projected onto the x-axis, however, if one is interested into the density on the front $\delta_F$*[10] *one has to normalize the result from Eq. 16.11 by the norm of the tangent for points of the front, i.e., $\sqrt{1+f'(x)^2}$. Therefore, the density on the front is*

$$\delta_F(x) = \frac{\sqrt{-f'(x)w(x,f(x))}}{\int_{x_{\min}}^{x_{\max}} \sqrt{-f'(x)w(x,f(x))}dx} \frac{1}{\sqrt{1+f'(x)^2}}. \tag{16.12}$$

**16.3.11 Example**
*Let us consider the test problem ZDT2 [378, see also Fig. 16.9] the Pareto front of which can be described by $f(x) = 1 - x^2$ with $x_{\min} = 0$ and $x_{\max} = 1$ and $f'(x) = -2x$ [25]. Considering the unweighted case, the density on the x-axis according to Eq. 16.11 is $\delta(x) = \frac{3}{2}\sqrt{x}$ and the density on the front according to Eq. 16.12 is $\delta_F(x) = \frac{3}{2}\frac{\sqrt{x}}{\sqrt{1+4x^2}}$, see Fig. 16.9 for an illustration.*

To summarize, we have seen that the density follows as a limit result from the fact that the integral between the attainment function of the solution set with $\mu$ points and the front itself (Fig. 16.4(d)) has to be minimized and the optimal $\mu$-distribution for $I_{H,w}$ and a finite number of points converges to the density when $\mu$ increases. Furthermore, we can conclude that the proportion of points of an optimal $\mu$-distribution with $x$-values within a certain interval $[a,b]$ converges to $\int_a^b \delta(x)dx$ if the number of points $\mu$ goes to infinity. How this relates to practice will be presented in Sec. 16.5 where analytical and experimental results on the density for specific well-known test problems are shown.

Instead of applying the results to specific test functions, the above results on the hypervolume indicator can also be interpreted in a broader sense: From (16.11), we know that it is only the weight function and the slope of the front that influences the density of the points of an optimal $\mu$-distribution—contrary to several prevalent beliefs as stated in the beginning of this section. Since the density of points does not depend on the position on the front but only on the gradient and the weight at the respective point, the density close to the extreme points of the front can be very high or very low—it only depends on the front shape. Section 16.4.1 will even present conditions under which the extreme points will never be included in an optimal $\mu$-distribution for $I_{H,\tilde{w}}$—in contrast to the statement by [61]. In the unweighted case, we observe that the density has its maximum for front parts where the tangent has a

[9] Note that in [25, Theorem 7] and its proof, the density should belong to $L^2(0, x_{\max})$ but also, $1/\delta \in L^2(0, x_{\max})$.

[10] The density on the front gives for any curve on the front (a piece of the front) C, the proportion of points of the optimal $\mu$-distribution (for $\mu$ to infinity) contained in this curve by integration on the curve: $\int_C \delta_F ds$. Since we know that for any parametrization of C, say $t \in [a,b] \to \gamma(t) \in \mathbb{R}^2$, we have $\int_C \delta_F ds = \int_a^b \delta_F(\gamma(t))\|\gamma'(t)\|_2 dt$, we can for instance use the natural parametrization of the front given by $\gamma(t) = (t, f(t))$ giving $\|\gamma'(t)\|_2 = \sqrt{1+f'(t)^2}$ that therefore implies that $\delta(x) = \delta_F(x)\sqrt{1+f'(x)^2}$. Note that we do a small abuse of notation writing $\delta_F(x)$ instead of $\delta_F(\gamma(x)) = \delta_F((x, f(x)))$.

gradient of -45° [25]. Therefore, and compliant with the statement by [61], optimizing the unweighted hypervolume indicator stresses so-called knee-points—parts of the Pareto front decision makers believe to be interesting regions [111, 82]. However, choosing a non-constant weight can highly change the distribution of points and makes it possible to include several user preferences into the search. The new result in (16.11) now explains *how* the distribution of points changes: for a fixed front, it is the square root of the weight that is directly reflected in the optimal density.

## 16.4 Influence of the Reference Point on the Extremes

Clearly, optimal $\mu$-distributions for $I_{H,w}$ are in some way influenced by the choice of the reference point $r$ as the definition of $I_{H,w}$ in Eq. 16.3 depends on $r$ and it is well-known from experiments that the reference point can influence the outcomes of multiobjective evolutionary algorithms drastically [222]. How in general, the outcomes of hypervolume-based algorithms are influenced by the choice of the reference point, however, has not been investigated from a theoretical perspective. In particular, it could not be observed from practical investigations how the reference point has to be set to ensure to find the extremes of the Pareto front.

In practice, mainly rules-of-thumb exist on how to choose the reference point. Many authors recommend to use the corner of a space that is a little bit larger than the actual objective space as the reference point. Examples include the corner of a box 1% larger than the objective space [219] or a box that is larger by an additive term of 1 than the extremal objective values obtained [61]. In various publications where the hypervolume indicator is used for performance assessment, the reference point is chosen as the nadir point[11] of the investigated solution set [310, 309, 195], while others recommend a rescaling of the objective values everytime the hypervolume indicator is computed [379].

In this section, we ask the question of how the choice of the reference point influences optimal $\mu$-distributions and theoretically investigate in particular whether there exists a choice for the reference point that implies that the extremes of the Pareto front are included in optimal $\mu$-distributions. The presented results generalize the statements by [25] to the weighted hypervolume indicator and give insights into how the reference point should be chosen if the weight function does not equal 1 everywhere. Our main result, stated in Theorem 16.4.3 and Theorem 16.4.7, shows that for continuous and differentiable Pareto fronts we can give implicit lower bounds on the $\mathcal{F}_1$ and $\mathcal{F}_2$ value for the reference point (possibly infinite depending on $f$ and $w$) such that all choices above this lower bound ensure the existence of the extremes in an optimal $\mu$-distribution for $I_{H,w}$. For the special case of the unweighted hypervolume indicator, these lower bounds turn into explicit lower bounds (Corollaries 16.4.5 and 16.4.8). Moreover, Sec. 16.4.1 shows that it is necessary to have a finite derivative on the left extreme and a non-zero one on the right extreme to ensure that the extremes are contained in an optimal $\mu$-distribution. This result contradicts the common belief that it is sufficient to choose the reference point slightly above and to the right to the nadir point or the border of the objective space to obtain the extremes as indicated above. A new result (Theorem 16.4.10), not covered by [25], shows that a point slightly worse than the nadir point in all objectives starts to become a good choice for the reference point as soon as $\mu$ is large enough.

Before we present the results, recall that $r = (r_1, r_2)$ denotes the reference point and $y = f(x)$ with $x \in [x_{\min}, x_{\max}]$ represents the Pareto front where therefore $(x_{\min}, f(x_{\min}))$ and $(x_{\max}, f(x_{\max}))$ are the left and right extremal points. Since we want that all Pareto-optimal solutions have a contribution to the hypervolume of the front in order to be possibly part of the optimal $\mu$-distribution, we assume that the reference point is dominated by all Pareto-optimal solutions, i.e., $r_1 > x_{\max}$ and $r_2 > f(x_{\min})$.

[11] In our notation, the nadir point equals $(x_{\max}, f(x_{\min}))$, i.e., is the smallest objective vector that is weakly dominated by all Pareto-optimal points.

### 16.4.1 Finite Number of Points

For the moment, we assume that the number of points $\mu$ is finite and provide necessary and sufficient conditions for finding a finite reference point such that the extremes are included in any optimal $\mu$-distribution for $I_{H,w}$. In Sec. 16.4.2, we later on derive further results in case $\mu$ goes to infinity.

#### Fronts for Which It Is Impossible to Have the Extremes

A previous belief was that choosing the reference point of the hypervolume indicator in a way, such that it is dominated by all Pareto-optimal points, is enough to ensure that the extremes can be reached by an indicator-based algorithm aiming at maximizing the hypervolume indicator. The main reason for this belief is that with such a choice of reference point, the extremes of the Pareto front always have a positive contribution to the overall hypervolume indicator and should be therefore chosen by the algorithm's environmental selection. However, theoretical investigations revealed that we cannot always ensure that the extreme points of the Pareto front are contained in an optimal $\mu$-distribution for the unweighted hypervolume indicator [25]. In particular, a necessary condition to have the left (resp. right) extreme included in optimal $\mu$-distributions is to have a finite (resp. non-zero) derivative on the left extreme (resp. right extreme). The following theorem generalizes this result and shows that also for the weighted hypervolume indicator, the same necessary condition holds.

**16.4.1 Theorem**
*Let $\mu$ be a positive integer. Assume that $f$ is continuous on $[x_{\min}, x_{\max}]$, non-increasing, differentiable on $]x_{\min}, x_{\max}[$ and that $f'$ is continuous on $]x_{\min}, x_{\max}[$ and that the weight function $w$ is continuous and positive. If $\lim_{x \to x_{\min}} f'(x) = -\infty$, the left extremal point of the front is never included in an optimal $\mu$-distribution for $I_{H,w}$. Likewise, if $f'(x_{\max}) = 0$, the right extremal point of the front is never included in an optimal $\mu$-distribution for $I_{H,w}$.*

*Proof.* The idea behind the proof is to assume the extreme point to be contained in an optimal $\mu$-distribution and to show a contradiction. In particular, the gain and loss in hypervolume if the extreme point is shifted can be computed analytically. A limit result for the case that $\lim_{x \to x_{\min}} f'(x) = -\infty$ (and $f'(x_{\max}) = 0$ respectively) shows that one can always increase the overall hypervolume indicator value if the outmost point is shifted, see also Fig. 16.11. For the technical details, including a technical lemma, we refer to Appendix 16.7.3 on page 224. □

**16.4.2 Example**
*Consider the test problem ZDT1 [378] with a Pareto front described by $f(x) = 1 - \sqrt{x}$ with $x_{\min} = 0$ and $x_{\max} = 1$, see Figure 16.9(a). The derivative $f'(x) = -1/(2\sqrt{x})$ equals $-\infty$ at the left extreme $x_{\min}$ and the left extreme is therefore never included in an optimal $\mu$-distribution for $I_{H,w}$ according to Theorem 16.4.1.*

Although one should keep the previous result in mind when using the hypervolume indicator, the fact that the extreme can never be obtained in the cases of Theorem 16.4.1 is less restrictive in practice. Due to the continuous search space for most of the test problems, no algorithm will obtain a specific solution exactly—and the extreme in particular—and if the number of points is high enough, a solution close to the extreme[12] will be found also by hypervolume-based algorithms. However, if the number of points is low the choice of the reference point is crucial and choosing it too close to the nadir point will massively change the optimal $\mu$-distribution as can be seen exemplary for the ZDT1 problem in Fig. 16.5[13]. Moreover, when using the weight function in the weighted hypervolume indicator to model

[12] Although the distance of solutions to the extremes might be sufficiently small in practice also for the scenario of Theorem 16.4.1, the theoretical result shows that for a finite $\mu$, we cannot expect that the solutions approach the extremes arbitrarily close.

[13] The shown approximations of the optimal $\mu$-distribution have been obtained by using the algorithm CMA-ES [180, version 3.40beta with standard settings] to solve the 2-dimensional optimization problem of Remark 16.3.7 with the two leftmost points as variables and a boundary handling with penalties if the leftmost or rightmost point is outside $[x_{\min}, x_{\max}]$ (population size 20, best result over 100 runs shown).

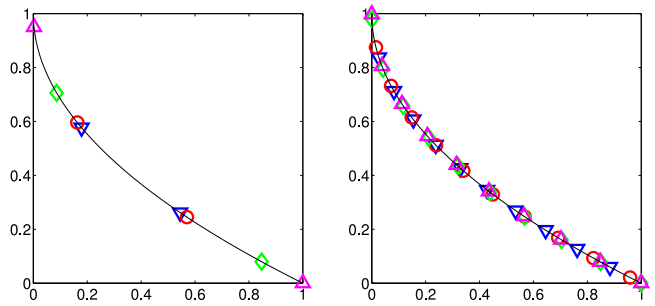


Figure 16.5: Influence of the choice of the reference point $r = (r_1, r_2)$ on optimal 2- (left) and optimal 10-distributions on the ZDT1 problem, in particular on the left extreme. Shown are the best approximations found within 100 CMA-ES runs for $r = (1.01, 1.01)$ ($\triangledown$), $r = (1.1, 1.1)$ ($\bigcirc$), $r = (2, 2)$ ($\diamond$), and $r = (11, 11)$ ($\triangle$). Note that according to theory, the left extreme is never included in optimal $\mu$-distributions and the lower bound on $r_1$ to ensure the right extreme is $\mathcal{R}_1 = 3$ [25].

preferences of the user towards certain regions of the objective search, one should pay attention to this fact by increasing the weight drastically close to such extremes if they are desired, see [24] for examples.

### Lower Bound for Choosing the Reference Point for Obtaining the Extremes

We have seen in the previous section that if the limit of the derivative of the front at the left extreme equals $-\infty$ (resp. if the derivative of the front at the right extreme equals zero) there is no choice of reference point that allows to have the extremes included in optimal $\mu$-distributions for $I_{H,w}$. We assume now that the limit of the derivative of the front at the left extreme is finite (resp. the derivative of the front at the right extreme is not zero) and investigate conditions ensuring that there exists (finite) reference points ensuring to have the extremes in the optimal $\mu$-distributions.

### Lower Bound for Left Extreme

**16.4.3 Theorem (Lower bound for left extreme)**
*Let $\mu$ be an integer larger or equal 2. Assume that $f$ is continuous on $[x_{\min}, x_{\max}]$, non-increasing, differentiable on $]x_{\min}, x_{\max}[$ and that $f'$ is continuous on $]x_{\min}, x_{\max}[$ and $\lim_{x \to x_{\min}} -f'(x) < \infty$. If there exists a $\mathcal{K}_2 \in \mathbb{R}$ such that for all $x_1 \in ]x_{\min}, x_{\max}]$*

$$\int_{f(x_1)}^{\mathcal{K}_2} w(x_1, y)dy > -f'(x_1) \int_{x_1}^{x_{\max}} w(x, f(x_1))dx \ , \tag{16.13}$$

*then for all reference points $r = (r_1, r_2)$ such that $r_2 \geq \mathcal{K}_2$ and $r_1 > x_{\max}$, the leftmost extremal point is contained in optimal $\mu$-distributions for $I_{H,w}$. In other words, defining $\mathcal{R}_2$ as*

$$\mathcal{R}_2 = \inf\{\mathcal{K}_2 \text{ satisfying } Eq.\ 16.13\} \ , \tag{16.14}$$

*the leftmost extremal point is contained in optimal $\mu$-distributions if $r_2 > \mathcal{R}_2$, and $r_1 > x_{\max}$.*

*Proof.* This proof is presented in Appendix 16.7.4 on page 226. □

**16.4.4 Remark**
*The previous theorem states only an implicit condition for $\mathcal{K}_2$ and it is not always obvious whether a finite $\mathcal{K}_2$ with the stated properties exists. There are different reasons for a non-existence of a finite $\mathcal{K}_2$—although we assume that $\lim_{x \to x_{\min}} -f'(x) < \infty$. One reason can be the fact that $f'(x_1)$ is infinite for*

*some $x_1 \in ]x_{\min}, x_{\max}]$ such that the right-hand side of Eq. 16.13 is not finite and therefore $\mathcal{K}_2$ cannot be finite as well. Example 16.4.6, however, shows an example where $f'(x_1) = -\infty$ for an $x_1 \in ]x_{\min}, x_{\max}]$ and $\mathcal{K}_2$ is still finite. Another possible reason for the non-existence of a finite $\mathcal{K}_2$ can be a choice of $w$ such that the left-hand side of Eq. 16.13 is always smaller than the right-hand side—even assuming that $w$ is continuous does not prevent such a choice of $w$.*

We will now apply the previous theorem to the unweighted hypervolume and prove an *explicit* lower bound for setting the reference point so as to have the left extreme. This results recovers [25, Theorem 2].

**16.4.5 Corollary (Lower bound for left extreme)**
*Let $\mu$ be an integer larger or equal 2. Assume that $f$ is continuous on $[x_{\min}, x_{\max}]$, non-increasing, differentiable on $]x_{\min}, x_{\max}[$ and that $f'$ is continuous on $[x_{\min}, x_{\max}[$. Let us assume that $\lim_{x \to x_{\min}} -f'(x) < \infty$. If*

$$\mathcal{R}_2 = \sup\{f'(x)(x - x_{\max}) + f(x) : x \in ]x_{\min}, x_{\max}]\} \tag{16.15}$$

*is finite, then the leftmost extremal point is contained in optimal $\mu$-distributions for $I_H$ if the reference point $r = (r_1, r_2)$ is such that $r_2$ is strictly larger than $\mathcal{R}_2$ and $r_1 > x_{\max}$.*

*Proof.* The proof is presented in Appendix 16.7.5 page 227. □

**16.4.6 Example**
*Consider again the DTLZ2 test function from Example 16.2.1 with $f(x) = \sqrt{1 - x^2}$ and $f'(x) = -\frac{x}{\sqrt{1-x^2}}$ where $x_{\min} = 0$ and $x_{\max} = 1$. Assume $w = 1$, i.e., the unweighted hypervolume indicator $I_H$. We see that $f'(x_{\max}) = -\infty$ but nevertheless, $\mathcal{R}_2$ is finite according to Eq. 16.15, namely*

$$\mathcal{R}_2 = \sup\left\{-\frac{x}{\sqrt{1 - x^2}}(x - x_{\max}) + \sqrt{1 - x^2} : x \in ]x_{\min}, x_{\max}]\right\} = \sqrt{6\sqrt{3} - 9} \approx 1.18 \ ,$$

*which can be obtained for example with a computer algebra system such as Maple.*

**Lower Bound for Right Extreme**
We now turn to the case of the right extreme and address the same question as for the left extreme: assuming that $f'(x_{\max}) \neq 0$, can we find an explicit lower bound for the first coordinate of the reference point ensuring that the right extreme is included in optimal $\mu$-distributions? The following result holds.

**16.4.7 Theorem (Lower bound for right extreme)**
*Let $\mu$ be an integer larger or equal 2. Assume that $f$ is continuous on $[x_{\min}, x_{\max}]$, non-increasing, differentiable on $]x_{\min}, x_{\max}[$ and that $f'$ is continuous on $]x_{\min}, x_{\max}[$ and $f'(x_{\max}) \neq 0$. If there exists a $\mathcal{K}_1 \in \mathbb{R}$ such that for all $x_\mu \in [x_{\min}, x_{\max}[$*

$$-f'(x_\mu) \int_{x_\mu}^{\mathcal{K}_1} w(x, f(x_\mu))dx > \int_{f(x_\mu)}^{f(x_{\min})} w(x_\mu, y)dy \ , \tag{16.16}$$

*then for all reference points $r = (r_1, r_2)$ such that $r_1 \geq \mathcal{K}_1$ and $r_2 > f(x_{\min})$, the rightmost extremal point is contained in optimal $\mu$-distributions. In other words, defining $\mathcal{R}_1$ as*

$$\mathcal{R}_1 = \inf\{\mathcal{K}_1 \text{ satisfying Eq. } 16.16\} \ , \tag{16.17}$$

*the rightmost extremal point is contained in optimal $\mu$-distributions if $r_1 > \mathcal{R}_1$, and $r_2 > f(x_{\min})$.*

*Proof.* This proof is presented in Appendix 16.7.6 on page 228. □

We will now apply the previous theorem to the unweighted hypervolume and prove an explicit lower bound for setting the reference point so as to have the right extreme. This results recovers [25, Theorem 2].

**16.4.8 Corollary (Lower bound for right extreme)**
*Let $\mu$ be an integer larger or equal 2. Assume that $f$ is continuous on $[x_{\min}, x_{\max}]$, non-increasing, differentiable on $]x_{\min}, x_{\max}[$ and that $f'$ is continuous and strictly negative on $]x_{\min}, x_{\max}]$. If*

$$\mathcal{R}_1 = \sup\left\{x + \frac{f(x) - f(x_{\min})}{f'(x)} : x \in [x_{\min}, x_{\max}[\right\} \tag{16.18}$$

*is finite, then the rightmost extremal point is contained in optimal $\mu$-distributions for $I_H$ if the reference point $r = (r_1, r_2)$ is such that $r_1 > \mathcal{R}_1$ and $r_2 > f(x_{\min})$.*

*Proof.* The proof is presented in Appendix 16.7.7 page 228. □

## 16.4.2 Number of Points Going to Infinity

The lower bounds we have derived for the reference point such that the extremes are included are independent of $\mu$. It can be seen in the proof that those bounds are not tight if $\mu$ is larger than 2. Deriving tight bounds is, however, difficult because it would require to know for a given $\mu$ where the second point of optimal $\mu$-distributions is located. It can be certainly achieved in the linear case, but it might be impossible in more general cases. However, we want to investigate now how $\mu$ influences the choice of the reference point so as to have the extremes. In this section, we will denote $\mathcal{R}_1^{\text{Nadir}}$ and $\mathcal{R}_2^{\text{Nadir}}$ the first and second coordinates of the nadir point, namely $\mathcal{R}_1^{\text{Nadir}} = x_{\max}$ and $\mathcal{R}_2^{\text{Nadir}} = f(x_{\min})$.

We will prove that for any reference point dominated by the nadir point, there exists a $\mu_0$ such that for all $\mu$ larger than $\mu_0$, optimal $\mu$-distributions associated to this reference point include the extremes in case the extremes can be contained in optimal $\mu$-distributions, i.e., if $-f'(x_{\min}) < \infty$ and $f'(x_{\max}) < 0$. Before, we establish a lemma saying that if there exists a reference point $R^1$ allowing to have the extremes, then all reference points $R^2$ dominated by this reference point $R^1$ will also allow to have the extremes.

**16.4.9 Lemma**
*Let $R^1 = (r_1^1, r_2^1)$ and $R^2 = (r_1^2, r_2^2)$ be two reference points with $r_1^1 < r_1^2$ and $r_2^1 < r_2^2$. If both extremes are included in optimal $\mu$-distributions for $I_{H,w}$ associated with $R^1$ then both extremes are included in optimal $\mu$-distributions for $I_{H,w}$ associated with $R^2$.*

*Proof.* The proof is presented in Appendix 16.7.8 page 229. □

**16.4.10 Theorem**
*Let us assume that $f$ is continuous, differentiable with $f'$ continuous on $[x_{\min}, x_{\max}]$, $f'(x_{\max}) < 0$, and $w$ is bounded, i.e., there exists $W > 0$ such that $w(x, y) \leq W$ for all $(x, y)$. For all $\varepsilon = (\varepsilon_1, \varepsilon_2) \in \mathbb{R}^2_{>0}$,*

1. *there exists a $\mu_1$ such that for all $\mu \geq \mu_1$, and any reference point $R$ dominated by the nadir point such that $R_2 \geq \mathcal{R}_2^{\text{Nadir}} + \varepsilon_2$, the left extreme is included in optimal $\mu$-distributions,*
2. *there exists a $\mu_2$ such that for all $\mu \geq \mu_2$, and any reference point $R$ dominated by the nadir point such that $R_1 \geq \mathcal{R}_1^{\text{Nadir}} + \varepsilon_1$, the right extreme is included in optimal $\mu$-distributions.*

*Proof.* The proof is presented in Appendix 16.7.9 page 229. □

As a corollary, we obtain the following result for obtaining both extremes simultaneously:

**16.4.11 Corollary**
*Let us assume that $f$ is continuous, differentiable with $f'$ continuous on $[x_{\min}, x_{\max}]$, $f'(x_{\max}) < 0$, and $w$ is bounded, i.e., there exists a $W > 0$ such that $w(x, y) \leq W$ for all $(x, y)$. For all $\varepsilon = (\varepsilon_1, \varepsilon_2) \in \mathbb{R}^2_{>0}$, there exists a $\mu_0 \in \mathbb{N}$ such that for $\mu$ larger than $\mu_0$ and for all reference points weakly dominated by $(\mathcal{R}_1^{\text{Nadir}} + \varepsilon_1, \mathcal{R}_2^{\text{Nadir}} + \varepsilon_2)$, both the left and right extremes are included in optimal $\mu$-distributions.*

*Proof.* The proof is straightforward taking for $\mu_0$ the maximum of $\mu_1$ and $\mu_2$ in Theorem 16.4.10. □

Theorem 16.4.10 and Corollary 16.4.11 state that for bi-objective Pareto fronts which are continuous on the interval $[x_{\min}, x_{\max}]$ and a bounded weight, we can expect to have the extremes in optimal $\mu$-distributions for any reference point dominated by the nadir point if $\mu$ is large enough, i.e., larger than $\mu_0$. Unfortunately, the proof does not allow to state how large $\mu_0$ has to be chosen for a given reference point but it is expected that $\mu_0$ depends on the reference point as well as on the front shape and $w$. Recently, for linear Pareto fronts, this dependency could be shown explicitly and we will briefly summarize this result in the following.

## 16.5 Application to Multiobjective Test Problems

Besides being used within indicator-based algorithms, the hypervolume indicator has been also frequently used for performance assessment when comparing multiobjective optimizers—mainly because of its refinement property [383] and its resulting ability to map both information about the proximity of a solution set to the Pareto front and about the set's spread in objective space into a single scalar. Also here, knowing the optimal $\mu$-distribution and its corresponding hypervolume value for certain test problems is crucial. On the one hand, knowing the largest hypervolume value obtainable by $\mu$ solutions allows to compare the achieved hypervolume values of different algorithms not only relatively but also absolutely in terms of the difference between the achieved and the achievable hypervolume value. On the other hand, only knowing the actual optimal $\mu$-distributions for a certain test problem allows to investigate whether hypervolume-based algorithms really converge to their inherent optimization goal (or get stuck in local optima of (16.3) and (16.4)) which has not been investigated yet. In this section, we therefore apply the theoretical concepts derived in Sections 16.3 and 16.4 to several known test problems. First, we recapitulate results from [25] and in Sec. 16.5.1 and investigate optimal $\mu$-distributions for the unweighted hypervolume indicator $I_H$ in case of a linear Pareto front. Then, we apply the results to the test function suites ZDT, DTLZ, and WFG in Sec. 16.5.2.

### 16.5.1 Linear Fronts

In this section, we have again a closer look at linear Pareto fronts, i.e., fronts that can be formally defined as $f : x \in [x_{\min}, x_{\max}] \mapsto \alpha x + \beta$ where $\alpha < 0$ and $\beta \in \mathbb{R}$. For linear fronts with slope $\alpha = -1$, [60] (and later on [137] for a more restricted front of shape $f(x) = 1 - x$) already proved that a set of $\mu$ points maximizes the unweighted hypervolume if and only if the points are equally spaced. However, the used proof techniques do not allow to state where the leftmost and rightmost point have to be placed in order to maximize the hypervolume with respect to a certain reference point—an assumption that later results do not require [25]. We will recapitulate those recent results briefly and in particular show for linear fronts of arbitrary slope, how the—in this case unique—optimal $\mu$-distribution for $I_H$ looks like without making assumptions on the positions of extreme solutions.

First of all, we formalize the result of Example 16.3.8 that, as a direct consequence of Corollary 16.3.6, the distance between two neighbored solutions is constant for arbitrary linear fronts:


Figure 16.6: Optimal $\mu$-distribution for $\mu = 4$ points and the unweighted hypervolume indicator if the reference point is not dominated by the extreme points of the Pareto front (Theorem 16.5.2, left) and in the most general case (Theorem 16.5.3, right) for a front with slope $f'(x) = \alpha = -\frac{1}{3}$. The dotted lines in the right plot limit the regions where the leftmost point, the rightmost point, or both are included in the optimal $\mu$-distributions for $\mu = 4$ (see also Fig. 16.7).

**16.5.1 Theorem**
*If the Pareto front is a (connected) line, the optimal $\mu$-distribution with respect to the unweighted hypervolume indicator is such that the distance is the same between all neighbored solutions.*

*Proof.* Applying Eq. 16.7 to $f(x) = \alpha x + \beta$ implies that $\alpha(\overline{x}^\mu_{i+1} - \overline{x}^\mu_i) = f(\overline{x}^\mu_i) - f(\overline{x}^\mu_{i-1}) = \alpha(\overline{x}^\mu_i - \overline{x}^\mu_{i-1})$ for $i = 2, \ldots, \mu - 1$ and therefore the distance between consecutive points of the optimal $\mu$-distribution for $I_H$ is constant. □

Moreover, in case the reference point is not dominated by the extreme points of the Pareto front, i.e., $r_1 < x_{\max}$ and $r_2$ is such that there exists (a unique) $\overline{x}^\mu_0 \in [x_{\min}, x_{\max}]$ with $\overline{x}^\mu_0 = f^{-1}(r_2)$, the exact position of the optimal $\mu$-distribution for $I_H$ on the linear front can be determined, see also the left plot of Fig. 16.6:

**16.5.2 Theorem**
*If the Pareto front is a (connected) line and the reference point $(r_1, r_2)$ is not dominated by the extremes of the Pareto front, the optimal $\mu$-distribution with respect to the unweighted hypervolume indicator is unique and satisfies for all $i = 1, \ldots, \mu$*

$$\overline{x}^\mu_i = f^{-1}(r_2) + \frac{i}{\mu + 1} \cdot (r_1 - f^{-1}(r_2)) \ . \tag{16.19}$$

*Proof.* From Eq. 16.7 and the previous proof we know that $\alpha\left(\overline{x}^\mu_{i+1} - \overline{x}^\mu_i\right) = f(\overline{x}^\mu_i) - f(\overline{x}^\mu_{i-1}) = \alpha(\overline{x}^\mu_i - \overline{x}^\mu_{i-1})$ , for $i = 1, \ldots, \mu$ while $f(\overline{x}^\mu_0) = r_2$ and $\overline{x}^\mu_{\mu+1} = r_1$ are defined as in Corollary 16.3.6; in other words, the distances between $\overline{x}^\mu_i$ and its two neighbors $\overline{x}^\mu_{i-1}$ and $\overline{x}^\mu_{i+1}$ are the same for each $1 \leq i \leq \mu$. Therefore, the points $(\overline{x}^\mu_i)_{1 \leq i \leq \mu}$ partition the interval $[\overline{x}^\mu_0, \overline{x}^\mu_{\mu+1}]$ into $\mu + 1$ sections of equal size and we obtain Eq. 16.19. □

Although Theorem 16.5.2 proves the exact unique positions of the $\mu$ points maximizing the unweighted hypervolume indicator in the restricted case where the reference point $r$ is not dominated by the extremes of the front, the result can be used to obtain the exact distributions also in the most general case for any reasonable[14] choice of the reference point and any $\mu \in \mathbb{N}$ if the linear front is defined in the interval $[0, x_{\max}]$[15].

**16.5.3 Theorem**
*Given $\mu \in \mathbb{N}_{\geq 2}$, $\alpha \in \mathbb{R}_{<0}$, $\beta \in \mathbb{R}_{>0}$, and a linear Pareto front $f(x) = \alpha x + \beta$ within $[0, x_{\max} = -\frac{\beta}{\alpha}]$, the unique optimal $\mu$-distribution $(\overline{x}^\mu_1, \ldots, \overline{x}^\mu_\mu)$ for the unweighted hypervolume indicator $I_H$ with reference point*

[14] Again, choosing the reference point such that it dominates Pareto-optimal points does not make sense as no solution will have positive hypervolume contributions.
[15] Assuming $x_{\min} = 0$ is not a restriction as the result for other choices of $x_{\min}$ can be derived by a simple coordinate transformation.

*$(r_1, r_2) \in \mathbb{R}^2_{>0}$ can be described by*

$$\overline{x}_i^\mu = f^{-1}(F_l) + \frac{i}{\mu+1}\left(F_r - f^{-1}(F_l)\right) \tag{16.20}$$

*for all $1 \leq i \leq \mu$ where*

$$F_l = \min\left\{r_2,\ \frac{\mu+1}{\mu}\beta - \frac{1}{\mu}f(r_1),\ \frac{\mu}{\mu-1}\beta\right\} \quad \text{and} \quad F_r = \min\left\{r_1,\ \frac{\mu+1}{\mu}x_{\max} - \frac{1}{\mu}f^{-1}(r_2),\ \frac{\mu}{\mu-1}x_{\max}\right\}$$

*if the reference point is dominated by at least one Pareto-optimal point.*

*Proof.* The proof idea is the following. We can elongate the linear front beyond $x_{\min}$ and $x_{\max}$ and use the result of Theorem 16.5.2 to obtain the optimal placement dependent on $r_1$ and $r_2$—keeping in mind that all points are restricted to the interval $[x_{\min}, x_{\max}]$. In case $r_1$ and $r_2$ are too far away from the nadir point $(x_{\max}, \beta)$ such that Theorem 16.5.2 gives us $\overline{x}_1^\mu < x_{\min}$ or $\overline{x}_\mu^\mu > x_{\max}$, we have to make sure that these constraints are fulfilled by restricting the values $F_l$ and $F_r$ in Eq. 16.20 accordingly. □

Right from the technicalities in the proof of Theorem 16.5.3 we see for which choices of the reference point the left and/or the right extreme are contained in the optimal $\mu$-distribution.

**16.5.4 Corollary**
*Given $\mu \in \mathbb{N}_{\geq 2}$, $\alpha \in \mathbb{R}_{<0}$, $\beta \in \mathbb{R}_{>0}$, and a linear Pareto front $f(x) = \alpha x + \beta$ within $[0, x_{\max} = -\frac{\beta}{\alpha}]$,*

- *the left extreme point $(0, \beta)$ is included in the optimal $\mu$-distribution for the unweighted hypervolume indicator if the reference point $(r_1, r_2) \in \mathbb{R}^2_{>0}$ lies above the line $L(x) = \frac{\mu+1}{\mu}\beta - \frac{1}{\mu}f(x) = \beta - \frac{\alpha}{\mu}x$ or if $r_2 > \frac{\mu}{\mu-1}\beta$ and*
- *the right extreme point $(x_{\max}, 0)$ is included if the reference point lies below the line $R(x) = \frac{\mu+1}{\mu}x_{\max} - \frac{1}{\mu}f^{-1}(x) = -\alpha\mu x - \mu\beta$ or if $r_1 > \frac{\mu}{\mu-1}x_{\max}$.*

Figure 16.7 gives an example for the front $f(x) = 2 - \frac{x}{3}$ and shows the regions within which the reference point ensures the left and/or the right extreme of the front for various choices of $\mu$. Note that in the specific case of linear Pareto fronts, we not only know that the reference point to obtain both extremes approaches the nadir point if $\mu$ goes to infinity as proven in Sec. 16.4.2 but with the previous corollary, we also know *how fast* this happens.

As pointed out before, we do not know in general whether an optimal $\mu$-distribution for a given indicator is unique or not. The example of a linear front is a case where we can ensure the uniqueness due to the concavity of the hypervolume indicator [62]. Note also that besides for linear fronts, only one front shape is known so far for which we can also determine optimal $\mu$-distributions exactly: for front shapes of the form $f(x) = \beta/x$ with $\beta > 1$, $x_{\min} = -\beta$, and $x_{\max} = -1$ and when the reference point is in $(0, 0)$ [156]. On the other hand, even in the case of convex Pareto fronts, examples are known where the hypervolume indicator is not concave anymore and therefore the uniqueness of optimal $\mu$-distributions is not known [62].

### 16.5.2 Test Function Suites ZDT, DTLZ, and WFG

In this section, we apply the presented results to problems in the ZDT [378], the DTLZ [119], and the WFG [193] test function suites. All results are derived for the unweighted case of $I_H$, but they can also be derived for any other weight function $w(x, y) \neq 1$. In particular, we derive the function $f(x)$

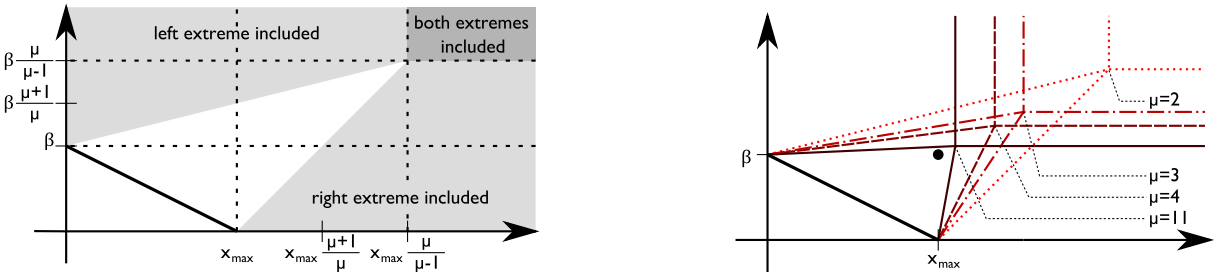



Figure 16.7: Influence of the reference point on the extremes for problems with linear Pareto fronts: the left plot shows the different regions within which the reference point ensures one (light gray), both (dark gray) or none (white) of the extremes in the optimal $\mu$-distribution for $\mu = 2$ and the example front of $f(x) = 2 - \frac{x}{3}$. The right plot shows the borders of these regions for $\mu = 2$ (dotted), $\mu = 3$ (dash-dotted), $\mu = 4$ (dashed), and $\mu = 11$ (solid) for the same front. For clarity, the nadir point is shown as a black circle.

describing the Pareto front and its derivative $f'(x)$ which directly leads to the density $\delta_F(x)$ with constant $C$. Furthermore, we derive a lower bound $\mathcal{R}$ for the choice of the reference point such that the extremes are included and compute an approximation of the optimal $\mu$-distribution for $\mu = 20$ points. For the latter, the approximation schemes as proposed by [25] are used to get a precise picture for a given $\mu$[16]. The densities and the lower bounds $\mathcal{R}$ for the reference point are obtained by the commercial computer algebra system Maple 12.0.

Figure 16.8 summarizes the results on the density and the lower bounds for the reference point for all investigated problems whereas we refer to the appendix for more detailed derivations (Appendix 16.7.10 presents the ZDT, Appendix 16.7.11 the DTLZ, and Appendix 16.7.12 the WFG results). Moreover, Fig. 16.9 shows a plot of the Pareto front, the obtained approximation of an optimal $\mu$-distribution for $\mu = 20$, and the derived density $\delta_F(x)$ (as the hatched area on top of the front $f(x)$) for all investigated test problems.

The presented results show that for several of the considered test problems, analytical results for the density and the lower bounds for the reference point can be given easily—at least if a computer algebra system such as Maple is used. Otherwise, numerical results can be provided that approximate the mathematical results with an arbitrary high precision (up to the machine precision) which also holds for the approximations of the optimal $\mu$-distributions shown in Fig. 16.9. Note that in the latter case, the approximation schemes used do not guarantee that the actual maximum of Eq. 16.3 and Eq. 16.4 is found as already discussed by [25]. However, the distributions shown in Fig. 16.9 have been cross-checked by using the robust stochastic search optimizer CMA-ES [180] in a similar manner as for the plots in Fig. 16.5. Moreover, the resulting optimal $\mu$-distributions are independent of the starting conditions of the approximation schemes which is a strong indicator that the distributions found are indeed good approximations of the optimal distributions of $\mu$ points [25].

Last, we give an additional interpretation of the density results: the density not only gives information about the bias of the hypervolume indicator for a given front, but can also be used to assess the number of solutions to be expected on a given segment of the front, as the following example illustrates.

### 16.5.5 Example

*Consider again ZDT2 as in Example 16.3.11. We would like to answer the question what is the fraction of points $r_F$ of an optimal $\mu$-distribution with the first and second objective being smaller or equal* 0.5 *and* 0.95 *respectively, see the highlighted front part in Figure 16.10. From $f^{-1}(y) = \sqrt{1-y}$ and $f^{-1}(0.95) = \sqrt{0.05}$ follows, that for the considered front segment $x \in [\sqrt{0.05}, 0.5]$ holds. Using $\delta(x)$ given*

[16]For the test suites ZDT and DTLZ, additional approximations of the optimal $\mu$-distribution for other typical numbers of points can be downloaded at `http://www.tik.ee.ethz.ch/sop/mudistributions`.

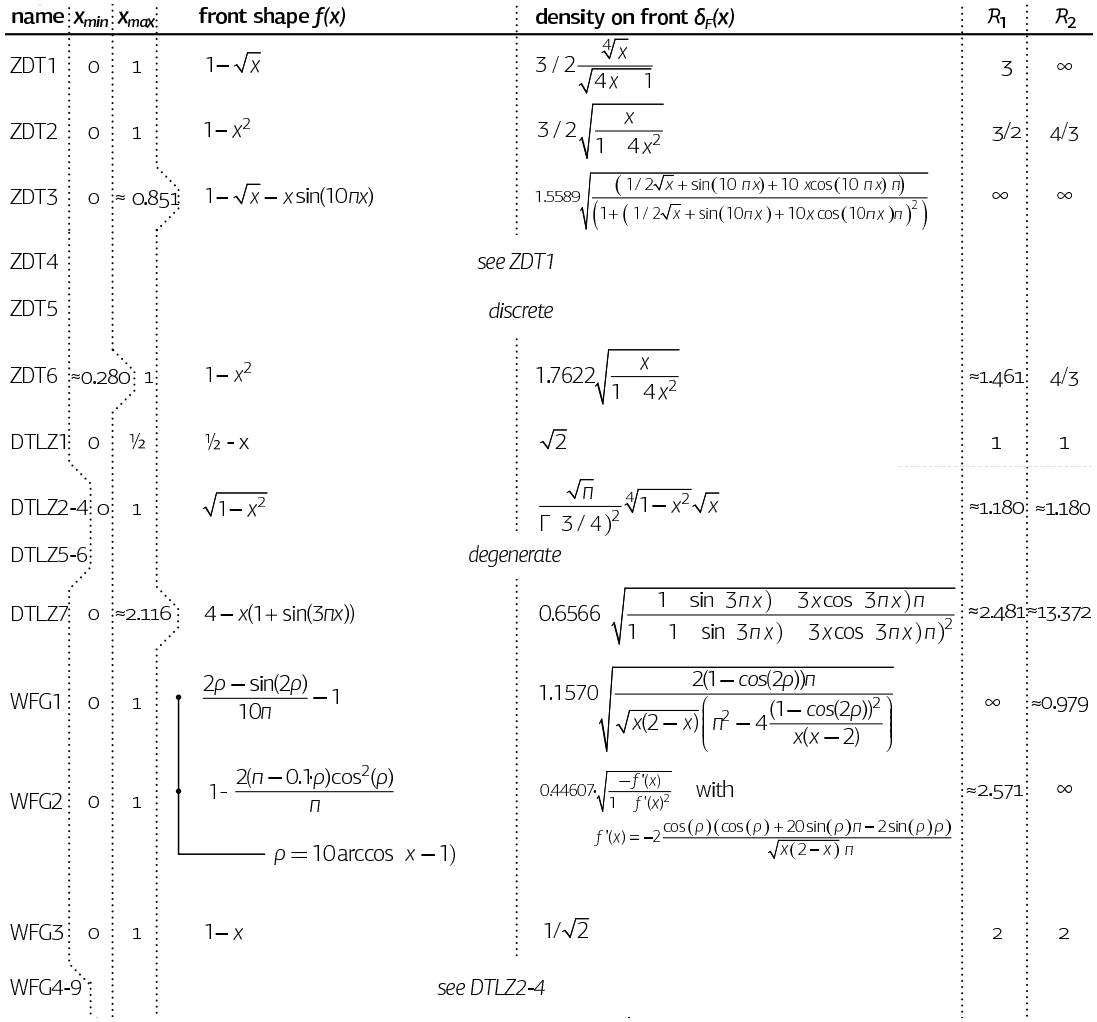


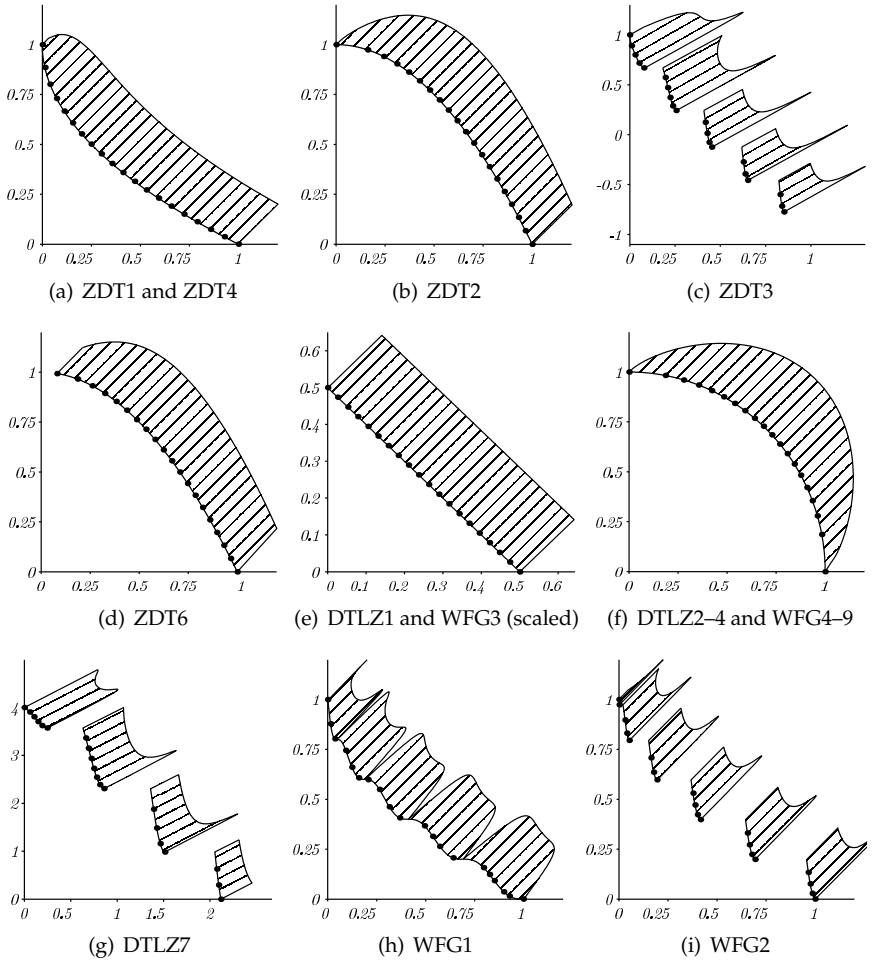

Figure 16.8: Lists for all ZDT, DTLZ, and WFG test problems and the unweighted hypervolume indicator $I_H$: (i) the Pareto front as $x \in [x_{\min}, x_{\max}] \mapsto f(x)$, (ii) the density $\delta_F(x)$ on the front according to Eq. 16.12, and (iii) a lower bound $\mathcal{R} = (\mathcal{R}_1, \mathcal{R}_2)$ of the reference point to obtain the extremes (Eq. 16.18 and 16.15 respectively).

*in Example 16.3.11 and integrating over* $[\sqrt{0.05}, 0.5]$ *yields:*

$$r_F = \int_{\sqrt{0.05}}^{0.5} \delta(x)dx = \int_{\sqrt{0.05}}^{0.5} \frac{3}{2}\sqrt{x}dx = \frac{1}{4}\sqrt{2} - 0.05^{3/4} \approx 24.78\% \ .$$

*The same result can be obtained by taking the line integral of the density on the front over the considered front segment. Let* $\delta_F^s(x, f(x)) := \delta_F(x)$ *denote the density on the front for a given point* $(x, f(x))$*, then* $r_F = \int_\gamma \delta_F^s(s)ds = \int_a^b \delta_F^s(\gamma(t))\,\|\dot{\gamma}(t)\|_2\,dt$ *where the path* $\gamma$ *denotes the considered line segment on the front, i.e.,* $\gamma : [a = \sqrt{0.05}, b = 0.5] \to \mathbb{R}^2$, $t \mapsto (t, 1-t^2)$*. With* $\|\dot{\gamma}(t)\|_2 = \sqrt{1+f'(t)^2}$ *and* $\delta_F(\gamma(t)) = \delta_F(t)$ *we have* $r_F = \int_{\sqrt{0.05}}^{0.5} \delta_F(t)\sqrt{1+f'(t)^2}dt = \int_{\sqrt{0.05}}^{0.5} \delta(t)dt \approx 24.78\%$*. Note that for the approximated optimal* $\mu$*-distribution of a finite number of* $\mu = 100$ *points*[17] *we obtained* 24 *points in the considered line segment, which is close to the predicted percentage of* $r_F = 24.78\%$*.*

[17] see `http://www.tik.ee.ethz.ch/sop/download/supplementary/testproblems/zdt2/data/mu100.txt`



(a) ZDT1 and ZDT4 (b) ZDT2 (c) ZDT3

(d) ZDT6 (e) DTLZ1 and WFG3 (scaled) (f) DTLZ2–4 and WFG4–9

(g) DTLZ7 (h) WFG1 (i) WFG2

Figure 16.9: Pareto front shape $f(x)$, approximate optimal distribution of 20 points (black dots) for the unweighted hypervolume indicator, and the density $\delta_F(x)$ (hatched area) for different test problems.

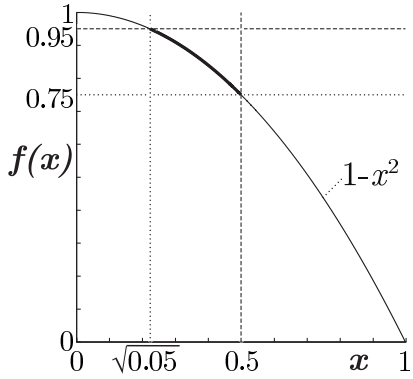



Figure 16.10: The density of points $\delta(x)$ and $\delta_F(x)$ can be used to assess the number of points to be expected in a given part of the front. The plot shows the thick line segment of the Pareto-front of ZDT2 for which $f(x) \leq 0.95$ and $x \leq 0.5$ hold, see Example 16.5.5.

## 16.6 Conclusions

Indicator-based evolutionary algorithms transform a multiobjective optimization problem into a single-objective one that corresponds to finding a set of $\mu$ points that maximizes the underlying quality indicator. Theoretically understanding these so-called *optimal $\mu$-distributions* for a given indicator is a fundamental issue both for performance assessment of multiobjective optimizers and for the decision which indicator to take for the optimization in practice such that the search bias introduced by the indicator meets the user's preferences.

In this report, we theoretically characterize optimal $\mu$-distributions for the weighted hypervolume indicator in case of bi-objective problems. The results generalize previous work on the unweighted hypervolume indicator and are, in addition, applied to several known test problems. In particular, we investigate the sets of $\mu$ points that maximize the weighted hypervolume indicator and, besides general investigations for finite $\mu$, we derive a limit result for $\mu$ going to infinity in terms of a density of points. Furthermore, we investigate the influence of the reference point on optimal $\mu$-distributions, i.e., we derive lower bounds for placing the reference point (possibly infinite) for guaranteeing the Pareto front's extreme points in an optimal $\mu$-distribution and investigate cases where the extremes are never contained in an optimal $\mu$-distribution. In addition, we show that the belief, the best choice for the reference point corresponds to a point that is slightly worse than the nadir point in all objectives, can be founded theoretically if the number of points goes to infinity. Last, we apply the theoretical results to problems of the ZDT, DTLZ, and WFG test problem suites resulting in recommended choices of the reference point including numerical and sometimes analytical expressions for the resulting density of points on the front.

We believe the results presented in this report are important for several reasons. On the one hand, we prove that several previous beliefs about the bias of the hypervolume indicator and the choice of the reference point to obtain the extremes of the front have been wrong. On the other hand, the results on optimal $\mu$-distributions are highly useful in performance assessment if the hypervolume indicator is used as a quality measure. For the first time, approximations of optimal $\mu$-distributions for finite $\mu$ allow to compare the outcome of indicator-based evolutionary algorithms to the actual optimization goal. Moreover, the actual hypervolume indicator of optimal $\mu$-distributions (or the approximations we provide) offers a way to interpret the obtained hypervolume indicator values in an absolute fashion as the hypervolume of an optimal $\mu$-distribution is a better estimate of the best achievable hypervolume than the hypervolume of the entire Pareto front. Last, we would like to mention that the presented results for the weighted hypervolume indicator also provide a basis for a better understanding of how to articulate user preferences with the weighted hypervolume indicator in terms of the question on how to choose the weight function in practice.

## 16.7 Appendix

### 16.7.1 Proof of Theorem 16.3.4 stated on page 205

Before to prove the result, we rewrite Eq. 16.3 (page 201) in the following way

$$I_{H,w}(x_1, \ldots, x_\mu) = \sum_{i=1}^{\mu} g(x_i, x_{i+1}) \ , \tag{16.21}$$

where $g$ is the 2-dimensional function defined as

$$g(\alpha, \beta) = \int_{\alpha}^{\beta} \left( \int_{f(\alpha)}^{f(x_0)} w(x, y) dy \right) dx \ . \tag{16.22}$$

The derivation of the gradient of $I_{H,w}$ thus relies on computing the partial derivatives of $g$. The following lemma gives the expressions of the partial derivatives of $g$:

**16.7.1 Lemma**
*Let $w$ be a weight function for the weighted hypervolume indicator $I_{H,w}$ and $f : [x_{\min}, x_{\max}] \to \mathbb{R}$ be a continuous and differentiable function describing a 2-dimensional Pareto front. Let $g$ be defined as $g(\alpha, \beta) = \int_\alpha^\beta \left( \int_{f(\alpha)}^{f(x_0)} w(x, y)dy \right) dx$ where $f(x_0) = r_2$. Then,*

$$\partial_1 g(\alpha, \beta) = -f'(\alpha) \int_\alpha^\beta w(x, f(\alpha))dx - \int_{f(\alpha)}^{f(x_0)} w(\alpha, y)dy \tag{16.23}$$

$$\partial_2 g(\alpha, \beta) = \int_{f(\alpha)}^{f(x_0)} w(\beta, y)dy \tag{16.24}$$

*Proof.*

To compute the first partial derivative of $g$, we need to compute the derivative of the function $g_1 : \alpha \to g(\alpha, \beta)$. Let us define $\gamma(l, m) = \int_{f(m)}^{f(x_0)} w(l, y)dy$ such that $g_1(\alpha) = \int_\alpha^\beta \gamma(x, \alpha)dx$. Define $K(\bar{x}, \bar{y}) = \int_{\bar{x}}^\beta \gamma(x, \bar{y})dx$ and be $\Phi : \alpha \in \mathbb{R} \to (\alpha, \alpha) \in \mathbb{R}^2$. Then $g_1(\alpha) = K \circ \Phi(\alpha)$ such that we can apply the chain rule to find the derivative of $g_1$. Since $g_1$ maps $\mathbb{R}$ into $\mathbb{R}$, the differential of $g_1$ in $\alpha$ applied in $h$ equals the derivative of $g_1$ in alpha times $h$. We thus have that for any $h \in \mathbb{R}$

$$g_1'(\alpha)h = (D_\alpha g_1)(h) = D_{\Phi(\alpha)}K \circ D_\alpha \Phi(h) \tag{16.25}$$

where $D_\alpha \Phi$ (resp. $D_{\Phi(\alpha)}K$) are the differential of $\Phi$ (resp. $K$) in $\alpha$ (resp. $\Phi(\alpha)$). We therefore need to compute the differentials of $\Phi$ and $K$. Since $\Phi$ is linear, $D_\alpha \Phi = \Phi$ and thus

$$D_\alpha \Phi(h) = (h, h) \ . \tag{16.26}$$

Moreover, the differential of $K$ can be expressed with the partial derivatives of K, i.e., $D_{(\bar{x},\bar{y})}K(h_1, h_2) = (\nabla K) \cdot (h_1, h_2)$ where $\nabla$ is the vector differential operator $\nabla = \left( \frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n} \right) = (\partial_1, \ldots, \partial_n)$ and $(h_1, h_2) \in \mathbb{R}^2$. Hence,

$$D_{(\bar{x},\bar{y})}K(h_1, h_2) = \partial_1 K(\bar{x}, \bar{y})\, h_1 + \partial_2 K(\bar{x}, \bar{y})\, h_2.$$

We thus need to compute the partial derivatives of K. From the fundamental theorem of calculus, $\partial_1 K(\bar{x}, \bar{y}) = -\gamma(\bar{x}, \bar{y})$. Besides, $\partial_2 K(\bar{x}, \bar{y}) = \int_{\bar{x}}^\beta \partial_2 \gamma(x, \bar{y})dx$ and therefore

$$D_{(\bar{x},\bar{y})}K(h_1, h_2) = -\gamma(\bar{x}, \bar{y})h_1 + \left( \int_{\bar{x}}^\beta \partial_2 \gamma(x, \bar{y})dx \right) h_2.$$

Applying again the fundamental theorem of calculus to compute the second partial derivative of $\gamma$, we find that

$$\partial_2 \gamma(x, \bar{y}) = -f'(\bar{y})w(x, f(\bar{y}))$$

and thus

$$D_{(\bar{x},\bar{y})}K(h_1, h_2) = \left( -\int_{f(\bar{y})}^{f(x_0)} w(\bar{x}, y)dy \right) h_1 + \left( \int_{\bar{x}}^\beta -f'(\bar{y})w(x, f(\bar{y}))dx \right) h_2. \tag{16.27}$$

Combining Eq. 16.27 and Eq. 16.26 in Eq. 16.25 we obtain

$$\partial_1 g(\alpha, \beta) = g_1'(\alpha) = -f'(\alpha) \int_\alpha^\beta w(x, f(\alpha))dx - \int_{f(\alpha)}^{f(x_0)} w(\alpha, y)dy$$

which gives Eq. 16.23.

To compute the second partial derivative of $g$, we need to compute, for any $\alpha$, the derivative of the function $g_2 : \beta \to g(\alpha, \beta)$. The function $g_2$ can be rewritten as $g_2 : \beta \to \int_\alpha^\beta \theta(x)dx$ where $\theta(x) = \int_{f(\alpha)}^{f(x_0)} w(x, y)dy$. Therefore, from the fundamental theorem of calculus we have that $\partial_2 g(\alpha, \beta) = g_2'(\beta) = \theta(\beta)$ and thus

$$\partial_2 g(\alpha, \beta) = \int_{f(\alpha)}^{f(x_0)} w(\beta, y)dy \ .$$

□

We are now ready to prove Theorem 16.3.4 *Proof.* From the first order necessary optimality conditions, we know that if $(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu)$ maximizes Eq. 16.3, then either $\overline{x}_i^\mu$ belongs to $]x_{\min}, x_{\max}[$ and the i-th partial derivative of $I_{H,w}(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu)$ equals zero in $\overline{x}_i^\mu$, or $\overline{x}_i^\mu$ belongs to the boundary of $[x_{\min}, x_{\max}]$, i.e., $\overline{x}_i^\mu = x_{\min}$ or $\overline{x}_i^\mu = x_{\max}$. Therefore, we need to compute the partial derivatives of $I_{H,w}$. From Eq. 16.21, we have that $\partial_1 I_{H,w}(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu) = \partial_1 g(\overline{x}_1^\mu, \overline{x}_2^\mu)$ and from Lemma 16.7.1 we therefore obtain that

$$\partial_1 I_{H,w}(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu) = -f'(\overline{x}_1^\mu) \int_{\overline{x}_1^\mu}^{\overline{x}_2^\mu} w(x, f(\overline{x}_1^\mu))dx - \int_{f(\overline{x}_1^\mu)}^{f(\overline{x}_0^\mu)} w(\overline{x}_1^\mu, y)dy$$

and thus if $\overline{x}_1^\mu \neq x_{\min}$ and $\overline{x}_1^\mu \neq x_{\max}$. By setting the previous equation to zero, we obtain

$$-f'(\overline{x}_1^\mu) \int_{\overline{x}_1^\mu}^{\overline{x}_2^\mu} w(x, f(\overline{x}_1^\mu)dx = \int_{f(\overline{x}_1^\mu)}^{f(\overline{x}_0^\mu)} w(\overline{x}_1^\mu, y)dy \ .$$

For $2 \leq i \leq \mu$, $\partial_i I_{H,w}(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu) = \partial_2 g(\overline{x}_{i-1}^\mu, \overline{x}_i^\mu) + \partial_1 g(\overline{x}_i^\mu, \overline{x}_{i+1}^\mu)$. Using Lemma 16.7.1 we obtain

$$\partial_i I_{H,w}(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu) = \int_{f(\overline{x}_{i-1}^\mu)}^{f(\overline{x}_0^\mu)} w(\overline{x}_i^\mu, y)dy - f'(\overline{x}_i^\mu) \int_{\overline{x}_i^\mu}^{\overline{x}_{i+1}^\mu} w(x, f(\overline{x}_i^\mu))dx - \int_{f(\overline{x}_i^\mu)}^{f(\overline{x}_0^\mu)} w(\overline{x}_i^\mu, y)dy \ .$$

Gathering the first and last term of the right-hand side, we obtain

$$\partial_i I_{H,w}(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu) = \int_{f(\overline{x}_{i-1}^\mu)}^{f(\overline{x}_i^\mu)} w(\overline{x}_i^\mu, y)dy - f'(\overline{x}_i^\mu) \int_{\overline{x}_i^\mu}^{\overline{x}_{i+1}^\mu} w(x, f(\overline{x}_i^\mu))dx \tag{16.28}$$

and thus if $\overline{x}_{i+1}^\mu \neq x_{\min}$ and $\overline{x}_{i+1}^\mu \neq x_{\max}$, by setting the previous equation to zero, we obtain

$$\int_{f(\overline{x}_{i-1}^\mu)}^{f(\overline{x}_i^\mu)} w(\overline{x}_i^\mu, y)dy = f'(\overline{x}_i^\mu) \int_{\overline{x}_i^\mu}^{\overline{x}_{i+1}^\mu} w(x, f(\overline{x}_i^\mu))dx \ .$$

□

### 16.7.2 Proof of Lemma 16.3.9 stated on page 207

*Proof.* Let us first note that the Cauchy-Schwarz inequality implies that

$$\int_0^{x_{\max}} \frac{|f'(x)w\big(x, f(x)\big)|}{|\delta(x)|}dx \leq \sqrt{\int_0^{x_{\max}} \Big(f'(x)w(x, f(x))\Big)^2 dx \int_0^{x_{\max}} (1/\delta(x))^2 dx} \tag{16.29}$$

and since $x \to f'(x)w(x, f(x)) \in L^2(0, x_{\max})$ and $\frac{1}{\delta} \in L^2(0, x_{\max})$, the right-hand side of Eq. 16.29 is finite and Eq. 16.10 is well-defined. The proof is divided into two steps. First, we rewrite $E_\mu$ and, in a second step, the limit result is derived by using this new characterization of $E_\mu$.

**Step 1.** In a first step we are going to prove that $E_\mu$ defined in Eq. 16.9 satisfies

$$E_\mu = \mu \sum_{i=0}^{\mu} \Big( -\tfrac{1}{2} f'(\overline{x}_i^\mu) w(\overline{x}_i^\mu, f(\overline{x}_i^\mu))(\overline{x}_{i+1}^\mu - \overline{x}_i^\mu)^2 + O\Big((\overline{x}_{i+1}^\mu - \overline{x}_i^\mu)^3\Big)\Big) \ . \tag{16.30}$$

To this end, we elongate the front to the right such that $f$ equals $f(x_{\max}) = 0$ for $x \in [x_{\max}, \overline{x}_{\mu+1}^\mu]$. Like that, we can decompose $\int_0^{x_{\max}} \int_0^{f(x)} w(x,y)dydx$ into $\sum_{i=0}^{\mu} \int_{\overline{x}_i^\mu}^{\overline{x}_{i+1}^\mu} \int_0^{f(x)} w(x,y)\,dy\,dx$, while using the fact that $\int_{x_{\max}}^{\overline{x}_{\mu+1}^\mu} \int_0^{f(x)} w(x,y)dydx = 0$. Using the right-hand side of the previous equation in Eq. 16.9, we find that

$$E_\mu = \mu \left[ \sum_{i=0}^{\mu} \int_{\overline{x}_i^\mu}^{\overline{x}_{i+1}^\mu} \left( \int_0^{f(\overline{x}_i^\mu)} w(x,y)\,dy \right) dx - \sum_{i=0}^{\mu} \int_{\overline{x}_i^\mu}^{\overline{x}_{i+1}^\mu} \left( \int_0^{f(x)} w(x,y)\,dy \right) dx \right] = \mu \sum_{i=0}^{\mu} \int_{\overline{x}_i^\mu}^{\overline{x}_{i+1}^\mu} \int_{f(x)}^{f(\overline{x}_i^\mu)} w(x,y)\,dy\,dx \ . \tag{16.31}$$

At the first order, we have that

$$\int_{f(x)}^{f(\overline{x}_i^\mu)} w(x,y)dy = w(\overline{x}_i^\mu, f(\overline{x}_i^\mu))(f(\overline{x}_i^\mu) - f(x)) + O((x - \overline{x}_i^\mu)) \ . \tag{16.32}$$

Since $f$ is differentiable, we can use a Taylor approximation of $f$ in each interval $[\overline{x}_i^\mu, \overline{x}_{i+1}^\mu]$ and write $f(x) = f(\overline{x}_i^\mu) + f'(\overline{x}_i^\mu)(x - \overline{x}_i^\mu) + O((x - \overline{x}_i^\mu)^2)$, which thus implies that $f(\overline{x}_i^\mu) - f(x) = -f'(\overline{x}_i^\mu)(x - \overline{x}_i^\mu) + O((x - \overline{x}_i^\mu)^2)$ and thus the left-hand side of Eq. 16.32 becomes $-w(\overline{x}_i^\mu, f(\overline{x}_i^\mu)) f'(\overline{x}_i^\mu)(x - \overline{x}_i^\mu) + O((x - \overline{x}_i^\mu)^2)$. By integrating the previous equation between $\overline{x}_i^\mu$ and $\overline{x}_{i+1}^\mu$ we obtain

$$\int_{\overline{x}_i^\mu}^{\overline{x}_{i+1}^\mu} \int_{f(x)}^{f(\overline{x}_i^\mu)} w(x,y)\,dy\,dx = -\frac{1}{2} w(\overline{x}_i^\mu, f(\overline{x}_i^\mu)) f'(\overline{x}_i^\mu)(\overline{x}_{i+1}^\mu - \overline{x}_i^\mu)^2 + O((\overline{x}_{i+1}^\mu - \overline{x}_i^\mu)^3) \ .$$

Summing up for $i = 0$ to $i = \mu$, multiplying by $\mu$ and using Eq. 16.31, we obtain Eq. 16.30, which concludes Step 1.

**Step 2.** We now decompose $\frac{1}{2} \int_0^{x_{\max}} \frac{f'(x) w(x, f(x))}{\delta(x)} dx$ into

$$\frac{1}{2} \sum_{i=0}^{\mu-1} \int_{\overline{x}_i^\mu}^{\overline{x}_{i+1}^\mu} \frac{f'(x) w(x, f(x))}{\delta(x)} dx + \frac{1}{2} \int_{\overline{x}_\mu^\mu}^{x_{\max}} \frac{f'(x) w(x, f(x))}{\delta(x)} dx \ .$$

For the sake of convenience in the notations, for the remainder of the proof, we redefine $\overline{x}_{\mu+1}^\mu$ as $x_{\max}$ such that the previous equation becomes

$$\frac{1}{2} \int_0^{x_{\max}} \frac{f'(x) w(x, f(x))}{\delta(x)} dx = \frac{1}{2} \sum_{i=0}^{\mu} \int_{\overline{x}_i^\mu}^{\overline{x}_{i+1}^\mu} \frac{f'(x) w(x, f(x))}{\delta(x)} dx \tag{16.33}$$

For $\mu$ to $\infty$, the assumption $\mu \sup((\sup_{0 \le i \le \mu-1} |\overline{x}_{i+1}^\mu - \overline{x}_i^\mu|), |x_{\max} - \overline{x}_\mu^\mu|) \to c$ implies that the distance between two consecutive points $|\overline{x}_{i+1}^\mu - \overline{x}_i^\mu|$ as well as $|\overline{x}_\mu^\mu - x_{\max}|$ converges to zero. Let $x \in [0, x_{\max}]$ and let us define for a given $\mu$, $\varphi(\mu)$ as the index of the points such that $\overline{x}_{\varphi(\mu)}^\mu$ and $\overline{x}_{\varphi(\mu)+1}^\mu$ surround $x$, i.e., $\overline{x}_{\varphi(\mu)}^\mu \le x < \overline{x}_{\varphi(\mu)+1}^\mu$. Since we assume that $\delta$ is continuous, a first order approximation of $\delta(x)$ is $\delta(\overline{x}_\mu^\mu)$, i.e., $\delta(x) = \delta(\overline{x}_{\varphi(\mu)}^\mu) + O(\overline{x}_{\varphi(\mu)+1}^\mu - \overline{x}_{\varphi(\mu)}^\mu)$ and therefore by integrating between $\overline{x}_{\varphi(\mu)}^\mu$ and $\overline{x}_{\varphi(\mu)+1}^\mu$ we obtain

$$\int_{\overline{x}_{\varphi(\mu)}^\mu}^{\overline{x}_{\varphi(\mu)+1}^\mu} \delta(x)dx = \delta(\overline{x}_{\varphi(\mu)}^\mu)(\overline{x}_{\varphi(\mu)+1}^\mu - \overline{x}_{\varphi(\mu)}^\mu) + O(\overline{x}_{\varphi(\mu)+1}^\mu - \overline{x}_{\varphi(\mu)}^\mu)^2) \ . \tag{16.34}$$

Moreover by definition of the density $\delta$, $\int_{\overline{x}_{\varphi(\mu)}^\mu}^{\overline{x}_{\varphi(\mu)+1}^\mu} \delta(x)dx$ approximates the number of points contained in the interval $[\overline{x}_{\varphi(\mu)}^\mu, \overline{x}_{\varphi(\mu)+1}^\mu[$ (i.e., one) normalized by $\mu$:

$$\mu \int_{\overline{x}_{\varphi(\mu)}^\mu}^{\overline{x}_{\varphi(\mu)+1}^\mu} \delta(x)dx = 1 + O((\overline{x}_{\varphi(\mu)+1}^\mu - \overline{x}_{\varphi(\mu)}^\mu)) \ . \tag{16.35}$$

Using Eq. 16.34 and Eq. 16.35, we thus have $1/\delta(\overline{x}^{\mu}_{\varphi(\mu)}) = \mu(\overline{x}^{\mu}_{\varphi(\mu)+1} - \overline{x}^{\mu}_{\varphi(\mu)}) + O(\mu(\overline{x}^{\mu}_{\varphi(\mu)+1} - \overline{x}^{\mu}_{\varphi(\mu)})^2)$. Therefore for every $i$ we have that

$$\frac{1}{\delta(\overline{x}^{\mu}_{i})} = \mu(\overline{x}^{\mu}_{i+1} - \overline{x}^{\mu}_{i}) + O(\mu(\overline{x}^{\mu}_{i+1} - \overline{x}^{\mu}_{i})^2) \ . \tag{16.36}$$

Since $x \to f'(x)w(x, f(x))/\delta(x)$ is continuous, we also obtain

$$\int_{\overline{x}^{\mu}_{i}}^{\overline{x}^{\mu}_{i+1}} \frac{f'(x)w(x, f(x))}{\delta(x)} dx = \frac{f'(\overline{x}^{\mu}_{i})w(\overline{x}^{\mu}_{i}, f(\overline{x}^{\mu}_{i}))}{\delta(\overline{x}^{\mu}_{i})}(\overline{x}^{\mu}_{i+1} - \overline{x}^{\mu}_{i}) + O((\overline{x}^{\mu}_{i+1} - \overline{x}^{\mu}_{i})^2) \ .$$

Injecting Eq. 16.36 in the previous equation, we obtain

$$\int_{\overline{x}^{\mu}_{i}}^{\overline{x}^{\mu}_{i+1}} \frac{f'(x)w(x, f(x))}{\delta(x)} dx = \mu f'(\overline{x}^{\mu}_{i})w(\overline{x}^{\mu}_{i}, f(\overline{x}^{\mu}_{i}))(\overline{x}^{\mu}_{i+1} - \overline{x}^{\mu}_{i})^2 + O(\mu(\overline{x}^{\mu}_{i+1} - \overline{x}^{\mu}_{i})^3) \ .$$

Multiplying by 1/2 and summing up for $i$ from 0 to $\mu$ and using Eq. 16.30 and Eq. 16.33, we obtain

$$\frac{1}{2}\int_{0}^{x_{\max}} \frac{f'(x)w(x, f(x))}{\delta(x)} = -E_{\mu} + \sum_{i=0}^{\mu} O(\mu(\overline{x}^{\mu}_{i+1} - \overline{x}^{\mu}_{i})^3) \ . \tag{16.37}$$

Let us define $\Delta_{\mu}$ as $\sup((\sup_{0 \le i \le \mu-1} |\overline{x}^{\mu}_{i+1} - \overline{x}^{\mu}_{i}|), |x_{\max} - \overline{x}^{\mu}_{\mu}|)$. By assumption, we know that $\mu\Delta_{\mu}$ converges to a positive constant $c$. The last term of Eq. 16.37 satisfies

$$\left| \sum_{i=0}^{\mu} O(\mu(\overline{x}^{\mu}_{i+1} - \overline{x}^{\mu}_{i})^3) \right| \le K\mu^2(\Delta_{\mu})^3$$

where $K > 0$. Since $\mu\Delta_{\mu}$ converges to $c$, $(\mu\Delta_{\mu})^2$ converges to $c^2$. With $\Delta_{\mu}$ converging to 0, we therefore have that $\mu^2\Delta^3_{\mu}$ converges to 0. Taking the limit in Eq. 16.37, we therefore obtain

$$-\frac{1}{2}\int_{0}^{x_{\max}} \frac{f'(x)w(x, f(x))}{\delta(x)} dx = \lim_{\mu \to \infty} E_{\mu} \ .$$

□

### 16.7.3 Proof of Theorem 16.4.1 stated on page 210

Before to state and prove Theorem 16.4.1, we need to establish a technical lemma.

**16.7.2 Lemma**
*Let us assume that $f$ is continuous on $[x_{\min}, x_{\max}]$ and differentiable on $]x_{\min}, x_{\max}[$. Let $x_2 \in ]x_{\min}, r_1]$ and let us define the function $\Theta : [0, x_{\max} - x_{\min}] \to \mathbb{R}$ as*

$$\Theta(\varepsilon) = \int_{x_{\min}+\varepsilon}^{x_2} \left( \int_{f(x_{\min}+\varepsilon)}^{f(x_{\min})} w(x, y)dy \right) dx$$

*and $\Gamma : [0, x_2 - x_{\min}] \to \mathbb{R}$ as*

$$\Gamma(\varepsilon) = \int_{x_{\min}}^{x_{\min}+\varepsilon} \left( \int_{f(x_{\min})}^{r_2} w(x, y)dy \right) dx \ .$$

*If $w$ is continuous, positive and $\lim_{x \to x_{\min}} f'(x) = -\infty$ then for any $r_2 > f(x_{\min})$*

$$\lim_{\varepsilon \to 0} \frac{\Theta(\varepsilon)}{\Gamma(\varepsilon)} = +\infty \ .$$

*Proof.* The limits of $\Theta$ and $\Gamma$ for $\varepsilon$ converging to 0 equal 0. We will therefore apply the l'Hôpital rule to compute $\lim_{\varepsilon\to 0}\frac{\Theta(\varepsilon)}{\Gamma(\varepsilon)}$. First of all, note that since $f$ is differentiable on $]x_{\min}, x_{\max}[$, $\Theta$ and $\Gamma$ are differentiable on $]0, x_{\max}-x_{\min}]$. Moreover, we see that $\Theta(\varepsilon) = g(x_{\min}+\varepsilon, x_2)$ where $g$ is defined in Eq. 16.22 except for the change from $f(\overline{x}_0^\mu)$ to $f(x_{\min})$. The proof of Lemma 16.7.1, however, does not change if we exchange the constant $f(\overline{x}_0^\mu)$ to the constant $f(x_{\min})$ and we deduce that

$$\Theta'(\varepsilon) = -f'(x_{\min}+\varepsilon)\int_{x_{\min}+\varepsilon}^{x_2} w(x, f(x_{\min}+\varepsilon))dx - \int_{f(x_{\min}+\varepsilon)}^{f(x_{\min})} w(x_{\min}+\varepsilon, y)dy \ .$$

From the fundamental theorem of calculus, we also have that

$$\Gamma'(\varepsilon) = \int_{f(x_{\min})}^{r_2} w(x_{\min}+\varepsilon, y)dy \ .$$

From the l'Hôpital rule, we deduce that

$$\lim_{\varepsilon\to 0}\frac{\Theta(\varepsilon)}{\Gamma(\varepsilon)} = \lim_{\varepsilon\to 0}\frac{\Theta'(\varepsilon)}{\Gamma'(\varepsilon)} \ . \tag{16.38}$$

By continuity of $w$, we deduce that

$$\lim_{\varepsilon\to 0}\Gamma'(\varepsilon) = \lim_{\varepsilon\to 0}\int_{f(x_{\min})}^{r_2} w(x_{\min}+\varepsilon, y)dy = \int_{f(x_{\min})}^{r_2} w(x_{\min}, y)dy$$

and by continuity of $f$ and $w$, we deduce that

$$\lim_{\varepsilon\to 0}\int_{x_{\min}+\varepsilon}^{x_2} w(x, f(x_{\min}+\varepsilon))dx = \int_{x_{\min}}^{x_2} w(x, f(x_{\min}))dx \qquad \text{and} \qquad \lim_{\varepsilon\to 0}\int_{f(x_{\min}+\varepsilon)}^{f(x_{\min})} w(x_{\min}+\varepsilon, y)dy = 0 \ .$$

Therefore $\lim_{\varepsilon\to 0}\Theta'(\varepsilon) = \lim_{\varepsilon\to 0} -f'(x_{\min}+\varepsilon)\cdot\int_{x_{\min}}^{x_2} w(x, f(x_{\min}))dx = +\infty$ because $x_2$ is fixed, i.e., independent of $\varepsilon$, and therefore, the integral is constant. By Eq. 16.38 we obtain the result. □

Now, we are ready to prove Theorem 16.4.1. *Proof.* We first prove the result for the left extreme. We denote $\overline{x}_1^\mu$ and $\overline{x}_2^\mu$ the two leftmost points of an optimal $\mu$-distribution for $I_{H,w}$ if $\mu \geq 2$. In case of $\mu = 1$, let $\overline{x}_1^\mu$ be the optimal position of the (single) point. In this case, the contribution of $\overline{x}_1^\mu$ in the first dimension extends to the reference point, which we represent by setting $\overline{x}_2^\mu = r_1$ such that from now on, we can assume $\mu \geq 2$. We assume that $\lim_{x\to x_{\min}} f'(x) = -\infty$ and that $\overline{x}_1^\mu = x_{\min}$ in order to get a contradiction. Let $I_{H,w}(x_{\min})$ be the hypervolume solely dominated by the point $x_{\min}$. If we shift $\overline{x}_1^\mu$ to the right by $\varepsilon > 0$ (see Figure 16.11), then the new hypervolume contribution $I_{H,w}(x_{\min}+\varepsilon)$ satisfies

$$I_{H,w}(x_{\min}+\varepsilon) = I_{H,w}(x_{\min}) + \int_{x_{\min}+\varepsilon}^{\overline{x}_2^\mu}\int_{f(x_{\min}+\varepsilon)}^{f(x_{\min})} w(x,y)dydx - \int_{x_{\min}}^{x_{\min}+\varepsilon}\int_{f(x_{\min})}^{r_2} w(x,y)dydx \ .$$

Identifying $x_2$ with $\overline{x}_2^\mu$ in the definition of $\Theta$ in Lemma 16.7.2, the previous equation can be rewritten as

$$I_{H,w}(x_{\min}+\varepsilon) = I_{H,w}(x_{\min}) + \Theta(\varepsilon) - \Gamma(\varepsilon) \ .$$

From Lemma 16.7.2, for any $r_2 > f(x_{\min})$, there exists an $\varepsilon > 0$ such that $\frac{\Theta(\varepsilon)}{\Gamma(\varepsilon)} > 1$ and thus $\Theta(\varepsilon) - \Gamma(\varepsilon) > 0$. Thus, for any $r_2 > f(x_{\min})$, there exists an $\varepsilon$ such that $I_{H,w}(x_{\min}+\varepsilon) > I_{H,w}(x_{\min})$ and thus $I_{H,w}(x_{\min})$ is not maximal which contradicts the fact that $\overline{x}_1^\mu = x_{\min}$. In a similar way, we can prove the result for the right extreme. □

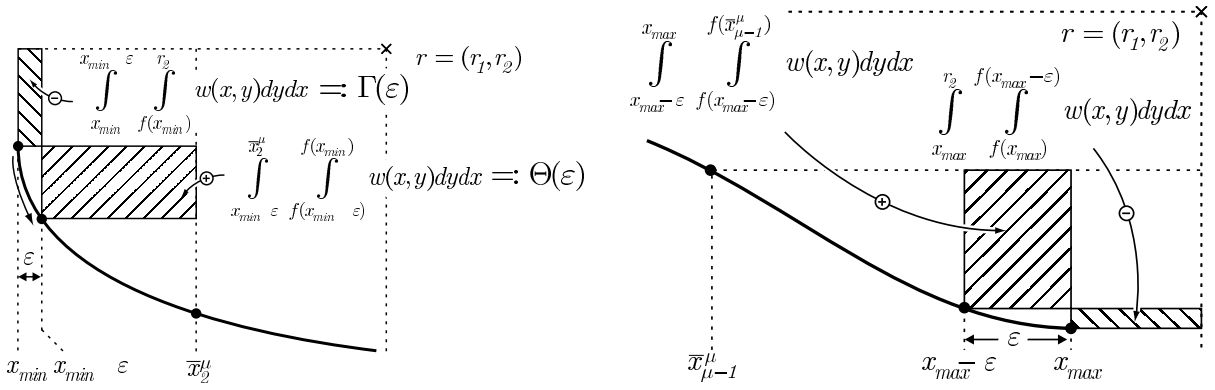


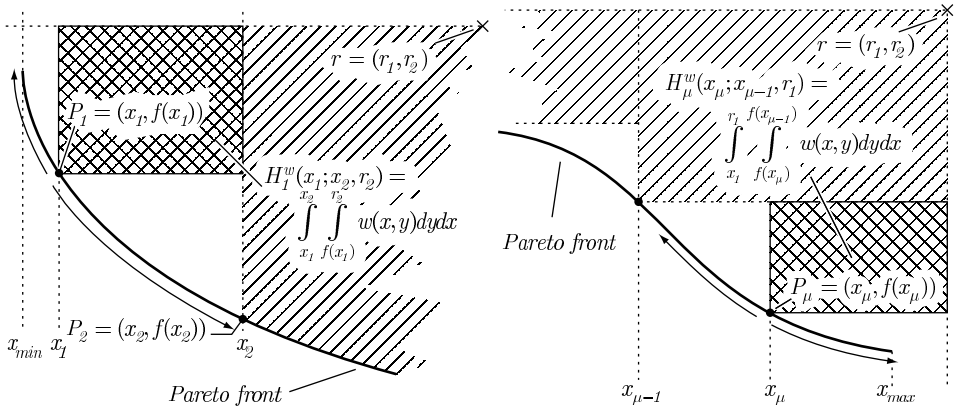

Figure 16.11: If the function $f(x)$ describing the Pareto front has an infinite derivative at its left extreme, the leftmost Pareto-optimal point at $x_{\min}$ will never coincide with the leftmost point $\overline{x}_1^\mu$ of an optimal $\mu$-distribution for $I_{H,w}$ (left); similarly, if the derivative is zero at the right extreme, the rightmost Pareto-optimal point at $x_{\max}$ will never coincide with the rightmost point $\overline{x}_\mu^\mu$ (right). The reason is in both cases that for any finite $r_1$, and $r_2$ respectively, there exists an $\varepsilon > 0$, such that the dominated space gained ($\oplus$) when moving $\overline{x}_1^\mu$ from $x_{\min}$ to $x_{\min} + \varepsilon$, and $\overline{x}_\mu^\mu$ from $x_{\max}$ to $x_{\max} - \varepsilon$ respectively, is larger than the space no longer dominated ($\ominus$).

### 16.7.4 Proof of Theorem 16.4.3 stated on page 211

The proof of the theorem requires to establish a technical proposition. We have assumed that the reference point is dominated by the Pareto front, i.e., at least $r_1 > x_{\max}$ and $r_2 > f(x_{\min})$. Let us consider a set of points on the front and the hypervolume contribution of the leftmost point $P_1 = (x_1, f(x_1))$ (see Figure 16.12). This hypervolume contribution is a function of $x_1$ itself, $x_2$, the x-coordinate of the second leftmost point, and $r_2$, the second coordinate of the reference point. For a fixed $x_2, r_2$, the hypervolume contribution of the leftmost point with coordinate $x_1 \in [x_{\min}, x_2[$ is denoted $H_1^w(x_1; x_2, r_2)$ and reads

$$H_1^w(x_1; x_2, r_2) = \int_{x_1}^{x_2}\int_{f(x_1)}^{r_2} w(x,y)dydx \ . \tag{16.39}$$

The following proposition establishes a key property of the function $H_1^w$.




(a) left extreme (b) right extreme

Figure 16.12: Shows the notation and formula to compute the hypervolume contributions of the leftmost and rightmost point $P_1$ and $P_\mu$ respectively.

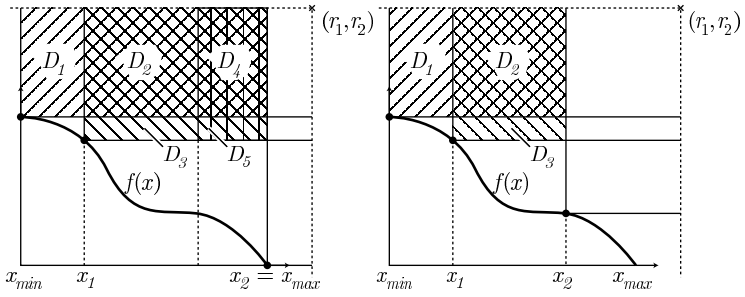



Figure 16.13: If the hypervolume indicator is larger for the choice of $x_1 = x_{\min}$ than when choosing $x_1 > x_{\min}$ if $x_2 = x_{\max}$ (left-hand side), it is also larger for $x_1 = x_{\min}$ for any $x_2 > x_1$ (right-hand side).

**16.7.3 Proposition**
*If $x_1 \to H_1^w(x_1; x_{\max}, r_2)$ is maximal for $x_1 = x_{\min}$, then for any $x_2 \in ]x_1, x_{\max}]$ the contribution $H_1^w(x_1; x_2, r_2)$ is maximal for $x_1 = x_{\min}$ too.*

*Proof.* [Proof] Assume that $H_1^w(x_1; x_{\max}, r_2)$ is maximal for $x_1 = x_{\min}$, i.e., $H_1^w(x_{\min}; x_{\max}, r_2) \geq H_1^w(x_1; x_{\max}, r_2)$, for all $x_1 \in ]x_{\min}, x_{\max}]$. Let $\{D_1, \ldots, D_5\}$ denote the weighted hypervolume indicator values of different non-overlapping rectangular areas shown in Fig. 16.13. Then for all $x_1$ in $]x_{\min}, x_{\max}]$, $H_1^w(x_{\min}; x_{\max}, r_2) \geq H_1^w(x_1; x_{\max}, r_2)$ can be rewritten using $D_1, \ldots, D_5$ as

$$D_1 + D_2 + D_4 \geq D_2 + D_3 + D_4 + D_5$$

which in turn implies that $D_1 + D_2 \geq D_2 + D_3 + D_5$. Since $D_5 \geq 0$ we have that $D_1 + D_2 \geq D_2 + D_3$, which corresponds to $H_1^w(x_{\min}; x_2, r_2) \geq H_1^w(x_1; x_2, r_2)$. Hence, $H_1^w(x_1; x_2, r_2)$ is also maximal for $x_1 = x_{\min}$ for any choice $x_2 \in ]x_1, x_{\max}]$. □

We are now ready to prove Theorem 16.4.3. *Proof.* [Proof of Theorem 16.4.3] Let $x_1$ and $x_2$ denote the $x$-coordinates of the two leftmost points $P_1 = (x_1, f(x_1))$ and $P_2 = (x_2, f(x_2))$. Then the hypervolume contribution of $P_1$ is given by Eq. 16.39. To prove that $P_1$ is the extremal point $(x_{\min}, f(x_{\min}))$, we need to prove that $x_1 \in [x_{\min}, x_2] \mapsto H_1^w(x_1; x_2, r_2)$ is maximal for $x_1 = x_{\min}$. By using Proposition 16.7.3, we know that if we prove that $x_1 \to H_1^w(x_1; x_{\max}, r_2)$ is maximal for $x_1 = x_{\min}$ then we will also have that $H_1^w : x_1 \in [x_{\min}, x_2] \mapsto H_1^w(x_1; x_2, r_2)$ is maximal for $x_1 = x_{\min}$. Therefore we will now prove that $x_1 \to H_1^w(x_1; x_{\max}, r_2)$ is maximal for $x_1 = x_{\min}$. To do so, we will show that $\frac{dH_1^w(x_1; x_{\max}, r_2)}{dx_1} \neq 0$ for all $x_{\min} < x_1 \leq x_{\max}$. According to Lemma 16.7.1, the derivative of the hypervolume contribution of $P_1$ is

$$\frac{dH_1^w(x_1; x_{\max}, r_2)}{dx_1} = -f'(x_1) \int_{x_1}^{x_{\max}} w(x, f(x_1))dx - \int_{f(x_1)}^{r_2} w(x_1, y)dy$$

Hence, by choosing $r_2 > \mathcal{K}_2$ according to Theorem 16.4.3, $\frac{dH_1^w(x_1; x_{\max}, r_2)}{dx_1} \neq 0$. □

## 16.7.5 Proof of Corollary 16.4.5 stated on page 212

*Proof.* We replace $w(x, y)$ in Eq. 16.13 of Theorem 16.4.3 by 1 and obtain that if there exists a $\mathcal{K}_2 \in \mathbb{R}$ such that

$$\forall x_1 \in ]x_{\min}, x_{\max}] : \mathcal{K}_2 - f(x_1) > -f'(x_1)(x_{\max} - x_1), \tag{16.40}$$

then for any $r_2 \geq \mathcal{K}_2$, the leftmost extreme is included. The previous equation writes $\mathcal{K}_2 > f(x_1) - f'(x_1)(x_{\max} - x_1)$ for all $x_1 \in ]x_{\min}, x_{\max}]$. However $-f'(x_1)(x_{\max} - x_1) = f'(x_1)(x_1 - x_{\max})$. Therefore Eq. 16.40 writes as

$$\forall x_1 \in ]x_{\min}, x_{\max}] : \mathcal{K}_2 > f(x_1) + f'(x_1)(x_1 - x_{\max}) \ . \tag{16.41}$$

Since $\mathcal{K}_2$ has to be larger than the right-hand side of Eq. 16.41 for all $x_1$ in $]x_{\min}, x_{\max}]$, it has to be larger than the supremum of $f(x_1) + f'(x_1)(x_1 - x_{\max})$ for $x_1$ in $]x_{\min}, x_{\max}]$ and thus

$$\mathcal{K}_2 > \sup\{f(x_1) + f'(x_1)(x_1 - x_{\max}) : x \in ]x_{\min}, x_{\max}]\} \ . \tag{16.42}$$

Defining $\mathcal{R}_2$ as the infimum over $\mathcal{K}_2$ satisfying Eq. 16.42 results in Eq. 16.15 which concludes the proof. □

### 16.7.6 Proof of Theorem 16.4.7 stated on page 212

Before to present the proof, we consider the hypervolume contribution of the rightmost point:

$$H^w_\mu(x_\mu; x_{\mu-1}, r_1) = \int_{x_\mu}^{r_1} \int_{f(x_\mu)}^{f(x_{\mu-1})} w(x, y) dy dx \tag{16.43}$$

Similar to Proposition 16.7.3 we can establish the following proposition:

**16.7.4 Proposition**
*If $x_\mu \to H^w_1(x_\mu; x_{\min}, r_1)$ is maximal for $x_\mu = x_{\max}$, then for any $x_\mu \in [x_{\min}, x_{\mu-1}[$ the contribution $H^w_\mu(x_\mu; x_{\mu-1}, r_1)$ is maximal for $x_\mu = x_{\max}$ too.*

We are now ready to prove Theorem 16.4.7. *Proof.* [Proof of Theorem 16.4.7] Let $x_\mu$ and $x_{\mu-1}$ denote the $x$-coordinates of the two rightmost points $P_\mu = (x_\mu, f(x_\mu))$ and $P_{\mu-1} = (x_{\mu-1}, f(x_{\mu-1}))$. Then the hypervolume contribution of $P_\mu$ is given by Eq. 16.43. To prove that $P_\mu$ is the extremal point $(x_{\max}, f(x_{\max}))$, we need to prove that $x_\mu \in [x_{\mu-1}, x_{\max}] \mapsto H^w_\mu(x_\mu; x_{\mu-1}, r_1)$ is maximal for $x_\mu = x_{\max}$. By using Proposition 16.7.4, we know that if we prove that $x_\mu \to H^w_\mu(x_\mu; x_{\min}, r_1)$ is maximal for $x_\mu = x_{\max}$ then we will also have that $H^w_\mu : x_\mu \in [x_{\mu-1}, x_{\max}] \mapsto H^w_\mu(x_\mu; x_{\mu-1}, r_1)$ is maximal for $x_\mu = x_{\max}$. Therefore, we will now prove that $x_\mu \to H^w_\mu(x_\mu; x_{\min}, r_1)$ is maximal for $x_\mu = x_{\max}$. To do so, we will show that $\frac{dH^w_\mu(x_\mu; x_{\min}, r_1)}{dx_\mu} \neq 0$ for all $x_{\min} \leq x_\mu < x_{\max}$. According to Lemma 16.7.1, the derivative of the hypervolume contribution of $P_\mu$ is

$$\frac{dH^w_\mu(x_\mu; x_{\min}, r_1)}{dx_\mu} = -f'(x_\mu) \int_{x_\mu}^{r_1} w(x, f(x_\mu)) dx - \int_{f(x_\mu)}^{f(x_{\min})} w(x_\mu, y) dy \ .$$

Hence, by choosing $r_1 > \mathcal{K}_1$ according to Theorem 16.4.7, $\frac{dH^w_\mu(x_\mu; x_{\min}, r_1)}{dx_\mu} \neq 0$. □

### 16.7.7 Proof of Corollary 16.4.8 stated on page 213

*Proof.* We replace $w(x, y)$ in Eq. 16.16 of Theorem 16.4.7 by 1 and obtain that if there exists a $\mathcal{K}_1 \in \mathbb{R}$ such that $-f'(x_\mu)(\mathcal{K}_1 - x_\mu) > (f(x_{\min}) - f(x_\mu))$ holds for all $x_\mu \in [x_{\min}, x_{\max}[$, then for every $r_1 \geq \mathcal{K}_1$, the rightmost extreme is included in optimal $\mu$-distributions for $I_H$. The previous inequality writes

$$\forall x_\mu \in [x_{\min}, x_{\max}[ : \mathcal{K}_1 > (f(x_\mu) - f(x_{\min}))/f'(x_\mu) + x_\mu \ . \tag{16.44}$$

Since $\mathcal{K}_1$ has to be larger than the right-hand side of Eq. 16.44 for all $x_\mu$ in $[x_{\min}, x_{\max}[$, it has to be larger than the supremum of the right-hand side of Eq. 16.44 for $x_\mu$ in $[x_{\min}, x_{\max}[$ and thus

$$\mathcal{K}_1 > \sup\left\{x + \frac{f(x) - f(x_{\min})}{f'(x)} : x \in [x_{\min}, x_{\max}[\right\} \ . \tag{16.45}$$

Defining $\mathcal{R}_1$ as the infimum over $\mathcal{K}_1$ satisfying Eq. 16.45 results in Eq. 16.18 which concludes the proof. □

### 16.7.8 Proof of Lemma 16.4.9 stated on page 213

*Proof.* Let us denote the leftmost and the rightmost point of an optimal $\mu$-distribution for $I_{H,w}$ as $\overline{x}_1^\mu(R)$ and $\overline{x}_\mu^\mu(R)$ respectively when the hypervolume indicator is computed with respect to a reference point $R$. By assumption, $\overline{x}_1^\mu(R^1) = x_{\min}$ and $\overline{x}_\mu^\mu(R^1) = x_{\max}$. Assume, in order to get a contradiction, that $\overline{x}_1^\mu(R^2) > x_{\min}$ (i.e., the leftmost point of the optimal $\mu$-distribution for $I_{H,w}$ and $R^2$ is not the left extreme) and assume that $\overline{x}_\mu^\mu(R^2) = x_{\max}$ for the moment. Let us denote $\overline{I_{H,w}^\mu}(R^2)$ the hypervolume associated with an optimal $\mu$-distribution for $I_{H,w}$ computed with respect to the reference point $R^2$ (and $\overline{I_{H,w}^\mu}(R^1)$ accordingly for $R^1$). We decompose $\overline{I_{H,w}^\mu}(R^2)$ in the following manner (see Figure 16.14)

$$\overline{I_{H,w}^\mu}(R^2) = A_1 + A_2 + A_3 \tag{16.46}$$

where $A_1$ is the hypervolume (computed with respect to $w$) enclosed in between the optimal $\mu$-distribution associated with $R^2$ and the reference point $R^1$, $A_2$ is the hypervolume (computed with respect to $w$) enclosed in the rectangle whose diagonal extremities are $R^2$ and $(\overline{x}_1^\mu(R^2), r_2^1)$ and $A_3$ is the hypervolume (again with respect to $w$) enclosed in the rectangle with diagonal $[(r_1^1, f(x_{\max})), (r_1^2, r_2^1)]$. Consider now an optimal $\mu$-distribution for $I_{H,w}$ associated with the reference point $R^1$ and denote this optimal $\mu$-distribution $(\overline{x}_1^\mu(R^1), \ldots, \overline{x}_\mu^\mu(R^1))$. The weighted hypervolume enclosed by this set of points and $R^2$ equals $\overline{I_{H,w}^\mu}(R^1) + A_2 + A_2' + A_3$ where $A_2'$ is the hypervolume (computed with respect to $w$) enclosed in the rectangle whose diagonal is $[(x_{\min}, r_2^1), (\overline{x}_1^\mu(R^2), r_2^2)]$ (Fig. 16.14). By definition of $\overline{I_{H,w}^\mu}(R^2)$ we have that

$$\overline{I_{H,w}^\mu}(R^2) \geq \overline{I_{H,w}^\mu}(R^1) + A_2 + A_2' + A_3 \ . \tag{16.47}$$

However, since $\overline{I_{H,w}^\mu}(R^1)$ is the maximal hypervolume value possible for the reference point $R^1$ and a set of $\mu$ points, we have that $A_1 \leq \overline{I_{H,w}^\mu}(R^1)$ and thus with Eq. 16.47 that $\overline{I_{H,w}^\mu}(R^2) \geq A_1 + A_2 + A_2' + A_3$ . From Eq. 16.46, we deduce that

$$\overline{I_{H,w}^\mu}(R^2) \geq \overline{I_{H,w}^\mu}(R^2) + A_2' \ . \tag{16.48}$$

Since we have assumed that $\overline{x}_1^\mu(R^2) > x_{\min}$ and that $r_2^2 > r_2^1$, we have $A_2' > 0$. And thus, Eq. 16.48 implies that $\overline{I_{H,w}^\mu}(R^2) > \overline{I_{H,w}^\mu}(R^2)$, which contradicts our assumption. In a similar way, we show a contradiction if we assume that both $\overline{x}_1^\mu(R^2) > x_{\min}$ and $\overline{x}_\mu^\mu(R^2) < x_{\max}$, i.e., if both extremes are not contained in an optimal $\mu$-distribution for $I_{H,w}$ and the reference point $R^2$. Also the proof for the right extreme is similar. □

### 16.7.9 Proof of Theorem 16.4.10 stated on page 213

*Proof.* Let us fix $\varepsilon_2 \in \mathbb{R}_{>0}$ and let $R = (R_1, R_2) = (r_1, \mathcal{R}_2^{\text{Nadir}} + \varepsilon_2)$ for $r_1$ arbitrarily chosen with $r_1 \geq \mathcal{R}_1^{\text{Nadir}}$. The optimal $\mu$-distributions for $I_{H,w}$ and the reference point $R$ obviously depend on $\mu$. Let $\overline{x}_2^\mu(R)$ denote the second point of an optimal $\mu$-distribution for $I_{H,w}$ when $R$ is chosen as reference point. We know that for $\mu$ to infinity, $\overline{x}_2^\mu(R)$ converges to $x_{\min}$. Also, because $f'$ is continuous on $[x_{\min}, x_{\max}]$, the extreme value theorem implies that there exists $\theta > 0$ such that $|f'(x)| \leq \theta$ for all $x \in [x_{\min}, x_{\max}]$. Since $f'$ is negative we therefore have

$$\forall x \in [x_{\min}, x_{\max}] : -f'(x) \leq \theta \ . \tag{16.49}$$

In order to prove that the leftmost point of an optimal $\mu$-distribution is $x_{\min}$, it is enough to show that the first partial derivative of $I_{H,w}$ is non-zero on $]x_{\min}, \overline{x}_2^\mu(R)]$. According to Eq. 16.3 and Lemma 16.7.1, the first partial derivative of $I_{H,w}((\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu))$ equals (we omit the dependence in $R$ for the following

equations)

$$\begin{aligned}\partial_1 I_{H,w} &= -f'(\overline{x}_1^\mu)\int_{\overline{x}_1^\mu}^{\overline{x}_2^\mu} w\Big(x, f(\overline{x}_1^\mu)\Big)dx - \int_{f(\overline{x}_1^\mu)}^{R_2} w(\overline{x}_1^\mu, y)dy \\ &= \Big(-f'(\overline{x}_1^\mu)\Big)\int_{x_{\min}}^{\overline{x}_2^\mu} w\Big(x, f(\overline{x}_1^\mu)\Big)dx - \Big(-f'(\overline{x}_1^\mu)\Big)\int_{x_{\min}}^{\overline{x}_1^\mu} w\Big(x, f(\overline{x}_1^\mu)\Big)dx - \int_{f(\overline{x}_1^\mu)}^{\mathcal{R}_2^{\text{Nadir}}} w(\overline{x}_1^\mu, y)dy - \int_{\mathcal{R}_2^{\text{Nadir}}}^{\mathcal{R}_2^{\text{Nadir}}+\varepsilon_2} w(\overline{x}_1^\mu, y)dy \ .\end{aligned} \tag{16.50}$$

Since the second and third summand are non-positive due to $w$ being strictly positive we have

$$\leq \Big(-f'(\overline{x}_1^\mu)\Big)\int_{x_{\min}}^{\overline{x}_2^\mu} w\Big(x, f(\overline{x}_1^\mu)\Big)dx - \int_{\mathcal{R}_2^{\text{Nadir}}}^{\mathcal{R}_2^{\text{Nadir}}+\varepsilon_2} w(\overline{x}_1^\mu, y)dy \tag{16.51}$$

and because $w \leq W$ and with Eq. 16.49, Eq. 16.51 can be upper bounded by

$$\leq \theta W(\overline{x}_2^\mu - x_{\min}) - \int_{\mathcal{R}_2^{\text{Nadir}}}^{\mathcal{R}_2^{\text{Nadir}}+\varepsilon_2} w(\overline{x}_1^\mu, y)dy \ . \tag{16.52}$$

Since $\overline{x}_2^\mu$ converges to $x_{\min}$ for $\mu$ to infinity, and $-\int_{\mathcal{R}_2^{\text{Nadir}}}^{\mathcal{R}_2^{\text{Nadir}}+\varepsilon_2} w(\overline{x}_1^\mu, y)dy < 0$ we deduce that there exists $\mu_1$ such that for all $\mu$ larger than $\mu_1$, Eq. 16.52 is strictly negative and thus for all $\mu$ larger than $\mu_1$, the first partial derivative of $I_{H,w}$ is non zero, i.e., $\overline{x}_1^\mu = x_{\min}$. With Lemma 16.4.9 we deduce that all reference points dominated by $R$ will also allow to obtain the left extreme.

We will now follow the same steps for the right extreme. Let us fix $\varepsilon_1 \in \mathbb{R}_{>0}$ and let $R = (\mathcal{R}_1^{\text{Nadir}} + \varepsilon_1, r_2)$ for $r_2 \geq \mathcal{R}_2^{\text{Nadir}}$. Following the same steps for the right extreme, we need to prove that the $\mu$-th partial derivative of $I_{H,w}$ is non zero for all $\overline{x}_\mu^\mu \in [\overline{x}_{\mu-1}^\mu, x_{\max}[$. According to Eq. 16.28,

$$\begin{aligned}\partial_\mu I_{H,w}(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu) &= -\int_{f(\overline{x}_\mu^\mu)}^{f(\overline{x}_{\mu-1}^\mu)} w(\overline{x}_\mu^\mu, y)dy - f'(\overline{x}_\mu^\mu)\int_{\overline{x}_\mu^\mu}^{\mathcal{R}_1^{\text{Nadir}}+\varepsilon_1} w(x, f(\overline{x}_\mu^\mu))dx \\ &\geq -W(f(\overline{x}_{\mu-1}^\mu) - f(\overline{x}_\mu^\mu)) - f'(\overline{x}_\mu^\mu)\int_{\overline{x}_\mu^\mu}^{\mathcal{R}_1^{\text{Nadir}}+\varepsilon_1} w(x, f(\overline{x}_\mu^\mu))dx\end{aligned} \tag{16.53}$$

and since $\overline{x}_\mu^\mu \leq \mathcal{R}_1^{\text{Nadir}}$, we obtain

$$\geq -W(f(\overline{x}_{\mu-1}^\mu) - f(\overline{x}_\mu^\mu)) - f'(\overline{x}_\mu^\mu)\int_{\mathcal{R}_1^{\text{Nadir}}}^{\mathcal{R}_1^{\text{Nadir}}+\varepsilon_1} w(x, f(\overline{x}_\mu^\mu))dx \tag{16.54}$$

By continuity of $f$ and the fact that both $\overline{x}_\mu^\mu$ and $\overline{x}_{\mu-1}^\mu$ converge to $x_{\max}$ the term $W(f(\overline{x}_{\mu-1}^\mu) - f(\overline{x}_\mu^\mu))$ converges to zero. Since $-f'(\overline{x}_\mu^\mu)\int_{\mathcal{R}_1^{\text{Nadir}}}^{\mathcal{R}_1^{\text{Nadir}}+\varepsilon_1} w(x, f(\overline{x}_\mu^\mu))dx$ is strictly positive, we deduce that there exists $\mu_2$ such that for all $\mu \geq \mu_2$, $\partial_\mu I_{H,w}(\overline{x}_1^\mu, \ldots, \overline{x}_\mu^\mu)$ is strictly positive and thus for all $\mu$ larger than $\mu_2$ the $\mu$-th partial derivative of $I_{H,w}$ is non zero, i.e., $\overline{x}_\mu^\mu = x_{\max}$. With Lemma 16.4.9 we deduce that all reference points dominated by $R$ allow to obtain the right extreme. □

### 16.7.10 Results for the ZDT Test Function Suite

There exist six ZDT test problems—ZDT1 to ZDT6—of which ZDT5 has a discrete Pareto front and is therefore excluded from our investigations [378]. In the following, let $d = (d_1, \ldots, d_n) \in \mathbb{R}^n$ denote the

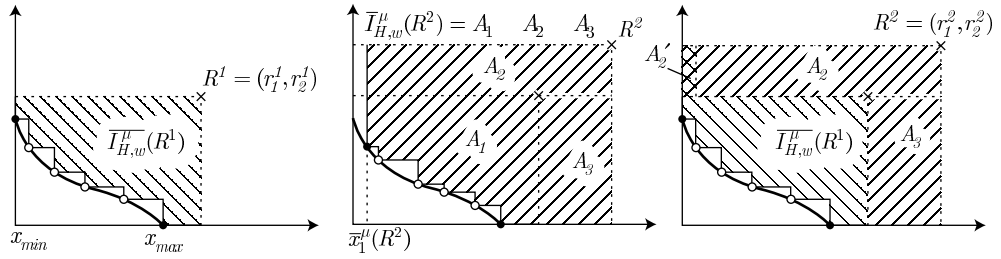



Figure 16.14: If the optimal distribution of $\mu$ points contains the extremes (left-hand side), then after increasing the reference point from $R^1$ to $R^2$ the extremes are still included in the optimal $\mu$-distribution (right-hand side). This can be proven by contradiction (middle).

decision vector of $n$ real-valued variables. The shapes of the Pareto fronts as stated below follow from the definition of the objectives including a function $g(d)$ and the fact that the Pareto front is obtained by setting $g(d) = 1$.

**ZDT1** From Example 16.4.2, we recapitulate the front shape of ZDT1 as $f(x) = 1 - \sqrt{x}$ with $x_{\min} = 0$ and $x_{\max} = 1$, see Figure 16.9(a). From $f'(x) = -1/(2\sqrt{x})$ the density on the front according to Eq. 16.12 is $\delta_F(x) = \frac{3\sqrt[4]{x}}{2\sqrt{4x+1}}$. Since $f'(x_{\min}) = -\infty$, the left extreme is never included as stated already in Example 16.4.2. The lower bound of the reference point $\mathcal{R} = (\mathcal{R}_1, \mathcal{R}_2)$ to have the right extreme, according to Eq. 16.18, equals $\mathcal{R}_1 = \sup_{x \in ]x_{\min}, x_{\max}]} x + \frac{1-\sqrt{x}-1}{-1/(2\sqrt{x})} = \sup_{x \in ]0,1]} 3x = 3$.

**ZDT2** From Example 16.3.11, we recapitulate the front shape of ZDT2 as $f(x) = 1 - x^2$ with $x_{\min} = 0$ and $x_{\max} = 1$ and the density of $\delta_F(x) = \frac{3\sqrt{x}}{2\sqrt{1+4x^2}}$ (see Fig. 16.9(b)). The lower bounds for the reference point $\mathcal{R} = (\mathcal{R}_1, \mathcal{R}_2)$ to obtain the extremes are according to the equations Eq. 16.18 and Eq. 16.15 $\mathcal{R}_1 = \sup_{x \in ]x_{\min}, x_{\max}]} x + \frac{1-x^2-1}{-2x} = \sup_{x \in ]0,1]} \frac{3}{2}x = \frac{3}{2}$ and $\mathcal{R}_2 = \sup_{x \in [x_{\min}, x_{\max}[} -2x \cdot (x-1) + 1 - x^2 = \sup_{x \in [0,1[} 2x - 3x^2 + 1 = \frac{4}{3}$ respectively.

**ZDT3** Due to the sine-function in the definition of ZDT3's second objective, the front is discontinuous where $f : D \to [-1, 1], x \mapsto 1 - \sqrt{x} - x \cdot \sin(10\pi x)$ where $D = [0, 0.0830] \cup (0.1823, 0.2578] \cup (0.4093, 0.4539] \cup (0.6184, 0.6525] \cup (0.8233, 0.8518]$ is derived numerically. Hence $x_{\min} = 0$ and $x_{\max} = 0.8518$. The density is

$$\delta_F(x) = C \cdot \sqrt{\frac{1}{2\sqrt{x}} + \sin(10\pi x) + 10\pi x \cos(10\pi x)} \Bigg/ \sqrt{1 + \left(\frac{1}{2\sqrt{x}} + \sin(10\pi x) + 10\pi x \cos(10\pi x)\right)^2} \text{ with } C \approx 1.5589$$

where $x \in D$ and $\delta_F(x) = 0$ otherwise. Figure 16.9(c) shows the Pareto front and the density. Since $f'(x_{\min}) = -\infty$ and $f'(x_{\max}) = 0$, the left and right extremes are never included.

**ZDT4** The Pareto front of ZDT4 is again given by $f(x) = 1 - \sqrt{x}$. Hence, the density and the choice of the reference point is the same as for ZDT1.

**ZDT6** The Pareto front of ZDT6 is $f : [x_{\min}, x_{\max}] \to [0, 1], x \mapsto 1 - x^2$ with $x_{\min} \approx 0.2808$ and $x_{\max} = 1$, see Fig. 16.9(d). Hence, the Pareto front coincides with the one of ZDT2 except for $x_{\min}$ which is shifted slightly to the right. From this, it follows that also the density is the same except for a constant factor,

i.e., $\delta_F(x)$ is larger than the density for ZDT2 by a factor of $\approx 1.25$. For the lower bound $\mathcal{R}$ of the reference point, we obtain

$$\mathcal{R}_1 = \sup_{x \in ]x_{\min}, x_{\max}]} x + \frac{1 - x^2 - (1 - x_{\min}^2)}{-2x} = \sup_{x \in ]0.2808, 1]} \frac{x_{\min}^2 - 3x^2}{-2x} = \frac{3 - x_{\min}^2}{2} \approx 1.461 \quad \text{and}$$

$$\mathcal{R}_2 = \sup_{x \in [x_{\min}, 1[} -2x(x - x_{\max}) + 1 - x = \sup_{x \in [x_{\min}, 1[} 2x - 3x^2 + 1 = \frac{4}{3} \ .$$

Hence, the lower bound $\mathcal{R}_2$ is the same as for ZDT2, but $\mathcal{R}_1$ differs slightly from ZDT2.

### 16.7.11 Results for the DTLZ Test Function Suite

The DTLZ test suite offers seven test problems which can be scaled to any number of objectives [119]. For the bi-objective variants, DTLZ5 and DTLZ6 are degenerated, i.e., the Pareto fronts consist of only a single point and are not examined in the following. For the definitions of the problems, we refer to [119] and only state the shapes of the Pareto fronts which can be obtained by setting $g(d) = 0$ similar to the ZDT problems.

**DTLZ1** The Pareto front of DTLZ1 is described by $f(x) = 1/2 - x$ with $x_{\min} = 0$ and $x_{\max} = 1/2$, see Fig. 16.9(e). According to Eq. 16.12, we have $\delta_F(x) = \sqrt{2}$. A lower bound for the reference point is given by $\mathcal{R}_1 = \sup_{x \in ]0,1/2]} 1 - x = 1$ and $\mathcal{R}_2 = \mathcal{R}_1$ for symmetry reasons.

**DTLZ2** From Example 16.2.1, we recapitulate the front shape of $f(x) = \sqrt{1 - x^2}$ with $x_{\min} = 0$ and $x_{\max} = 1$, see Fig. 16.9(f). According to Eq. 16.12, the density on the front is $\delta_F(x) = \sqrt{\pi x} \sqrt[4]{1 - x^2} \Big/ \Gamma(3/4)^2$ where $\Gamma$ denotes the gamma-function, i.e., $\Gamma(3/4) \approx 1.225$. A lower bound for the reference point is given by

$$\mathcal{R}_1 = \sup_{x \in ]x_{\min}, x_{\max}]} x + \frac{\sqrt{1-x^2} - \sqrt{1-x_{\min}^2}}{-x/\sqrt{1-x^2}} = \sup_{x \in ]0,1]} \frac{\sqrt{1-x^2} - 1 + 2x^2}{x} = 1/2 \left(\sqrt{3} - 1\right) 3^{3/4} \sqrt{2} \approx 1.18$$

and for symmetry reasons $\mathcal{R}_2 = \mathcal{R}_1$.

**DTLZ3** The problem formulation of DTLZ3 is the same as for DTLZ2 except for the function $g(d)$. However, the Pareto front is formed by the same decision vectors as for DTLZ2 and the fronts of DTLZ2 and DTLZ3 are identical. Hence, also the density and the choice of the reference point are the same as for DTLZ2.

**DTLZ4** In DTLZ4, the same functions as in DTLZ2 are used with an additional meta-variable mapping $m : [0, 1] \to [0, 1]$ of the decision variables, i.e., the decision variable $m(d_i) = d_i^\alpha$ is used instead of the original decision variable $d_i$ in the formulation of the DTLZ2 function. This transformation does not affect the shape of the Pareto front and the results on optimal $\mu$-distributions for the unweighted hypervolume indicator again coincide with the ones for DTLZ2.

**DTLZ7** The Pareto front of DTLZ7 is discontinuous and described by the function $f : D \to [0, 4]$, $x \mapsto 4 - x(1 + \sin(3\pi x))$ where $D = [0, 0.2514] \cup (0.6316, 0.8594] \cup (1.3596, 1.5148] \cup (2.0518, 2.1164]$ which is derived numerically, see Fig. 16.9(g). Hence, $x_{\min} = 0$ and $x_{\max} \approx 2.1164$. The derivative of $f(x)$ is $f'(x) = -1 - \sin(3\pi x) - 3\pi x \cos(3\pi x)$ and the density therefore is

$\delta_F(x) = C \cdot \sqrt{1 + \sin(3\,\pi\,x) + 3\pi x \cos(3\,\pi\,x)} \Big/ \sqrt{1 + \left(1 + \sin(3\,\pi\,x) + 3\pi x \cos(3\,\pi\,x)\right)^2}$ with $C \approx 0.6566$. For $\mathcal{R}$, we find $\mathcal{R}_1 \approx 2.481$ and $\mathcal{R}_2 \approx 13.3720$.

### 16.7.12 Results for the WFG Test Function Suite

The WFG test suite offers nine test problems which can be scaled to any number of objectives. In contrast to DTLZ and ZDT, the problem formulations are build using an arbitrary number of so-called *transformation functions*. We abstain from quoting these functions here and refer the interested reader to [193]. The resulting Pareto front shape is determined by parameterized *shape functions* $h_i$ mapping $[0, 1]$ to the range $[0, 1]$. All test functions WFG4 to WFG9 share the same shape functions and are therefore examined together in the following.

**WFG1** For WFG1, the shape functions are convex and mixed respectively which leads to the Pareto front $f(x) = \frac{2\rho - \sin(2\,\rho)}{10\pi} - 1$ with $\rho = 10 \arccos(1 - x)$, $x_{\min} = 0$ and $x_{\max} = 1$, see Fig. 16.9(h). The density becomes

$$\delta_F(x) = C \cdot \sqrt{\frac{2\,(1 - \cos(2\rho))\,\pi}{\sqrt{x\,(2 - x)}\left(\pi^2 - 4\,\frac{(1-\cos(2\rho))^2}{x(x-2)}\right)}}$$

with $C \approx 1.1569$. Since $\lim_{x \to x_{\max}} f'(x_{\max}) = 0$ the rightmost extreme point is never included in an optimal $\mu$-distribution for $I_{H,w}$. For the choice of $\mathcal{R}_2$ the analytical expression is very long and therefore omitted. A numerical approximation leads to $\mathcal{R}_2 \approx 0.9795$.

**WFG2** For WFG2, the shape functions are convex and discontinuous respectively which leads to the discontinuous Pareto front $f : D \to [0, 1]$, $x \mapsto 1 - 2\,\frac{(\pi - 0.1\rho)\cos^2(\rho)}{\pi}$ where $\rho = \arccos(x - 1)$, and with a numerically derived domain $D = [0, 0.0021] \cup (0.0206, 0.0537] \cup (0.1514, 0.1956] \cup (0.3674, 0.4164] \cup (0.6452, 0.6948] \cup (0.9567, 1]$, $x_{\min} = 0$ and $x_{\max} = 1$, see Fig. 16.9(i). The density becomes

$$\delta_F(x) = C \cdot \frac{\sqrt{-f'(x)}}{\sqrt{1 + f'(x)^2}} \quad \text{with } C \approx 0.44607 \text{ and } f'(x) = -2\,\frac{\cos(\rho)\,(\cos(\rho) + 20\,\sin(\rho)\,\pi - 2\,\sin(\rho)\,\rho)}{\sqrt{x\,(2 - x)}\pi}$$

for all $x \in D$ and $\delta_F(x) = 0$ otherwise. Again, $f'(0) = -\infty$ such that the leftmost extreme point is never included in an optimal $\mu$-distribution for $I_{H,w}$. For the rightmost extreme one finds $\mathcal{R}_1 \approx 2.571$.

**WFG3** For WFG3, the shape functions are both linear—leading to the linear Pareto front $f(x) = 1 - x$ with $x_{\min} = 0$ and $x_{\max} = 1$. Hence, the density is $\delta_F(x) = 1/\sqrt{2}$, see Fig. 16.9(e) for a scaled version of this Pareto front. For the choice of the reference point the same arguments as for DTLZ1 hold, which leads to $\mathcal{R} = (2, 2)$.

**WFG4 to WFG9** For the six remaining test problems WFG4 to WFG9, the shape functions $h_1$ and $h_2$ are both concave—resulting in a spherical Pareto front $f(x) = \sqrt{1 - x^2}$ with $x_{\min} = 0$ and $x_{\max} = 1$. Hence, the Pareto front coincides with the front of DTLZ2 and also the density and the choice of the reference point are the same.

# Part IV

# Promising Ideas

# Chapter 17

# Technological architecture evolutions of Information Systems: trade-off and optimization

Vassilis Giakoumakis, Daniel Krob, Leo Liberti, Fabio Roda

*Submitted to Concurrent Engineering Research and Applications*

*In the normal lifespan of large enterprises, the strategic management of IT often evolves. Existing services must be replaced with new services without impairing operations. The problem of scheduling such replacement is of critical importance for the success of the operation. We analyze this problem from a quantitative point of view, underlining the trade-off nature of its solutions. We formalize this multi-objective optimization problem as a mathematical programming formulation. We discuss its theoretical properties and show that real-world instances can be solved by standard off-the-shelf tools.*

## 17.1 Introduction

For any information system manager, a recurrent key challenge is to avoid creating more complexity within its existing information system through the numerous IT projects that are launched in order to respond to the needs of the business. Such an objective leads thus typically to the necessity of co-optimizing both creation *and* replacement/destruction — called usually *kills* in the IT language — of parts of the information system, and of prioritizing the IT responses to the business consequently.

This important question is well known in practice and quite often addressed in the IT literature, but basically only from an enterprise architecture or an IT technical management perspective [56, 94, 264]. Architectural and managerial techniques, however, are often only parts of the puzzle that one has to solve to handle these optimization problems. On the basis of budget, resource and time constraints given by the enterprise management, architecture provides the business and IT structure of these problems. This is however not sufficient to model them completely or solve them. Nevertheless, from a methodological point of view, real systems (especially information ones) can be described adopting different perspectives and different levels of details according to Systems Architecture methodologies that we can find in literature [68, 69, 225]. Some powerful techniques help to manage complex systems decoupling them into parts. *Abstraction* and *Concretization* are the processes that allow to shift from a specific level of details to a properly lower/higher one so that we can focus on the features which are really relevant and

"hide" the others. *Decomposition* and *Integration* are the ones that let the analyst to identify and solve sub-problems. Alternating these activities we can provide a series of partial *visions* which cooperate to produce the whole picture, instead of a single global description.

According to these seminal concepts, in a previous work [162] we proposed a first abstract model (or *vision*) of the problem and moved a first step towards the integration of architectural business and IT project management aspects. More precisely, we proposed an operational model and a Mathematical Programming (MP) formulation expressing a generic global prioritization problem occurring in the — limited, but practically relevant — context of a technological evolution of an information system, reducing to one all the different objectives of the different stakeholders.

In this paper we refine our analysis. In the interest of simplicity, the first model proposed understated the multi-objective nature of the problem. In particular the presence of many teams at work is worth a closer examination, because it can lead to puzzling situations if some specific conditions hold or resources are insufficient. If we introduce a bound on the duration of the whole transition, or, at least, on each step of it, and force the activation/deactivations to be done before a short deadline we generate a competition between departements. Thus, we employ multi-optimization techniques in order to model and numerically solve a wider part of this general problem. This is a second step towards a full formalization of the problem which promises to provide a valuable help for IT practitioners.

The rest of this paper is organized as follows: section 2 describes the problem and the key elements involved, section 3 proposes the Mathematical Programming formulation and introduces the necessary multi-optimization techniques, section 4 discusses the theoretical properties of the model and section 5 reports the results of the computational tests.

## 17.2 Operational model of an evolving information system

### 17.2.1 Elements of information system architecture

Any information system of an enterprise (consisting of a set $D$ of departments) is classically described by two architectural layers:

- the *business layer*: the description of the business services (forming a set $V$) offered by the information system;
- the *IT layer*: the description of the IT modules (forming a set $U$) on which business services rely on.

In general, the relationship $A \subseteq V \times U$ between these two layers is not one-to-one. A given business service can require any number of IT modules to be delivered and vice-versa a given IT module can be involved in the delivery of several business services, as shown in Fig. 17.1.

### 17.2.2 Evolution of an information system architecture

From time to time, an information system may evolve in its entirety due to the replacement of an existing software technology by a new one (e.g. passing from several independent/legacy software packages to an integrated one, migrating from an existing IT technology to a new one, and so on). These evolutions invariably have a strong impact at the IT layer level, where the existing IT modules $U^E = \{M_1, \ldots, M_n\}$ are replaced by new ones in a set $U^N = \{N_1, \ldots, N_{n'}\}$ (in the sequel, we assume $U = U^E \cup U^N$). This translates to a replacement of existing services (sometimes denoted ES) in $V$ by new services (sometimes denoted NS) in $W$ ensuring that the impact on the whole enterprise is kept low in order to avoid business discontinuity. This also induces a relation $B \subseteq W \times U^N$ expressing reliance of new services on IT modules.

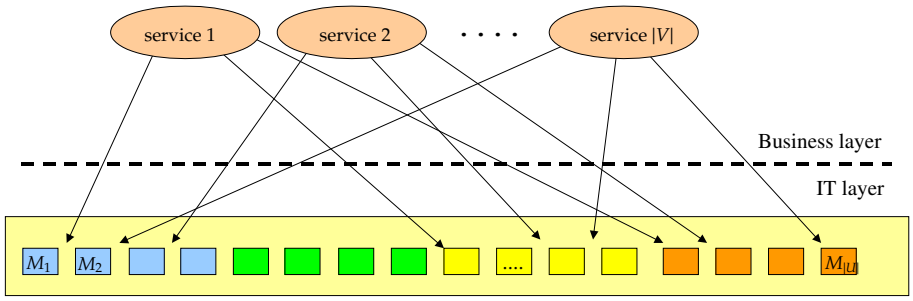



Figure 17.1: A simple two-layer information system architecture.

Note also that in this context, at the business level, there exists a relation (in $V \times W$) between existing services and new services which expresses the fact that a given existing service shall be replaced by a subset of new business services. We note in passing that this relation also induces another relation in $U^E \times U^N$ expressing the business covering of an existing IT module to a subset of new IT modules (see Fig. 17.2).

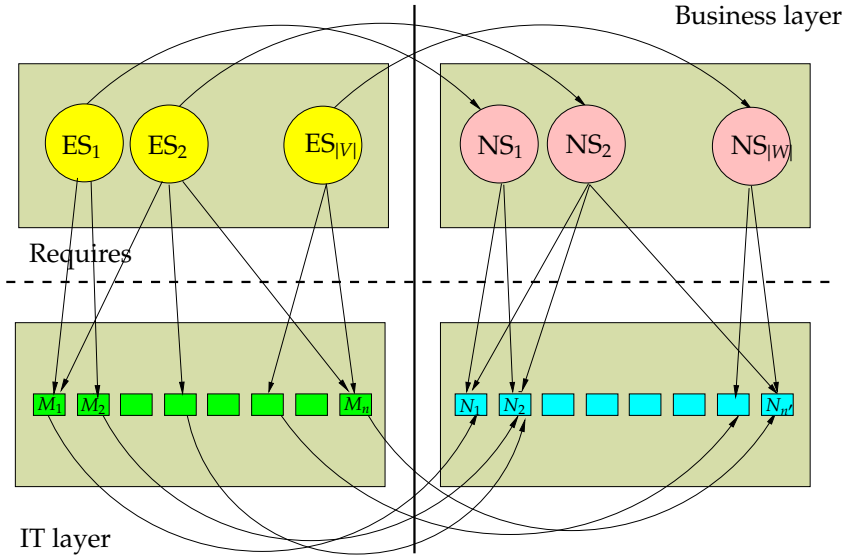



Figure 17.2: Evolution of an information system architecture.

### 17.2.3 Management of information system architecture evolutions

Mapping the above information system architecture on the organization of a company, it appears clear that three main types of enterprise actors are naturally involved in the management of these technological evolutions which are described below.

1. *Business department managers*: they are responsible of creating business value — within the perimeter of a business department in the set $D$ — through the new business services. This value might be measured by the amount of money they are ready to invest in the creation of these services (business services are usually bought internally by their users within the enterprise).

2. *IT project managers*: they are responsible for creating the new IT modules, which is a pre-requisite to creating the associated business services. The IT project manager has a project schedule usually organized in workpackages, each having a specific starting times and global budget (see Fig. 17.2): in our case, this schedule is presented as a set of deployment precedence rules for new modules.
3. *Kill managers*: they are responsible for destroying the old IT modules in order to avoid to duplicate the information system — and therefore its operating costs — when achieving its evolution. Kill managers have a budget for realizing such "kills", but they must ensure that any old IT module $i$ is only killed after the new services replacing those old ones relying on $i$ are put into service. The enterprise motivates the efficiency of kill managers by setting a monetary value on each deactivation: this provides a measure of the desirability of killing module $i$.

In this context, managing the technological evolution of an information system means being able of creating new IT modules within the time and budget constraints of the IT project manager in order to maximize both the IT modules business value brought by the new services and the associated kill value (i.e. the number of old services than can be killed).

### 17.2.4 The information system architecture evolution management problem

The architecture evolution of the IT system involves revenues, costs and schedules over a time horizon $t_{\max}$, as detailed below.

- *Time and budget constraints of the IT project manager*. Each new IT module $i \in U$ has a cost $a_i$ and a production schedule.
- *IT module business value*. Each department $\ell \in D$ is willing to pay $q_{\ell k}$ monetary units for a new service $k \in W$ from a departmental production budget $H^\ell = \sum_{k:(\ell,k)\in F} q_{\ell k}$; the business value of the new service $k$ is $c_k = \sum_{\ell:(\ell,k)\in F} q_{\ell k}$. We assume that this business value is transferred in a conservative way via the relation $B$ to the IT modules. Thus, there is a business contribution $\beta_{ik}$ over every $(i,k) \in B$ such that for each $k$ we have $c_k = \sum_{(i,k)\in B} \beta_{ik}$; furthermore, the global business value of module $i$ is $\sum_{k:(i,k)\in B} \beta_{ik}$. We also introduce a set $N \subseteq U$ of IT modules that are necessary to the new services.
- *Deployment schedule of new modules*. We are given a Directed Acyclic Graph (DAG) $(U, S)$ where each couple $(i,h) \in S \subset U \times U$ encodes a deployment precedence between the new modules $i, h$ (i.e. $h$ cannot be deployed before $i$).
- *Kill value*. Discontinuing (or *killing*) a module $i \in U$ has a cost $b_i$ due to the requirement, prior to the kill, of an analysis of the interactions between the module and the rest of the system architecture, in order to minimize the chances of the kill causing unexpected system behaviour. As mentioned above, it also has a monetary value (or desirability) $\phi_i$ .

The evolution involves several stakeholders. The department heads want to maximize the value of the required new services. The module managers want to produce the modules according to an assigned schedule whilst maximizing the business value for the new services to be activated. The kill managers want to maximize the monetary value of the deactivated modules within a certain kill budget. Thus, the rational planning of this evolution requires the solution of an optimization problem with several constraints and criteria, which we shall discuss in the next session.

## 17.3 Mathematical Programming based approach

Mathematical Programming (MP) is a formal language used for modelling and solving optimization problems [253, 363]. Each problem is modelled by means of a list of index sets, a list of known parameters

encoding the problem data (the *instance*), a list of decision variables, which will contain appropriate values after the optimization process has taken place, an objective function to be minimized or maximized, and a set of constraints. The objective and constraints are expressed in function of the decision variables and the parameters. The constraints might include integrality requirements on the decision variables. MPs are classified into Linear Programs (LP), Mixed-Integer Linear Programs (MILP), Nonlinear Programs (NLP), Mixed-Integer Nonlinear Programs (MINLP) according to the linearity of objective and constraints and to integrality requirements on the variables. MILPs and MINLPs are usually solved using a Branch-and-Bound (BB) method, explained at the beginning in Sect. 17.3.4. A *solution* is an assignment of numerical values to the decision variables. A solution is *feasible* if it satisfies the constraints. A feasible solution is *optimal* if it optimizes the objective function.

### 17.3.1 Multiobjective Programming

Multiobjective Programming (MOP) [158] is a modification of the MP language that allows for sets of objective functions to be supplied. A feasible solution $s$ dominates a second one $s'$ (denoted by $s \prec s'$ if it is at least as good as the first one on all objectives, and strictly better on at least one objective. MOP is dealt with either by looking for the *efficient set* (i.e. the set of all nondominated feasible solutions) or by reformulating the MOP to a single-objective MP which finds one efficient solution. All efficient points are possible choices but the ultimate decision maker usually wants only one or a few suggestions. In this case the resolution of a MOP includes a first phase, the production of all not dominated points and a second one, the selection of the most desirable ones in order to satisfy some kind of preference related to the users. The role played by these preferences determines four approaches present in literature (*no-preference*, *"a posteriori"*, *"a priori"*, *interactive*) according to [275]. In particular the *no-preference* methods do not use preferences and propose only the best solution. In this case, the two phases are not clearly distincted and the final choice is always based on a criterion which is independent from the user. A method which represents well this class is the $L_p$-metric method (see Figure 17.3 above). The basic idea of this method is to search in the objective space the point which has the minimal distance from a fixed point $h^*$. Typically, the chosen reference is the *ideal point (utopia)* , which corresponds to the maximal values of all objectives. The ideal point is the one we would chose if there were no constraints and no trade-offs between tasks. Formally it is defined as $h^* \in \mathbb{R}^n$:

$$h_i^* = max\{f_i(x)|x \in X\}, i = 1, \ldots, n; \tag{17.1}$$

Different metrics can be chosen but the most common choice is:

$$d_p(x_1, x_2) = \left( \sum_{i=1}^{n} |x_{1_i} - x_{2_i}|^p \right)^{1/p} \tag{17.2}$$

which determines the following minimization problem:

$$\min \left( \sum_{i=1}^{n} |f_i(x) - h_i^*|^p \right)^{1/p} \tag{17.3}$$

### 17.3.2 Introducing the MP formulation

As explained above, an enterprise in our context consists of a set $D$ of departments currently relying on existing services in $V$ and wishing to evolve to new services in $W$ within a time horizon $t_{\max}$. Each service relies on some IT module in $U$ (the set $N \subseteq U$ indexes those IT modules that are necessary). The relations between services and modules and, respectively, departments and services, are denoted as follows: $A \subseteq V \times U$, $B \subseteq W \times U$, $E \subseteq D \times V$ and $F \subseteq D \times W$. If an IT module $i \in U$ is required by a new service, then it must be produced (or activated) at a certain cost $a_i$. When an IT module $i \in U$ is no

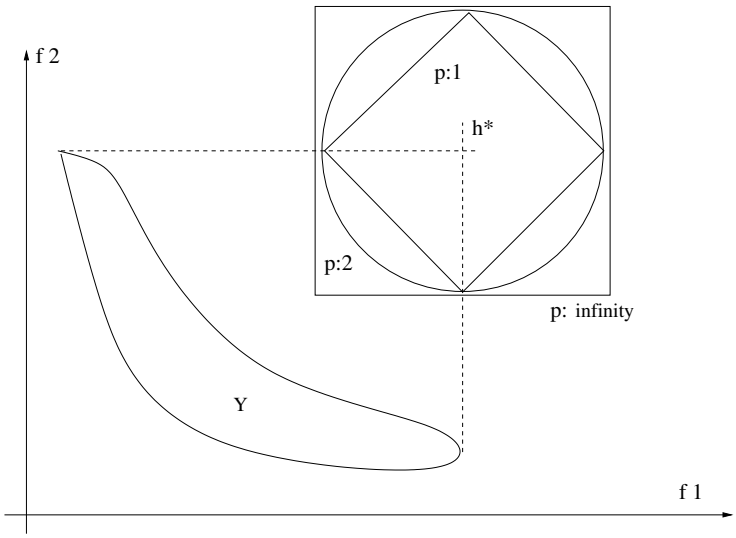



Figure 17.3: $L_p$-metric method

longer used by any service it must be killed at a certain cost $b_i$. Departments can discontinue using their existing services only when all new services providing the functionalities have been activated; when this happens, the service (and the corresponding IT modules) can be killed; a killed module $i$ contributes $\phi_i$ monetary units to the goal of the kill manager. Departments have budgets dedicated to producing and killing IT modules, which must be sufficient to perform their evolution to the new services; for the purposes of this paper, we suppose that departmental budgets are interchangeable, i.e. all departments credit and debit their costs and revenues to two unique enterprise-level budgets: a production budget $H_t$ and a kill budget $K_t$ indexed by the time period $t$. A new service $k \in W$ has a value $c_k$, and an IT module $i \in U$ contributes $\beta_{ik}$ to the value of the new service $k$ that relies on it. We use the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ shown in Fig. 17.4 to model departments, existing services, new services, IT modules and their relations. The vertices are $\mathcal{V} = U \cup V \cup W \cup D$, and the edges are $\mathcal{E} = A \cup B \cup E \cup F$. This graph is the union of the four bipartite graphs $(U, V, A)$, $(U, W, B)$, $(D, V, E)$ and $(D, W, F)$ encoding the respective relations. We remark that $E$ and $F$ collectively induce a relation between existing services and new services with a "replacement" semantics (an existing service can be killed if the related new services are active).

### 17.3.3 Sets, variables, objectives, constraints

We present here the MP formulation of the Architecture Evolution Problem (AEP). We recall that NS stands for new service and ES for existing service.

1. **Sets**:
   - $T = \{0, \ldots, t_{\max}\}$: set of time periods (Sect. 17.2.4, p. 240);
   - $U$: set of IT modules (Sect. 17.2.1, p. 238);
   - $N \subseteq U$: set of IT modules that are necessary for the NS (Sect. 17.2.4, p. 240);
   - $V$: set of existing services (Sect. 17.2.1, p. 238);
   - $W$: set of new services (Sect. 17.2.2, p. 238);
   - $A \subseteq V \times U$: relations between ES and IT modules (Sect. 17.2.1, p. 238);
   - $B \subseteq W \times U$: relations between NS and IT modules (Sect. 17.2.2, p. 238);
   - $D$: set of departments (Sect. 17.2.1, p. 238);

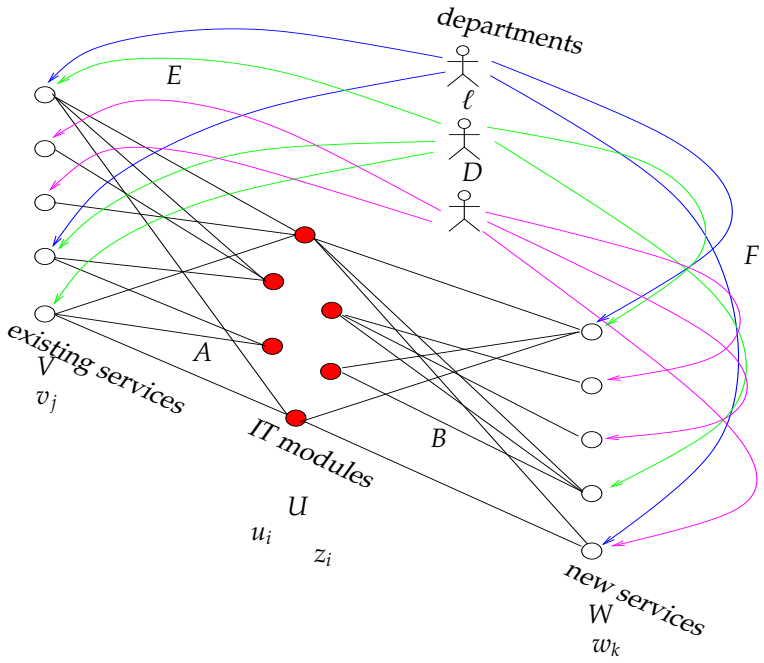



Figure 17.4: The bipartite graphs used to model the problem.

- $E \subseteq D \times V$: relations between departments and ES (Sect. 17.3.2, p. 242);
- $F \subseteq D \times W$: relations between departments and NS (Sect. 17.3.2, p. 242);
- $S \subset N \times N$: deployment precedences between new modules (Sect. 17.3.2, p. 240).

2. **Parameters**:

- $\forall i \in U$ $a_i$ = cost of producing an IT module (Sect. 17.2.4, p. 17.2.4);
- $\forall i \in U$ $b_i$ = cost of killing an IT module (Sect. 17.2.4, p. 17.2.4);
- $\forall i \in U$ $\phi_i$ = desirability (monetary units) of killing an IT module (Sect 17.2.4, p. 17.2.4)
- $\forall t \in T$ $H_t$ = production budget per time period (Sect. 17.3.2, p. 17.3.2);
- $\forall t \in T$ $K_t$ = kill budget per time period (Sect. 17.3.2, p. 17.3.2);
- $\forall (i,k) \in B$ $\beta_{ik}$ = monetary value given to NS $k$ by IT module $i$ (Sect. 17.2.4, p. 17.2.4).

3. **Decision variables**:

$$\forall i \in U, t \in T \; u_{it} = \begin{cases} 1 & \text{if IT module } i \text{ is used for a ES at time } t \\ 0 & \text{otherwise;} \end{cases} \tag{17.4}$$

$$\forall i \in U, t \in T \; z_{it} = \begin{cases} 1 & \text{if IT module } i \text{ is used for a NS at time } t \\ 0 & \text{otherwise;} \end{cases} \tag{17.5}$$

$$\forall j \in V, t \in T \; v_{jt} = \begin{cases} 1 & \text{if existing service } j \text{ is active at time } t \\ 0 & \text{otherwise;} \end{cases} \tag{17.6}$$

$$\forall k \in W, t \in T \; w_{kt} = \begin{cases} 1 & \text{if new service } k \text{ is active at time } t \\ 0 & \text{otherwise.} \end{cases} \tag{17.7}$$

4. **Objective functions**.

- Business gain: value contributed to new services by IT modules. This is the objective of the module managers, and, because $\beta$ is indexed on both modules and services, it agrees with the objective of the department heads.

$$\max_{u,v,w,z} \quad \sum_{\substack{t\in T\\(i,k)\in B}} \beta_{ik} z_{it} w_{kt}. \tag{17.8}$$

- Killing gain: objective of the kill managers.

$$\max_{u,v,w,z} \quad \sum_{\substack{t\in T\\i\in U}} \phi_i (1 - u_{it}). \tag{17.9}$$

5. **Constraints**.

- Production budget (cost of producing new IT modules; this is another objective of the module managers):

$$\forall t \in T \setminus \{t_{\max}\} \quad \sum_{i\in U} a_i (z_{i,t+1} - z_{it}) \leq H_t, \tag{17.10}$$

  where the term $z_{i,t+1} - z_{it}$ is only ever 1 when a new service requires production of an IT module — we remark that the next constraints prevent the term from ever taking value $-1$.
- Once an IT module is activated, do not deactivate it.

$$\forall t \in T \setminus \{t_{\max}\}, i \in U \quad z_{it} \leq z_{i,t+1}. \tag{17.11}$$

- Kill budget (cost of killing IT modules; this is part of the objective of the kill managers):

$$\forall t \in T \setminus \{t_{\max}\} \quad \sum_{i\in U} b_i (u_{it} - u_{i,t+1}) \leq K_t, \tag{17.12}$$

  where the term $u_{it} - u_{i,t+1}$ is only ever 1 when an IT module is killed — we remark that the next constraints prevent the term from ever taking value $-1$.
- Once an IT module is killed, cannot activate it again.

$$\forall t \in T \setminus \{t_{\max}\}, i \in U \quad u_{it} \geq u_{i,t+1}. \tag{17.13}$$

- If an existing service is active, the necessary IT modules must also be active:

$$\forall t \in T, (i,j) \in A \quad u_{it} \geq v_{jt}. \tag{17.14}$$

- If a new service is active, the necessary IT modules must also be active:

$$\forall t \in T, (i,k) \in B : i \in N \quad z_{it} \quad \geq \quad w_{kt}. \tag{17.15}$$

- An existing service can be deactivated once all departments relying on it have already switched to new services; for this purpose, we define sets $\mathcal{W}_j = \{k \in W \mid \exists \ell \in D\ ((\ell, j) \in E \wedge (\ell, k) \in F)\}$ for all $j \in V$:

$$\forall t \in T, j \in V \quad \sum_{k\in\mathcal{W}_j} (1 - w_{kt}) \leq |\mathcal{W}_j| v_{jt}. \tag{17.16}$$

- New modules must be deployed according to precedences: for a precedence $(i,h) \in S$, $i$ must be deployed at least one timestep before $h$ is; therefore, if $z_{it} = 0$ then $z_{hs} = 0$ for all $s \leq t$ (17.17), and if $t$ is the first timestep where $z_{it} = 1$ then $z_{ht} = 0$ (17.18):

$$\forall s \leq t \in T, (i,h) \in S \quad z_{hs} - z_{it} \quad \leq \quad 0 \tag{17.17}$$

$$\forall s \leq t \in T \setminus \{0\}, (i,h) \in S \quad z_{hs} + z_{it} - z_{i,t-1} \quad \leq \quad 1. \tag{17.18}$$

- Boundary conditions. To be consistent with the objectives of the module and kill managers, we postulate that:
  - at $t = 0$ all IT modules needed by existing services are active, all IT modules needed by new services are inactive:

$$\forall i \in U \; u_{i0} = 1 \quad \wedge \quad z_{i0} = 0; \tag{17.19}$$

$$\forall j \in V \; v_{j0} = 1 \quad \wedge \quad \forall k \in W \; w_{k0} = 0. \tag{17.20}$$

  - at $t = t_{\max}$ all IT modules needed by the existing services have been killed:

$$\forall i \in U \quad u_{it_{\max}} = 0. \tag{17.21}$$

  These boundary conditions are a simple implementation of the objectives of module and kill managers. Similar objectives can also be pursued by adjoining further constraints to the MP, such as for example that the number of IT modules serving ES must not exceed a given amount.

Apart from the fact that the formulation above has two objective functions, it belongs to the Binary Quadratic Programming (BQP) class, as a product of decision variables appears in the objective function and all variables are binary. BQPs can either be solved directly using standard BB-based solvers [51, 200, 320] or reformulated exactly (see [252] for a formal definition of *reformulation*) to a MILP, by means of the ProdBin reformulation [153, 253] prior to solving is with standard MILP solvers.

Intuitively, the multi-objective aspect of the BQP presented above is apparent if one takes into account the production and kill budget constraints. Moreover, in practice these budgets are both part of an enterprise-wide budget, so that (17.10) and (17.12) could be replaced by the following constraints:

$$\forall t \in T \setminus \{t_{\max}\} \quad \sum_{i \in U} (a_i(z_{i,t+1} - z_{it}) + b_i(u_{it} - u_{i,t+1})) \leq H_t + K_t. \tag{17.22}$$

Constraints (17.22) emphasize the trade-off nature of the two objectives (17.8)-(17.9), since they imply, respectively, $z$ and $u$ variables. Formally, we shall show in Sect. 17.4 that the trade-off nature of this problem is implied by constraints 17.16.

### 17.3.4 Valid cuts from implied properties

The BB method for for MPs with binary variables performs a binary tree-like recursive search. At every node, a lower bound to the optimal objective function value is computed by solving a continuous relaxation of the problem. If all integral variables happen to take integer values at the optimum of the relaxation, the node is *fathomed* with a feasible optimum. If this optimum has better objective function value than the feasible optima found previously, it replaces the *incumbent*, i.e. the best current optimum. Otherwise, a variable $x_j$ taking fractional value $\bar{x}_j$ is selected for branching. Two subnodes of the current node are created by imposing constraints $x_j \leq \lfloor \bar{x}_j \rfloor$ (left node) and $x_j \geq \lceil \bar{x}_j \rceil$ (right node) to the problem. If the relaxed objective function value at a node is worse than the current incumbent, the node is also fathomed. The step of BB which most deeply impacts its performance is the computation of the lower bound. To improve the relaxation quality, one often adjoins "redundant constraints" to the problem whenever their redundancy follows from the integrality constraints. Thus, such constraints will not be redundant with respect to the relaxation. An inequality is *valid* for a MP if it is satisfied by all its feasible points. If an inequality is valid for an MP but not for its relaxation, it is called a *valid cut*.

We shall now discuss two valid inequalities for the evolution problem. The first one stems from the following statement: *If a new service $k \in W$ is inactive, then all existing services linked to all departments relying on $k$ must be active*. We formalize this statement by defining the sets:

$$\forall k \in W \quad \mathcal{V}_k = \{j \in V \mid \exists \ell \in D \; ((\ell, j) \in E \wedge (\ell, k) \in F)\}.$$

The statement corresponds to the inequality:

$$\forall t \in T, k \in W \quad \sum_{j \in \mathcal{V}_k} (1 - v_{jt}) \leq |\mathcal{V}_k| w_{kt}. \tag{17.23}$$

**17.3.1 Lemma**
*Whenever* $(v, w)$ *are part of a feasible solution of the evolution problem,* (17.16) $\Leftrightarrow$ (17.23).

*Proof.* Firstly, we start proving that (17.16) $\Rightarrow$ (17.23). We proceed by contradiction: suppose (17.16) holds and (17.23) does not. Then there must be $t \in T, k \in W, j \in \mathcal{V}_k$ such that $w_{kt} = 0$ and $v_{jt} = 0$. By (17.16), $v_{jt} = 0$ implies $\forall h \in \mathcal{W}_j$ $(w_{ht} = 1)$. By definition of $\mathcal{V}_k$ and $\mathcal{W}_j$, we have that $k \in \mathcal{W}_j$, and hence $w_{kt} = 1$ against the assumption. Secondly, we observe that the converse, (17.16) $\Leftarrow$ (17.23), also holds. The proof is symmetric: it suffices to swap $j$ with $k$, $\mathcal{W}_j$ with $\mathcal{V}_k$, $v$ with $w$, (17.16) with (17.23). □

We remark that (17.16) $\Rightarrow$ (17.23) let us assert that (17.23) is a valid inequality for the AEP. The second inequality is a simple relation between $v$ and $w$.

**17.3.2 Proposition**
*The inequalities*

$$\forall t \in T, j \in V, k \in W \; \exists \ell \in D \; ((\ell, j) \in E \wedge (\ell, k) \in F) \quad v_{jt} + w_{kt} \geq 1 \tag{17.24}$$

*are valid for the AEP.*

*Proof.* Suppose (17.24) does not hold: hence there are $t \in T, j \in V, k \in W, \ell \in D$ with $(\ell, j) \in E$ and $(\ell, k) \in F$ such that $v_{jt} + w_{kt} = 0$. Since $v_{jt}, w_{kt} \geq 0$, this implies $v_{jt} = w_{kt} = 0$. It is easy to verify that if this is the case, (17.16) and (17.23) cannot both hold, contradicting (17.16) $\Leftrightarrow$ (17.23) (cf. Lemma (17.3.1)). □

Eq. (17.24) states that at any given time period no pair (ES, NS) related to a given department must be inactive (otherwise the department cannot be functional). We can add (17.23) and (17.24) to the MP formulation of the AEP, and hope they will improve the quality of the lower bound obtained via the LP relaxation. We remark that other valid inequalities similar to (17.23), (17.24) can be derived by the problem constraints; these will be studied in further works.

**17.3.3 Corollary**
*The inequalities*

$$\forall t \in T, i \in U \exists j \in V, k \in W, \ell \in D((i, j) \in A \wedge (i, k) \in B \wedge (\ell, j) \in E \wedge (\ell, k) \in F) \quad u_{it} + z_{it} \geq 1 \tag{17.25}$$

*are valid for the AEP.*

*Proof.* The result follows by summing (17.14) and (17.15) and then applying Prop. 17.3.2. □

Eq. (17.25) states that no pairs of (old,new) modules serving the same department can be inactive at the same time.

## 17.4 Formulation properties and trade-off

We investigate some decomposition properties of the original formulation (17.4)-(17.21), without considering (17.22) nor the valid cuts of Sect. 17.3.4. We shall show that (17.16) are the true source of the trade-off nature of (17.4)-(17.21).

Without (17.16), the formulation can be decomposed into two bi-objective subproblems involving only the $u, v$ and respectively $w, z$ variables. This suggests a Lagrangian relaxation [366] of (17.16) using Lagrangian coefficients $\lambda, \mu \geq 0$, which yields the two maximization objectives:

$$\begin{aligned}\zeta(\lambda,u,v,w,z) &= \sum_{\substack{(i,k)\in B\\ t\in T}} \beta_{ik} z_{it} w_{kt} + \sum_{\substack{t\in T\\ j\in V}} \lambda_{jt}\left(\sum_{k\in \mathcal{W}_j}(w_{kt}-1) + |\mathcal{W}_j| v_{jt}\right) = \\ &= \sum_{\substack{t\in T\\ (i,k)\in B}} \beta_{ik} z_{it} w_{kt} + \sum_{\substack{t\in T, j\in V\\ k\in \mathcal{W}_j}} \lambda_{jt} w_{kt} + \sum_{\substack{t\in T\\ j\in V}} |\mathcal{W}_j| \lambda_{jt}(v_{jt}-1),\end{aligned}$$

and

$$\begin{aligned}\eta(\mu,u,v,w,z) &= \sum_{\substack{t\in T\\ i\in U}} \phi_i u_{it} + \sum_{\substack{t\in T\\ j\in V}} \mu_{jt}\left(\sum_{k\in \mathcal{W}_j}(w_{kt}-1) + |\mathcal{W}_j| v_{jt}\right) = \\ &= \sum_{\substack{t\in T\\ i\in U}} \phi_i u_{it} + \sum_{\substack{t\in T, j\in V\\ k\in \mathcal{W}_j}} \mu_{jt} w_{kt} + \sum_{\substack{t\in T\\ j\in V}} |\mathcal{W}_j| \mu_{jt}(v_{jt}-1).\end{aligned}$$

Thus, we can decompose the problem of Sect. 17.3.2 into the two bi-objective Lagrangian subproblems $P, Q$:

$$\left.\begin{array}{rrcl} \max\limits_{u,v} & \sum\limits_{\substack{t\in T\\ j\in V}} |\mathcal{W}_j| \lambda_{jt}(v_{jt}-1) & & \\ \max\limits_{u,v} & \sum\limits_{\substack{t\in T\\ i\in U}} \phi_i u_{it} + \sum\limits_{\substack{t\in T\\ j\in V}} |\mathcal{W}_j| \mu_{jt}(v_{jt}-1) & & \\ \forall t \in T \setminus \{t_{\max}\} & \sum\limits_{i\in U} b_i(u_{it} - u_{i,t+1}) & \leq & K_t \\ \forall t \in T \setminus \{t_{\max}\}, i \in U & u_{it} & \geq & u_{i,t+1} \\ \forall t \in T, (i,j) \in A & u_{it} & \geq & v_{jt} \\ \forall i \in U & u_{i0} & = & 1 \\ \forall j \in V & v_{j0} & = & 1 \\ \forall i \in U & u_{it_{\max}} & = & 0, \end{array}\right\} \tag{17.26}$$

$$\left.\begin{array}{rrcl} \max\limits_{w,z} & \sum\limits_{\substack{t\in T\\ (i,k)\in B}} \beta_{ik} z_{it} w_{kt} + \sum\limits_{\substack{t\in T, j\in V\\ k\in \mathcal{W}_j}} \lambda_{jt} w_{kt} & & \\ \max\limits_{w,z} & \sum\limits_{\substack{t\in T, j\in V\\ k\in \mathcal{W}_j}} \mu_{jt} w_{kt} & & \\ \forall t \in T \setminus \{t_{\max}\} & \sum\limits_{i\in U} a_i(z_{i,t+1} - z_{it}) & \leq & H_t \\ \forall t \in T \setminus \{t_{\max}\}, i \in U & z_{it} & \leq & z_{i,t+1} \\ \forall t \in T, (i,k) \in B : i \in N & z_{it} & \geq & w_{kt} \\ \forall s \leq t \in T, (i,h) \in S & z_{hs} - z_{it} & \leq & 0 \\ \forall s \leq t \in T \setminus \{0\}, (i,h) \in S & z_{hs} + z_{it} - z_{i,t-1} & \leq & 1 \\ \forall i \in U & z_{i0} & = & 0 \\ \forall j \in V & w_{j0} & = & 0. \end{array}\right\} \tag{17.27}$$

We remark that the form of the objective functions for (17.26) and (17.27) is special: both cases conform to the general case

$$\left.\begin{array}{rl} \max\limits_{y\in Y} & f(y) + g(y) \\ \max\limits_{y\in Y} & g(y) \end{array}\right\}. \tag{17.28}$$

for functions $f, g$ and a set $Y$. We have the following results.

**17.4.1 Proposition**
*The efficient set of* (17.28) *is contained in that of:*

$$\left.\begin{array}{ll} \max\limits_{y\in Y} & f(y) \\ \max\limits_{y\in Y} & g(y) \end{array}\right\}. \tag{17.29}$$

*Proof.* Let $y, y'$ be in the efficient set of (17.28). Then, either $f(y)+g(y) \leq f(y')+g(y')$ $(*)$ and $g(y) \geq g(y')$ $(\dagger)$, or $f(y)+g(y) \geq f(y')+g(y')$ $(**)$ and $g(y) \leq g(y')$ $(\ddagger)$. By rearrangement of $(*)$ we have $f(y)-f(y') \leq g(y')-g(y)$; by $(\dagger)$, $g(y')-g(y) \leq 0$. Therefore, $f(y)-f(y') \leq 0$. By rearrangement of $(**)$ we have $f(y)-f(y') \geq g(y')-g(y)$, which by $(\ddagger)$ is $\geq 0$, hence $f(y)-f(y') \geq 0$. Thus $y, y'$ are in the efficient set of (17.29). □

By Prop. 17.4.1 one could find the efficient sets of:

$$\left.\begin{array}{ll} \max\limits_{u,v} & \sum\limits_{\substack{t\in T\\ j\in V}} |\mathcal{W}_j|\lambda_{jt}(v_{jt}-1) \\ \max\limits_{u,v} & \sum\limits_{\substack{t\in T\\ i\in U}} \phi_i u_{it} \\ & \text{constraints of (17.26)} \end{array}\right\} \tag{17.30}$$

$$\left.\begin{array}{ll} \max\limits_{w,z} & \sum\limits_{\substack{t\in T\\ (i,k)\in B}} \beta_{ik} z_{it} w_{kt} \\ \max\limits_{w,z} & \sum\limits_{\substack{t\in T, j\in V\\ k\in \mathcal{W}_j}} \mu_{jt} w_{kt} \\ & \text{constraints of (17.27)} \end{array}\right\} \tag{17.31}$$

and then simply filter out all the dominated solutions with respect to the objectives of (17.26) and, respectively, (17.27).

Formulations (17.30)-(17.31) are difficult to solve because, in accordance with Lagrangian duality theory, one would also have to minimize with respect to $\lambda, \mu$. In practice, one could employ a "pure decomposition" where $\lambda = \mu = 0$. This reduces (17.30)-(17.31) to the two following single-objective problems:

$$\left.\begin{array}{ll} \max\limits_{u,v} & \sum\limits_{\substack{t\in T\\ i\in U}} \phi_i u_{it} \\ & \text{constraints of (17.26)} \end{array}\right\} \tag{17.32}$$

$$\left.\begin{array}{ll} \max\limits_{w,z} & \sum\limits_{\substack{t\in T\\ (i,k)\in B}} \beta_{ik} z_{it} w_{kt} \\ & \text{constraints of (17.27)} \end{array}\right\} \tag{17.33}$$

This proves the following result:

**17.4.2 Theorem**
*Relaxing* (17.16) *yields a MP with the single objective function* (17.8) + (17.9).

In other words, the constraints (17.16) are the true source of the trade-off nature of the problem.

## 17.5 Computational results

In our previous work [162] we proposed a single objective model of the architecture evolutions problem and showed that it can be solved in a reasonable amount of time with regard to realistically sized instances. We aim to establish if we can solve the MO problem which involves the two objectives (17.8) and (17.9) as good as the sigle objective one which considered only (17.8). We are more interested in evaluating the computational effort required rather than in exactly modelling the preferences of the decision makers. Hence, we adopt a no-preference approach, the $L_p$-metric method, with p=1 and solve:

$$\min_{u,v,w,z} \left| \sum_{\substack{t\in T\\ (i,k)\in B}} \beta_{ik} z_{it} w_{kt} - h_1^* \right| + \left| \sum_{\substack{t\in T\\ i\in U}} \Phi_i u_{it} - h_2^* \right| \tag{17.34}$$

We consider a set of small instances, to be solved to guaranteed optimality, and one of larger instances where the BB algorithm is stopped either at BB termination or after 30 minutes of CPU time (whichever comes first). We use the AMPL modelling environment [154] and the off-the-shelf CPLEX 11 solver [200] running on four 2.4 GHz Intel Xeon CPUs with 8GB RAM. Ordinarily CPLEX's Quadratic Programming (QP) solver requires QPs with Positive Semi-Definite (PSD) quadratic forms only. Although in our case this may not be true, CPLEX can reformulate the problem exactly to the required form because all variables are binary.

We consider the same set of instances both for the single objective form of the problem and the bi-objective form. All instances have been randomly generated from a model that bears some similarity to data coming from an actual service industry. We consider three parameter categories: cardinalities (vertex set), graph density (edge creation probability) and monetary values. Each of the 64 instances in each set corresponds to a triplet (cardinality, edge creation probability, monetary value), each component of which ranges over a set of four elements.

### 17.5.1 CPU time

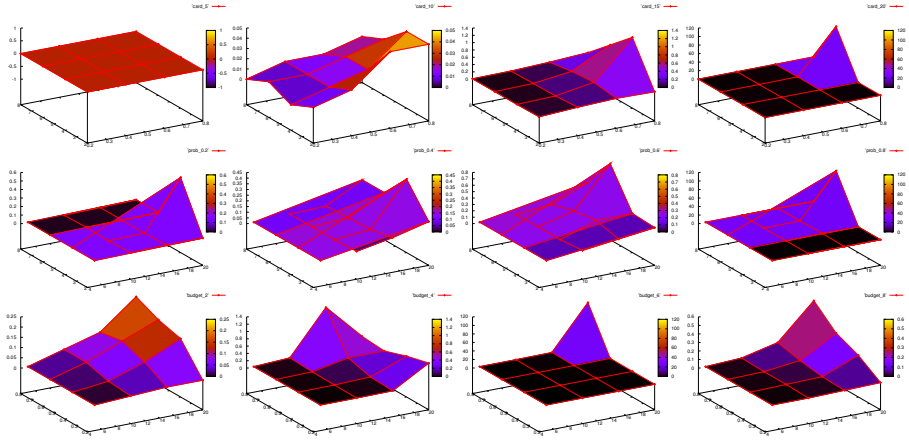


Figure 17.5: Single Objective Model: CPU time when solving small instances.

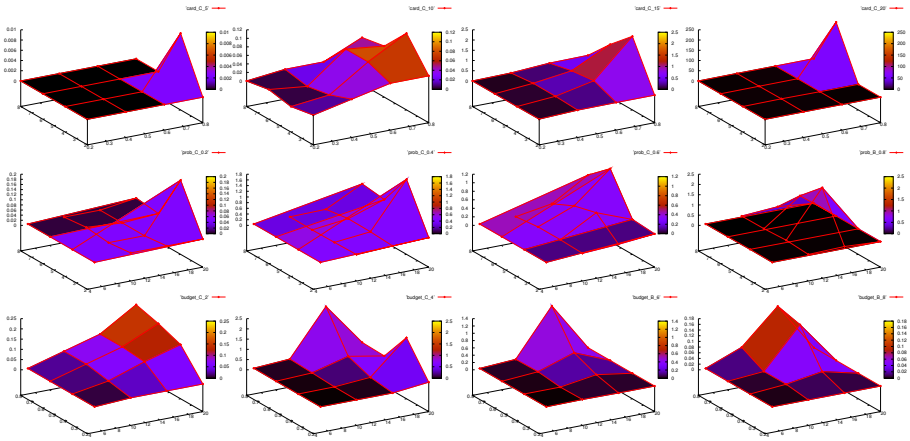


Figure 17.6: Bi-objective Model: CPU time when solving small instances.

In order to observe how CPU time scales when solving to guaranteed optimality, we present 12 plots referring to the small set, grouped by row. We have two groups of plots, the first (Fig. 17.5)

regards the results obtained with the single objective formulation and the second one (Fig. 17.6) the results obtained with the bi-objective formulation. We plot seconds of user CPU time: for each fixed cardinality, in function of edge creation probability and monetary value (Fig. 17.5 and Fig. 17.6, first row); for each fixed edge creation probability, in function of cardinality and monetary value (Fig. 17.5 and Fig. 17.6, second row); for each fixed monetary value, in function of cardinality and edge creation probability (Fig. 17.5 and Fig. 17.6, third row). The largest "small instance" corresponds to the triplet $(20, 0.8, 8)$. The plots show that the proposed methodology can solve a small instance to guaranteed optimality within half an hour; it is also possible to notice that denser graphs and smaller budgets yield more difficult instances. Sudden drops in CPU time might correspond to infeasible instances, which are usually detected quite fast. We notice also that the we can solve both the single objective and the bi-objective formulation. Table 17.1 shows the results of the comparison (with cardinality fixed at 20) and the relative increase of the cpu time needed to solve the new formulation. The effort is considerable higher but still manageable. Infeasibily is detected similarly in both models.

| Edge Probability | Budget | feasible/infeasible | cpu time increment |
|---|---|---|---|
| 0,4 | 2 | infeasible | 0,0 |
| 0,4 | 4 | feasible | 314,3 |
| 0,4 | 6 | feasible | 241,2 |
| 0,4 | 8 | feasible | 257,9 |
| 0,6 | 2 | infeasible | 12,5 |
| 0,6 | 4 | infeasible | 0,0 |
| 0,6 | 6 | feasible | 73,4 |
| 0,6 | 8 | feasible | 135,7 |
| 0,8 | 2 | infeasible | -4,5 |
| 0,8 | 4 | infeasible | 4,0 |
| 0,8 | 6 | feasible | 186,7 |
| 0,8 | 8 | feasible | 137,0 |

Table 17.1: Cpu time increment

### 17.5.2 Optimality Gap

Fig. 17.7 and Fig. 17.8 are also organized by rows, but we plot the *optimality gap* — an approximation ratio — at termination rather than the CPU time, which is in this case limited to 30 minutes. The largest "large instance" corresponds to the triplet $(40, 0.8, 16)$. The optimality gap, expressed in percentage, is defined as $\left(\frac{100|f^* - \bar{f}|}{|f^* + 10^{-10}|}\right)\%$, where $f^*$ is the objective function value of the best feasible solution found within the time limit, and $\bar{f}$ is the tightest overall lower bound. A gap of 0% corresponds to the instance being solved to optimality. The plots show that the proposed methodology is able to solve large instances to a gap of 12.8% within half an hour of CPU time at worst. It can solve and to an average gap of 1.13% both the single and the bi-objective formulation, within an average CPU time of 459s and 538s respectively (about 8 minutes). Tables 17.2 and 17.3 given in Appendix report the details of the comparison.

## 17.6 Conclusion

The *information system architecture evolution management problem*, namely the problem of scheduling the replacement of existing services with new services without discountinuity still need an exact and shared formulation, dispite of its practical importance. The presence of many deciders, as the *Business*

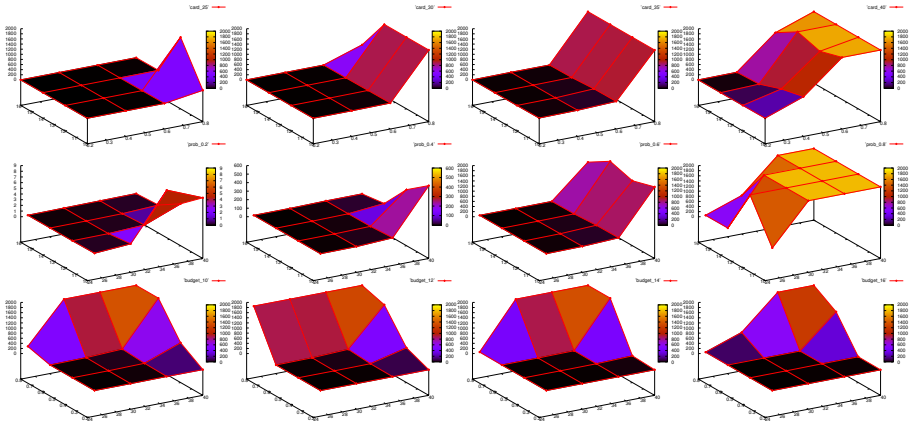


Figure 17.7: Single Objective Model: Optimality gap.

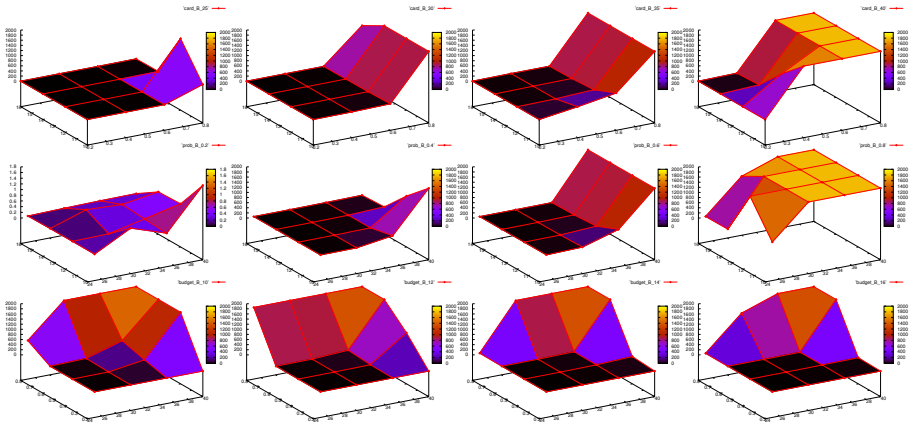


Figure 17.8: Bi-objective Model: Optimality gap.

*department managers*, the *IT project managers* and the *Kill managers* allows different perspectives. These different stakeholders have different needs that the evolution of the system has to satisfy and this causes conflicts between the respective tasks, especially when the scheduling of the activities is tight. The decision makers tipically aim to gain : (1) top business value produced by the new services, (2) the maximum number of new useful modules activated and (3) the maximum number of useless modules deactivated. In most situations, the objectives (1) and (2) are not really conflicting since the activation of new services require new modules, thus Business and IT managers push the activities in the same direction. On the contrary, the objectives (3) is potentially controversial, when there is a lack of time and resources. The activation of new modules and the deactivation of old ones require work. If the amount of workforce is limited (a bound of this kind is really plausible) we have to decide what has to be done first and, eventually, what is not necessary and can be planned for a later period. Business and Project managers on one side and Kill managers on the other side have to compete "to grab" the resouces and have diverging aims. The former can fully attain their tasks on time only forcing the latter to delay theirs and viceversa.

We proposed a model in a previous work which favoured the objectives (1) and (2) and now we refine it taking in account the objective (3) more carefully. We define a Mathematical Programming (MP) formulation and test that we can solve it for realistic instances observing the trend of both CPU time and optimality gap. This model is an evolution of the previous one, so we present a comparison of the results obtained for both of them. We also analyse some theoretical aspects of our formulation in order

to investigate futher developements.

From a wider perspective this work and its predecessor show a sequence of refinements which are an example of the real activity of modelling, which progress step by step towards the best model crossing different level of abstraction. We work at both improving the model and formalizing the practice of progressive refinements to make it a mature engineering method.

## 17.7 Appendix

| Card. | Prob. | Bud. | feas. (1) | cpu (1) | obj. (1) | gap (1) | feas. (2) | cpu (2) | obj. (2) | gap (2) |
|---|---|---|---|---|---|---|---|---|---|---|
| 25 | 0,20 | 10 | yes | 0,09 | 230 | 0,13 | yes | 0,03 | 529 | 0,37 |
| 25 | 0,20 | 12 | yes | 0,03 | 261 | 0,58 | yes | 0,04 | 561 | 0,23 |
| 25 | 0,20 | 14 | yes | 0,02 | 190 | 0,65 | yes | 0,03 | 490 | 0,25 |
| 25 | 0,20 | 16 | yes | 0,02 | 205 | 0,06 | yes | 0,02 | 505 | 0,02 |
| 25 | 0,40 | 10 | yes | 0,12 | 344 | 1,45 | yes | 0,69 | 497 | 0,63 |
| 25 | 0,40 | 12 | yes | 0,17 | 404 | 0,21 | yes | 0,35 | 629 | 0,84 |
| 25 | 0,40 | 14 | yes | 0,04 | 411 | 1,62 | yes | 0,26 | 636 | 1,25 |
| 25 | 0,40 | 16 | yes | 0,1 | 424 | 3,41 | yes | 0,44 | 652 | 0,3 |
| 25 | 0,60 | 10 | yes | 1,33 | 424 | 0,42 | yes | 3,14 | 574 | 0,03 |
| 25 | 0,60 | 12 | yes | 0,46 | 499 | 0,37 | yes | 2,48 | 649 | 0,76 |
| 25 | 0,60 | 14 | yes | 0,22 | 578 | 1,13 | yes | 0,64 | 803 | 0,03 |
| 25 | 0,60 | 16 | yes | 0,32 | 562 | 0,79 | yes | 0,72 | 787 | 0,44 |
| 25 | 0,80 | 10 | yes | 196,76 | 540 | 0,01 | yes | 504,4 | 690 | 0,01 |
| 25 | 0,80 | 12 | yes | 1801,15 | 631 | 0,55 | yes | 1800,55 | 781 | 0,83 |
| 25 | 0,80 | 14 | yes | 5,85 | 734 | 0,02 | yes | 13,4 | 959 | 0,01 |
| 25 | 0,80 | 16 | yes | 2,15 | 745 | 0,08 | yes | 6,33 | 970 | 0,04 |
| 30 | 0,20 | 10 | yes | 0,51 | 304 | 0,05 | yes | 0,69 | 612 | 1,04 |
| 30 | 0,20 | 12 | yes | 0,05 | 344 | 0,59 | yes | 0,05 | 704 | 0,24 |
| 30 | 0,20 | 14 | yes | 0,04 | 303 | 1,05 | yes | 0,44 | 627 | 0,21 |
| 30 | 0,20 | 16 | yes | 0,06 | 315 | 0,53 | yes | 0,05 | 674 | 0,47 |
| 30 | 0,40 | 10 | yes | 2,54 | 477 | 0,09 | yes | 7,34 | 657 | 0,06 |
| 30 | 0,40 | 12 | yes | 0,95 | 532 | 0,14 | yes | 4,36 | 712 | 0,11 |
| 30 | 0,40 | 14 | yes | 0,27 | 609 | 0,49 | yes | 0,76 | 879 | 0,89 |
| 30 | 0,40 | 16 | yes | 0,26 | 560 | 0,16 | yes | 1,94 | 830 | 0,06 |
| 30 | 0,60 | 10 | yes | 13,27 | 635 | 0,01 | yes | 37,76 | 815 | 0,01 |
| 30 | 0,60 | 12 | yes | 10,06 | 634 | 0,02 | yes | 11,43 | 814 | 0,07 |
| 30 | 0,60 | 14 | yes | 3,11 | 734 | 0,06 | yes | 5,81 | 914 | 0,11 |
| 30 | 0,60 | 16 | yes | 0,31 | 824 | 0,19 | yes | 1,2 | 1094 | 0,34 |
| 30 | 0,80 | 10 | yes | 1800,96 | 772 | 3,84 | yes | 1800,46 | 952 | 3,55 |
| 30 | 0,80 | 12 | yes | 1800,96 | 789 | 1,31 | yes | 1800,45 | 969 | 1,44 |
| 30 | 0,80 | 14 | yes | 1800,99 | 837 | 1,73 | yes | 1800,46 | 1017 | 1,78 |
| 30 | 0,80 | 16 | yes | 359,42 | 1099 | 0,01 | yes | 1296,32 | 1369 | 0,01 |

Table 17.2: Optimality gap (a)

| Card. | Prob. | Bud. | feas. (1) | cpu (1) | obj. (1) | gap (1) | feas. (2) | cpu (2) | obj. (2) | gap (2) |
|---|---|---|---|---|---|---|---|---|---|---|
| 35 | 0,20 | 10 | yes | 7,51 | 352 | 0,02 | yes | 0,28 | 764 | 0,02 |
| 35 | 0,20 | 12 | yes | 0,31 | 377 | 0,04 | yes | 0,33 | 716 | 0,07 |
| 35 | 0,20 | 14 | yes | 0,09 | 425 | 0,11 | yes | 0,39 | 769 | 0,11 |
| 35 | 0,20 | 16 | yes | 0,09 | 437 | 0,03 | yes | 0,07 | 857 | 0,32 |
| 35 | 0,40 | 10 | yes | 20,45 | 534 | 0,03 | yes | 212,1 | 641 | 0,02 |
| 35 | 0,40 | 12 | yes | 6,86 | 665 | 0,01 | yes | 20,34 | 875 | 0,01 |
| 35 | 0,40 | 14 | yes | 5,46 | 726 | 0,05 | yes | 15,19 | 936 | 0,04 |
| 35 | 0,40 | 16 | yes | 5,5 | 701 | 0,02 | yes | 10,72 | 914 | 0,07 |
| 35 | 0,60 | 10 | yes | 61,68 | 613 | 0,02 | yes | 459,1 | 718 | 0,01 |
| 35 | 0,60 | 12 | yes | 55,69 | 824 | 0,01 | yes | 87,78 | 1034 | 0,01 |
| 35 | 0,60 | 14 | yes | 12,79 | 816 | 0,02 | yes | 28,15 | 1026 | 0,01 |
| 35 | 0,60 | 16 | yes | 2,53 | 1012 | 0,13 | yes | 8,29 | 1222 | 0,03 |
| 35 | 0,80 | 10 | yes | 1800,83 | 579 | 7,74 | yes | 1800,45 | 684 | 7,32 |
| 35 | 0,80 | 12 | yes | 1800,85 | 978 | 4,97 | yes | 1800,4 | 1188 | 4,5 |
| 35 | 0,80 | 14 | yes | 1800,87 | 969 | 2,45 | yes | 1800,45 | 1179 | 2,33 |
| 35 | 0,80 | 16 | yes | 1800,72 | 1121 | 0,31 | yes | 1800,37 | 1331 | 0,56 |
| 40 | 0,20 | 10 | yes | 6,13 | 463 | 0,01 | yes | 1,72 | 812 | 0,04 |
| 40 | 0,20 | 12 | yes | 2,75 | 453 | 0,01 | yes | 0,62 | 837 | 0,04 |
| 40 | 0,20 | 14 | yes | 0,39 | 449 | 0,03 | yes | 0,59 | 822 | 0,01 |
| 40 | 0,20 | 16 | yes | 0,19 | 479 | 0,16 | yes | 0,08 | 948 | 0,1 |
| 40 | 0,40 | 10 | yes | 511,06 | 588 | 0,01 | yes | 1800,31 | 708 | 1,38 |
| 40 | 0,40 | 12 | yes | 324,62 | 686 | 0,01 | yes | 994,54 | 806 | 0,01 |
| 40 | 0,40 | 14 | yes | 41,64 | 818 | 0,01 | yes | 69,13 | 1058 | 0,01 |
| 40 | 0,40 | 16 | yes | 7,48 | 906 | 0,03 | yes | 25,12 | 1146 | 0,02 |
| 40 | 0,60 | 10 | yes | 1800,88 | 638 | 2,36 | yes | 1800,33 | 758 | 6,72 |
| 40 | 0,60 | 12 | yes | 1273,08 | 753 | 0,01 | yes | 1800,27 | 873 | 2,9 |
| 40 | 0,60 | 14 | yes | 1800,95 | 1061 | 0,73 | yes | 1800,37 | 1301 | 0,66 |
| 40 | 0,60 | 16 | yes | 1222,44 | 1105 | 0,01 | yes | 1800,36 | 1345 | 0,37 |
| 40 | 0,80 | 10 | yes | 1800,72 | 720 | 12,84 | yes | 1800,33 | 840 | 11,49 |
| 40 | 0,80 | 12 | yes | 1800,75 | 807 | 9,18 | yes | 1800,46 | 927 | 8,3 |
| 40 | 0,80 | 14 | yes | 1800,72 | 1340 | 6,08 | yes | 1800,39 | 1580 | 5,48 |
| 40 | 0,80 | 16 | yes | 1800,74 | 1315 | 3,59 | yes | 1800,53 | 1555 | 3,41 |

Table 17.3: Optimality gap (b)

# Chapter 18

# Recent advances on the discretizable molecular distance geometry problem

Carlile Lavor, Leo Liberti, Nelson Maculan, Antonio Mucherino

*Submitted to European Journal of Operational Research*

*Identifying 3D conformation is a crucial step to synthesizing useful proteins; in particular, the conformation of some of the proteins linked to photosynthesis is still largely unknown. Since photosynthesis allows the production of clean energy from light, this line of research is relevant to sustainable energy, although our approach is more general than that. The Molecular Distance Geometry Problem (MDGP) consists in finding an embedding in $\mathbb{R}^3$ of a nonnegatively weighted simple undirected graph with the property that the Euclidean distances between embedded adjacent vertices must be the same as the corresponding edge weights. The Discretizable Molecular Distance Geometry Problem (DMDGP) is a particular subset of the MDGP which can be solved using a discrete search occurring in continuous space; its main application is to find three-dimensional arrangements of proteins using Nuclear Magnetic Resonance (NMR) data. The model provided by the DMDGP is too theoretical for practical exploitation. In the last five years we strove to adapt the DMDGP to be an ever closer model of the actual difficulties posed by the problem of determining protein structures from NMR data, whilst always keeping the discrete search property valid. This survey lists recent developments on DMDGP related research.*

## 18.1 Introduction

The determination of the three-dimensional structure of a given protein is an all-important and formidable problem in biochemistry, mainly because the function of a protein is linked to its structure as well as to its atomic composition [325]. We consider here the subproblem of determining the protein structure with information arising from Nuclear Magnetic Resonance (NMR) data [170].

The output of a Nuclear Magnetic Resonance experiment on a given molecule can be taken to consist of a set of atomic labels (e.g. hydrogen, carbon and so on) and a set of pairs $(\ell, q)$, where $\ell \in \mathbb{N}$ and $q \in \mathbb{R}_+$, such that there are $\ell$ pairs of atoms in the molecule with labels in the given set and mutual distance equal to $q$ [54]. It turns out that NMR data can be manipulated so that it yields a list of pairs $\{u, v\}$ of atoms with a corresponding nonnegative distance $d_{uv}$. Unfortunately this manipulation is rather error-prone, resulting in interval-type errors, so that the exact inter-atomic distances $d_{uv}$ are in fact contained in given intervals $[d^L_{uv}, d^U_{uv}]$ [54]. For practical reasons, NMR experiments are often performed

on hydrogen atoms [54] (although sometimes carbons and nitrogens are also considered). Other known molecular information includes the number and type of atoms in the molecules, all the covalent bonds with corresponding Euclidean distances, all distances between atoms separated by exactly two covalent bonds, and the Euclidean coordinates of at least some of the atoms in the molecule [325].

We assume NMR data can be stored as a nonnegatively interval-weighted simple (i.e. without loops or parallel edges) undirected graph $G = (V, E, d)$ where $V$ represents a subset of atoms of the molecule for which distance measurements can be obtained, $\{u, v\} \in E$ if a distance measurement is present between atoms $u$ and $v$, and $d$ associates an edge $\{u, v\} \in E$ with the respective interval measurement $[d^L_{uv}, d^U_{uv}]$ (since precise distances are also known for certain edges, such as for covalent bonds, some intervals might have $d^L_{uv} = d^U_{uv}$). The main problem is that of finding a set (alternatively, all sets) of Cartesian coordinates for the atoms that are consistent with all the distance information. We shall call this problem the Protein Structure from NMR Data (PSNMR).

Our survey will focus on several variants of the PSNMR. Specifically, we shall consider the cases when: (a) $d$ maps $E$ into nonnegative real numbers (instead of intervals); (b) $V$ is the set of *all* atoms; (c) a particular order on $V$ guarantees the existence of an iterative search for the position of $v \in V$ given the positions of its adjacent predecessors; (d) the Euclidean space used for the embedding has an arbitrary number of dimensions (this is useful for applications other than to molecular structure prediction). Each case gives rise to different theoretical results; we show how we combined them in order to derive a very efficient discrete search in continuous space that addresses the main problem.

This paper is organized as follows: in the rest of Sect. 18.1 we give a very short review of continuous search based methods and illustrate their weaknesses as a motivation to work towards a discrete search. In Sect. 18.2 we introduce our discrete approach. In Sect. 18.3 we generalize the discrete search method to Euclidean spaces of arbitrary dimensions. In Sect. 18.4 we discuss automatic methods to find "good" orders for $V$ guaranteeing the existence of a discrete search method. In Sect. 18.5 we restrict $V$ to only contain hydrogen atoms. Sect. 18.7 presents our implementation to serial and parallel architectures. Sect. 18.8 concludes the paper and discusses future work.

### 18.1.1 Problems solved by continuous methods

Given a simple undirected graph $G = (V, E)$ and a positive integer $K$, an *embedding* of $G$ in $\mathbb{R}^K$ is a function $x : V \to \mathbb{R}^K$. Let $d : E \to \mathbb{R}_+$ be a given edge weight function on $G = (V, E, d)$. An embedding is *valid* for $G$ if

$$\forall \{u, v\} \in E \quad \|x_u - x_v\| = d_{uv}, \tag{18.1}$$

where $\|\cdot\|$ is the Euclidean norm, $x_v = x(v)$ for all $v \in V$ and $d_{uv} = d(\{u, v\})$ for all $\{u, v\} \in E$. For any $U \subseteq V$, an embedding of $G[U]$ (i.e. $(U, \{\{u, v\} \in E \mid u, v \in U\})$, the subgraph of $G$ induced by $U$) is a *partial embedding* of $G$. If $x$ is a partial embedding of $G$ and $y$ is an embedding of $G$ such that $\forall u \in U\ x_u = y_u$ then $y$ is an *extension* of $x$. With a slight abuse of notation, if $v \notin U$ and $y$ is an embedding of $G[U \cup \{v\}]$, we write $y = (x, y_v)$; in this case we also say that the point $y_v$ extends $x$.

The most basic model for the PSNMR problem is the following.

> Molecular Distance Geometry Problem (MDGP). Given a nonnegatively weighted simple undirected graph $G = (V, E, d)$ is there a valid embedding of $G$ in $\mathbb{R}^3$?

This is one of the foremost problems in distance geometry [71]; we shall call its generalization to $\mathbb{R}^K$ (with $K$ being given as part of the input) the Distance Geometry Problem (DGP), and denote the restriction of the DGP to a particular fixed dimension $K$ by DGP$_K$. If $d$ is an interval-valued function, i.e. $d(\{u, v\}) = [d^L_{uv}, d^U_{uv}]$ for all $\{u, v\} \in E$, we obtain a problem which is "closer" to the PSNMR.

> interval Molecular Distance Geometry Problem (*i*MDGP). Given a nonnegatively interval-

weighted simple undirected graph $G = (V, E, d)$, is there an embedding $x : V \to \mathbb{R}^3$ such that:

$$\forall \{u, v\} \in E \quad d^L_{uv} \leq \|x_u - x_v\| \leq d^U_{uv}? \tag{18.2}$$

In this case, an embedding is valid if it satisfies (18.2). Again, we consider the generalization to $\mathbb{R}^K$ and call it the interval Distance Geometry Problem ($i$DGP).

### 18.1.2 Characterization of the solution set

Let $\bar{X} = \{x : V \to \mathbb{R}^K \mid x \text{ satisfies (18.2)}\}$ be the set of all solutions to an $i$DGP instance. Then if $T$ is a translation or rotation of $\mathbb{R}^K$, for all $x \in \bar{X}$ we also have $T(x) \in \bar{X}$. Because there are continuously many such transformations, it follows that $|\bar{X}| = 2^{\aleph_0}$. We define an equivalence relation $\sim$ on $\bar{X}$ such that $x \sim y$ if and only if there is a translation or rotation $T$ such that $y = T(x)$. We then define $X = \bar{X}/\sim$ and identify the equivalence classes of $X$ with one of their representatives $x \in \bar{X}$. We can now consider $X$ as the "interesting" set of solutions of an $i$DGP instance. We remark that $|X|$ is not necessarily infinite. In fact, most of the $i$DGP variants considered in the sequel will have a finite $|X|$.

### 18.1.3 Problem complexity

A reduction from the Subset-Sum problem to the $\text{DGP}_1$ with unit weights was given in [323], showing that $\text{DGP}_1$ is **NP**-complete. For fixed values of $K$, [323] describes a reduction from 3-SAT to $\text{DGP}_1$ with integer weights and a reduction from $\text{DGP}_1$ with integer weights to $\text{DGP}_K$ with integer weights. In the same paper, Saxe also remarked that since YES certificates for the DGP generally involve irrational numbers for $K > 1$, it is not clear whether the DGP belongs to the class **NP** or not. From this it follows that the MDGP is **NP**-hard, and the same holds for $\text{DGP}_K$ for each integer $K > 1$. Considering formal decision problems as sets of instances, it is clear that $\text{DGP}_3 = \text{MDGP} \subset i\text{MDGP} \subset i\text{DGP}$ and $\text{DGP}_K \subset \text{DGP}$ for all $K \in \mathbb{N}$. Again, because singletons are also intervals, $\text{DGP} \subset i\text{DGP}$. Thus, by restriction ([157], Sect. 3.2.1), the DGP, $i$MDGP and $i$DGP are also **NP**-hard.

### 18.1.4 Euclidean distance matrices

A part of the distance geometry literature is concerned with determining whether a given $n \times n$ matrix $D$ is a *Euclidean distance matrix* (EDM), i.e. whether there exist points $p_1, \ldots, p_n$ of a finite-dimensional Euclidean space such that $d_{ij} = \|p_i - p_j\|^2$ for all $i, j \leq n$. This is called the EDM Problem (EDMP). The EDM Completion Problem (EDMCP) provides a partial matrix $D$ (i.e. a matrix with some unspecified entries) and asks whether there exists a EDM $D'$ such that $D_{ij} = D'_{ij}$ for all $(i, j)$ where $D$ is defined (see e.g. [235]).

Formally, the EDMCP differs from the DGP in that the dimension $K$ of the embedding Euclidean space is not provided as part of the input. Since the output YES certificate is the embedding itself, and writing out the embedding requires knowing $K$, the EDM and EDMCP implicitly make $K$ part of the output. For this reason no inclusion relation can be established between the EDMCP and the DGP variants presented above; in fact, the complexity status of the EDMCP is unknown, although some polynomial cases exist [235]. Since the EDMCP can be formulated as a Semidefinite Programming (SDP) problem, it can be solved to arbitrary precision in polynomial time [8].

### 18.1.5 Limitations of continuous methods

The problems listed above are naturally cast as nonlinear systems of equations and inequalities, and can therefore be reformulated to minimizing an objective function consisting of a sum of error terms, which is a Global Optimization (GO) problem. Some continuous methods for solving such problems are surveyed in [239, 258]. These methods often exhibit the following disadvantages.

- *Reliability*. All computations are floating-point; this yields inaccurate solutions. Moreover, it is well known that floating point errors often accumulate, which in the long run may invalidate the solution.
- *Efficiency*. GO methods often involve locally solving a (nonconvex) Nonlinear Programming (NLP) subproblem; local NLP solvers are complex pieces of software which may take a long time to converge.
- *Completeness*. To the best of our knowledge, there is no continuous method which is able to compute all solutions of an *i*DGP instance; and in fact most continuous methods are actually designed to compute at most one solution.

Of course these disadvantages are due to a trade-off against generality. In the rest of this paper we shall present mixed combinatorial methods for solving *subclasses* of the *i*DGP. It is this restriction that allows our methods to be more reliable, efficient and complete than continuous methods. Moreover, the subclasses for which our methods work are a good model for solving the *i*MDGP on proteins, which are in fact the main motivation for the PSNMR.

## 18.2 The Discretizable Molecular Distance Geometry Problem

Although the DGP implicitly requires a search in continuous space, if an appropriate order is given on $V$, we can show that the search space has a finite number of valid embeddings, up to translations and rotations. For an order $<$ on $V$ and for each $v \in V$, let $\rho(v) = |\{u \in V \mid u \leq v\}|$ be the *rank* of $v$ in $V$ with respect to $<$. Since the rank defines a bijection between $V$ and $\{1, \ldots, |V|\}$, we can identify $v$ with its rank and extend arithmetic notation to $V$ so that for some appropriate $i \in \mathbb{Z}$, $v + i$ denotes the vertex $u \in V$ with $\rho(u) = \rho(v) + i$.

We now outline an iterative algorithm for solving the DMDGP, a subset of the MDGP which will be defined below. We assume that an order is given on $V$. Suppose we want to embed a vertex $v \in V$ of rank greater than three in $\mathbb{R}^3$, and suppose also that: (a) we already know a valid embedding for all vertices preceding $v$; (b) the edges $\{v-3, v\}$, $\{v-2, v\}$, $\{v-1, v\}$ are in $E$. This means that the embedding of $v$, denoted by $x_v$, belongs to the three spheres centered at $x_{v-3}, x_{v-2}, x_{v-1}$ with respective radii $d_{v-3,v}, d_{v-2,v}, d_{v-1,v}$. The intersection of three spheres in $\mathbb{R}^3$ can either be empty, or consist of exactly one point, or of exactly two points (see Fig. 18.1), or of uncountably many points [104] (see Fig. 18.2). Because we assume all vertices preceding $v$ are already embedded prior to $v$, we know all their mutual distances. In particular, we know $d_{v-3,v-1}, d_{v-3,v-2}, d_{v-2,v-1}$. As long as the strict triangular inequality $d_{v-3,v-1} < d_{v-3,v-2} + d_{v-2,v-1}$ holds, then the intersection can only have either one or two points, depending on whether the discriminant of a certain quadratic polynomial in $x_v$ is zero or nonzero [104]: we call this the *finite sphere intersection property*. Because this discriminant can in general take any value in $\mathbb{R}_+$, and a singleton set has Lebesgue measure zero in $\mathbb{R}_+$, the sphere intersection has one point with probability 0 and two points with probability 1. We remark that the strict triangular inequality condition can only be checked once the predecessors of $v$ have been embedded; this prevents us from recognizing aprioristically whether an MDGP instance conforms to this condition or not. We address this limitation by requiring that all 4-cliques of consecutive vertices are subgraphs of $G$. Thus, each 3-(sub)clique $\mathbf{K}_3^v = \{v-3, v-2, v-1\}$ is used to verify the strict triangular inequality, and the edges from $\mathbf{K}_3^v$ to $v$ guarantee the finite sphere intersection property. If we

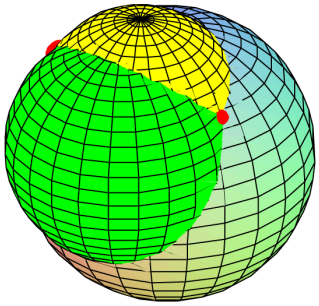


Figure 18.1: Three spheres intersect in exactly two points.

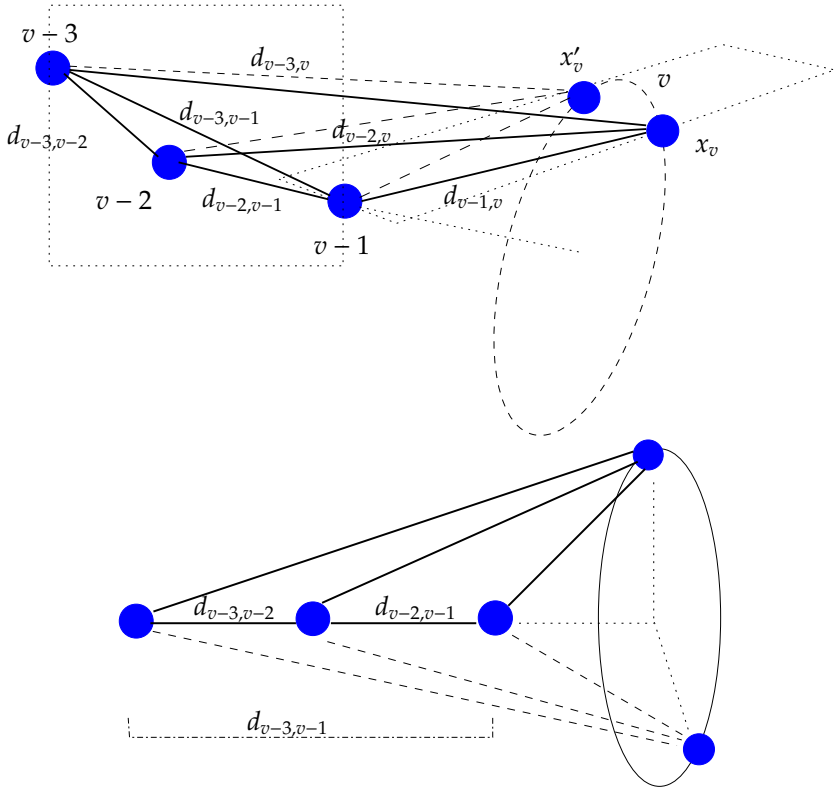



Figure 18.2: Locus of the intersection of three spheres: exactly two points (above) with $d_{v-3,v-1} < d_{v-3,v-2} + d_{v-2,v-1}$ and uncountably many (below) with $d_{v-3,v-1} = d_{v-3,v-2} + d_{v-2,v-1}$.

proceed by embedding vertices iteratively this way we end up with a tree of possibilities where each embedded vertex gives rise to either one or two new positions for the embedding of the next vertex in the order. Since the first vertex triplet has only one possible embedding up to translations and rotations (because $E$ contains a clique on the first four vertices), $|X|$ is finite [238, 240]. Moreover, with probability 1, it is a binary tree from the fourth level downwards.

A number of existing works exploits the finite sphere intersection property, but considering *four* (rather than three, as in our case) spheres [127, 128, 139, 368, 331, 114]; in general, the non-empty intersection of four spheres in $\mathbb{R}^3$ contains exactly *one* point: this follows because the system $\forall j \in$

$\{1,2,3,4\}$ $\|x_{v-j} - x_v\|^2 = d^2_{v-j,v}$ can be reduced to a square $3 \times 3$ linear system which is nonsingular under simple geometric regularity conditions. This ensures that the worst-case running time of an iterative algorithm based on this idea is $O(|V|)$. In [127] $G$ is assumed to be a clique. In [128] this requirement is weakened: the so-called *geometric build-up algorithm* can only find a valid embedding if for the current vertex one can find at least four previously embedded adjacent vertices; depending on the instance, however, the algorithm in [128] may fail to find a valid embedding even if one exists. In [331] the geometric build-up algorithm is modified to deal with some restricted types of measurement errors in the data. In [139] the finite sphere intersection property is introduced in the framework of wireless sensor networks.

Naturally, requiring known distances to four previously embedded adjacent vertices limits the extent of the iterative embedding algorithm to instances with relatively dense graphs. Because distances are usually hard to obtain (this is true for both molecules and sensor networks), an effort should be made in order to weaken this requirement. Although similar concepts were already known in rigidity [318], the first work providing an iterative discrete search algorithm for the MDGP that only requires *three* (rather than four) previously embedded adjacent vertices is [238, 240]. Other methods based on this weaker assumption are given in [256, 93, 369]. The following defines a subclass of MDGP instances conforming to these weaker assumptions [238, 240].

> Discretizable Molecular Distance Geometry Problem (DMDGP). Given a nonnegatively weighted simple undirected graph $G = (V, E, d)$, an order $<$ on $V$ and a mapping $x' : \{1,2,3\} \to \mathbb{R}^3$ such that:
>
> 1. $x'$ is a valid embedding of $G[\{1,2,3\}]$ (Start)
> 2. $G$ contains all 4-cliques of $<$-consecutive vertices as induced subgraphs (Discretization)
> 3. $\forall v \in V$ of rank greater than 3, $d_{v-3,v-1} < d_{v-3,v-2} + d_{v-2,v-1}$ (Strict Triangular Inequalities),
>
> is there a valid embedding $x$ of $G$ in $\mathbb{R}^3$ extending $x'$?

We remark that the formal definition of the DMDGP introduces an order on $V$ as an essential part of the input data; this marks a fundamental difference between [238, 240] and [256, 93, 369]. We shall discuss this further in Sect. 18.4.

### 18.2.1 Problem complexity

While it is clear that DMDGP $\subset$ MDGP, the DMDGP does not include any of the **NP**-hard classes described in Sect. 18.1.3, so restriction cannot be used to establish its **NP**-hardness. An explicit reduction from the Subset-Sum problem to the DMDGP was, however, provided in [238, 240].

### 18.2.2 Mathematical programming formulation

Given a nonnegatively weighted simple undirected graph $G = (V, E, d)$ where $V = \{1, \ldots, n\}$, a valid partial embedding $x' : \{1,2,3\} \to \mathbb{R}^3$ and a map $c : V \setminus \{1,2,3\} :\to [-1,1]$, consider the problem $\mathsf{C}$ of determining whether there exists a valid embedding $x$ of $G$ extending $x'$ and such that the torsion angles $\phi_i$, determined by each quadruplet $(x_{i-3}, x_{i-2}, x_{i-1}, x_i)$ (see Fig. 18.3), satisfy $\cos(\phi_i) = c(i)$ (denoted by $c_i$) for all $i > 3$. Because each $c_i$ can be computed from $d$ in constant time using formula (2.15) in [185], there is a trivial reduction from DMDGP to $\mathsf{C}$. Conversely, if one is given precise values for the torsion angle cosines, then every quadruplet $(x_{i-3}, x_{i-2}, x_{i-1}, x_i)$ must be a rigid framework (for $i > 3$), which implies both Discretization and Strict Triangular Inequalities in the DMDGP definition and shows $\mathsf{C} \subseteq$ DMDGP. Thus, $\mathsf{C}$ is the same set of instances as DMDGP.

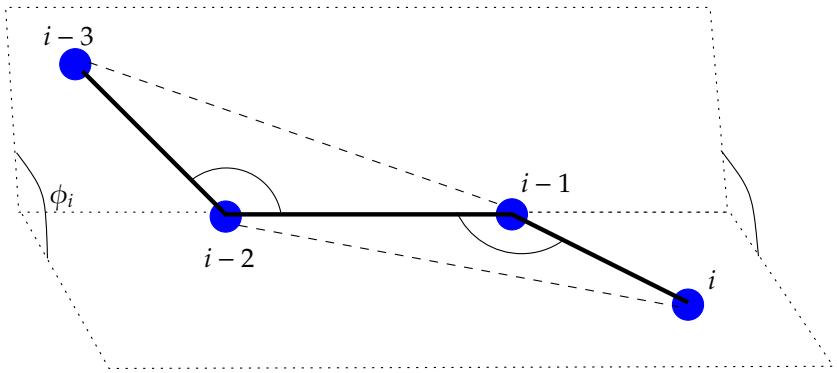



Figure 18.3: The torsion angle $\phi_i$.

We let $\alpha : V \setminus \{1,2\} \to \mathbb{R}^3$ compute the normal vector to the plane defined by three consecutive atoms:

$$\forall i \geq 3 \quad \alpha_i = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ x_{i-2,1} - x_{i-1,1} & x_{i-2,2} - x_{i-1,2} & x_{i-2,3} - x_{i-1,3} \\ x_{i,1} - x_{i-1,1} & x_{i,2} - x_{i-1,2} & x_{i,3} - x_{i-1,3} \end{vmatrix} \tag{18.3}$$

$$= \begin{pmatrix} (x_{i-2,2} - x_{i-1,2})(x_{i,3} - x_{i-1,3}) - (x_{i-2,3} - x_{i-1,3})(x_{i,2} - x_{i-1,2}) \\ (x_{i-2,1} - x_{i-1,1})(x_{i,3} - x_{i-1,3}) - (x_{i-2,3} - x_{i-1,3})(x_{i,1} - x_{i-1,1}) \\ (x_{i-2,1} - x_{i-1,1})(x_{i,2} - x_{i-1,2}) - (x_{i-2,2} - x_{i-1,2})(x_{i,1} - x_{i-1,1}) \end{pmatrix}, \tag{18.4}$$

so that $\alpha_i = \alpha_i(x)$ is a function of $x$, which is itself considered a matrix with entries $x_{ik}$. Now, for every $i > 3$ the cosine of the torsion angle $\phi_i$ is given by the scalar product of the normal vectors $\alpha_{i-1}$ and $\alpha_i$:

$$\forall i > 3 \quad \alpha_{i-1}(x) \cdot \alpha_i(x) = \cos \phi_i.$$

Thus, the following provides a mathematical programming formulation for the DMDGP:

$$\left.\begin{array}{rrcl} \min_x & \sum\limits_{\{i,j\} \in E} (\|x_i - x_j\|^2 - d_{ij}^2)^2 & & \\ \text{s.t.} & \forall i > 3 \quad \alpha_{i-1}(x) \cdot \alpha_i(x) & = & c_i. \end{array}\right\} \tag{18.5}$$

### 18.2.3 Branch-and-Prune framework

We describe the Branch-and-Prune (BP) algorithm for solving the DMDGP [238, 240, 256]. The version given here is recursive (for clarity); it is also parameterized so that its variants, described in the rest of this paper, can be presented as configurations or simple modifications of Alg. 10. We recall that given $G = (V, E)$ and $U \subseteq V$, $G[U]$ denotes the subgraph of $G$ induced by $U$. For $v \in V$, $N(v) = \{u \in V \mid \{u, v\} \in E\}$ is set of vertices adjacent to $v$. We denote by $S^{K-1}(y, r)$ the sphere in $\mathbb{R}^K$ (where $K = 3$) centered at $y$ with radius $r$.

The BranchAndPrune call has five arguments: the weighted simple undirected graph $G = (V, E, d)$ given as part of the DMDGP instance, a current vertex $v$ being embedded, a subset $U \subseteq N(v)$ with $|U| = K$ (where $K$ is the dimension of the embedding space), a valid embedding $x'$ of a subgraph of $G$ containing $G[U]$, and the set $X$ of valid embeddings of $G$ currently found. The recursion starts with the call BranchAndPrune$(G, 4, \{1, 2, 3\}, y, \emptyset)$ where $y$ is the valid embedding of $\{1, 2, 3\}$ given as part of the DMDGP instance.

The BP algorithm shown in Alg. 10 builds a binary search tree whose nodes at level $v$ represent possible spatial positions $p$ for the vertex $v$. Whenever the test in Step 5 for validity of an embedding fails, the branch of $p$ is pruned; pruning techniques are discussed in Sect. 18.2.3.

**Algorithm 10**: The Branch-and-Prune algorithmic framework.

```
1: BranchAndPrune(G, v, U, x', X):
2: Let P be the intersection of the K spheres S^{K-1}(x'_u, d_uv) for u ∈ U
3: for p ∈ P do
4:    Extend the current embedding to x = (x', p)
5:    if x is a valid embedding of G[{1, ..., v}] then
6:       if (v is the last vertex) then
7:          Append x to X
8:       else
9:          Let U' = (U \ {min U}) ∪ {v}
10:         BranchAndPrune(G, v + 1, U', x, X)
11:      end if
12:   end if
13: end for
```

**18.2.1 Theorem ([238, 240])**
*At termination of the BP shown in Alg. 10, X contains all valid embeddings of G extending x'.*

#### Completeness

The BP algorithm generates a search tree. For each leaf node of this tree, the unique path to the root node encodes an embedding of $G$. By Thm. 18.2.1, the unique paths from each leaf node at level $|V|$ encode all valid embeddings of $G$ extending $x'$. We remark that the BP can be stopped after the first valid embedding has been found when just one solution of the DMDGP is needed. It can also be allowed to proceed until all valid embeddings have been identified. This makes the BP algorithm complete, in the sense of Sect. 18.1.5.

#### Algorithmic complexity

In the worst case, when no pruning occurs, $|P| = 2$ at each iteration, which means that the search tree is a full binary tree. This makes the BP worst-case complexity exponential in $|V|$.

#### Performance

In order to asses the empirical behaviour of the BP algorithm we measure its efficiency in terms of seconds of user CPU time, and its reliability in terms of the Largest Distance Error (LDE):

$$\frac{1}{|E|} \sum_{\{u,v\} \in E} \frac{|\|x_u - x_v\| - d_{uv}|}{d_{uv}}. \tag{18.6}$$

The computational results shown in [238, 240] are markedly different from most continuous approaches: they scale up with instance size considerably better both in terms of CPU time and reliability. On the `1epw` PDB [55] instance, for example, which has 3861 atoms and 35028 distances, the BP took 0.25s to find all solutions, and yielded an LDE of $4 \times 10^{-12}$. By comparison, DGSOL [282] took 2038s and produced an embedding with an LDE value around 0.5 (we remark that LDE values greater than one usually denote a wrong protein structure). This instance is not an isolated case: the BP consistently outperforms all continuous approaches we have tested [282, 237, 257].

In very recent work [261] we argue that on average protein instances the BP search tree has bounded width, thus yielding a polynomial-time algorithm; this makes the BP efficient in the sense mentioned in

Sect. 18.1.5.

As for the reliability, our implementation employs several devices in order to limit the propagation of floating point errors given by the computation of $P$, from the choice of appropriate vertex orders minimizing the range of values taken by the entries in the distance matrix (Sect. 18.4) to the exploitation of repeated vertices (Sect. 18.5), for which the zero distance between a vertex and its repetition is used as verification device for the embedding of nearby vertices. This makes the BP reliable in the sense mentioned in Sect. 18.1.5.

**Pruning the BP search tree**

In this section we use the notation of Alg. 10. Pruning out infeasible branches of the BP search tree reduces the CPU time taken by the BP algorithm. Consider the point $p \in P \subseteq \mathbb{R}^K$ embedding vertex $v$ (line 2 in Alg. 10): if $p$ extends $x'$ to a valid embedding $x = (x', p)$ of $G[\{1, \ldots, v\}]$ then $p$ is *feasible*, otherwise it is *infeasible*. In the latter case, the whole sub-tree rooted at $p$ can be pruned.

The most natural pruning test at the iteration when the BP places vertex $v$ is to consider the vertex subset $\bar{U} = \{u \in V \mid u < v \wedge u \in N(v) \wedge u \notin U\}$. Vertices in $\bar{U}$ provide distances to $v$, they will already have been embedded in previous iterations, and have not been used to compute $P$: thus their positions $x_u$ can be matched against $x_v$ to check for consistency of $x_v$. If $\|x_u - x_v\| \neq d_{uv}$ then $(x', x_v)$ is not a valid embedding of $G$ and the node encoding $x_v$ can be pruned. In practical implementation, the pruning condition translates to $|\|x_u - x_v\| - d_{uv}| < \varepsilon$, for a constant tolerance $\varepsilon > 0$ [238, 240, 256]. We shall call this pruning device *Direct Distance Feasibility* (DDF).

Another pruning device that can be used during the discrete search is based on the point-to-point Dijkstra shortest-path searches on Euclidean graphs [242]. Consider the vertices $u, v, w$ with $u < v < w$ such that $\{u, w\} \in E$, i.e. the distance $d_{uw}$ is known. Suppose that a position for the vertex $u$ is already available, and that the feasibility of the node $x_v$ needs to be verified. Let $D(v, w)$ be an upper bound to the distance $\|x_v - x_w\|$ for all possible valid embeddings. Then, if $\|x_u - x_v\| > d_{uw} + D(v, w)$ holds, the node $x_v$ can be pruned [242] because the triangular inequality is negated. A valid upper bound $D(v, w)$ can be computed by finding the shortest path between the vertex $v$ and the vertex $w$ in $G$. We call this pruning device *Dijkstra Shortest Path* (DSP).

Computational experiments showed that the DSP is more efficient than the DDF in detecting infeasible embeddings, but it is also more computationally expensive. From a worst-case complexity point of view, the DDF is $O(1)$, whereas in a naive implementation the DSP is $O(|V|^2)$; of course, all shortest paths in $G$ can be computed as a pre-processing step to the BP in $O(|V|^3)$, so that the DSP can also be reduced to $O(1)$; but since the BP is very fast in practice, the CPU time for the pre-processing is often remarkedly noticeable.

Other pruning devices could be conceived in the application of the DMDGP to protein structure determination using NMR data. As an example, we can model atoms (i.e., vertices) using their van der Waals radii [325]: if the two atoms are not bound, they should not be embedded at points with shorter distance than the threshold given by van der Waals radii. Naturally, such thresholds depend on the kind of atoms involved. Moreover, when considering DMDGPs restricted to *backbone* atoms only (the part of the protein formed by the group of atoms which each amino acid has in common, excluding the side chains), an auxiliary problem could be solved during the search. Every time a $C_\alpha$ carbon is placed, the conformation of the side chain attached to the carbon could be found by solving a Side Chain Placement Problem (SCPP) [322]. If such a problem has no solutions, then the atomic position for the $C_\alpha$ carbon is deemed infeasible.

### 18.2.4 Cardinality of the solution set

It was empirically observed that for most DMDGP instances, BP always finds a number of solution that is a power of two [256]. Counterexamples to this conjecture are given in Lemma 5.1 in [238, 240] and in Sect. 6 in [260]. It was shown in [260] that, for the DMDGP, $|X|$ is a power of two with probability 1.

In this section we give a summary of the argument.

Since the DMDGP definition requires $G$ to have at least those edges used to satisfy the Discretization axiom, we partition $E$ into the sets $E_D = \{\{u, v\} \mid |\rho(v) - \rho(u)| \leq K\}$ and $E_P = E \setminus E_D$. With a slight abuse of notation we call $E_D$ the *discretization distances* and $E_P$ the *pruning distances*. An MDGP instance with all discretization distances and which satisfies Strict Triangular Inequalities is a DMDGP instance. Pruning distances are used to reduce the BP search space by pruning its tree. In practice, pruning distances might make the set $P$ in Alg. 10 have cardinality 0 or 1 instead of 2. Let $G_D = (V, E_D, d)$ and $X_D$ be the set of valid embeddings of $G_D$; since $G_D$ has no pruning distances, the BP search tree for $G_D$ is a full binary tree and $|X_D| = 2^{|V|-3}$.

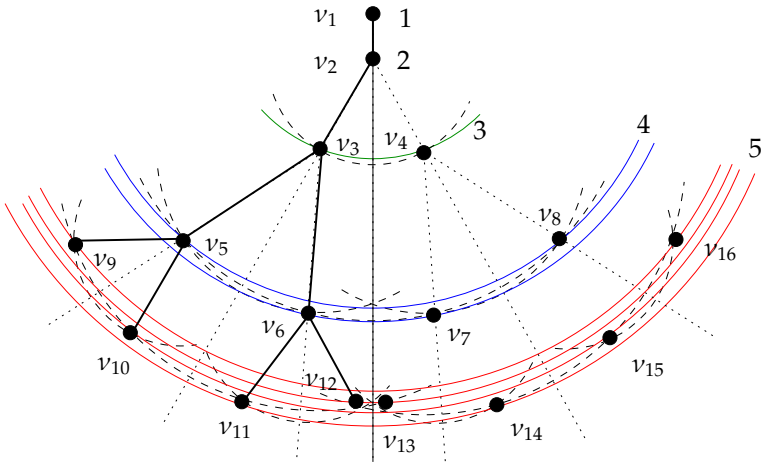



Figure 18.4: An example in 2D; a pruning distance $\{1, 4\}$ prunes either $v_6, v_7$ or $v_5, v_8$.

The discretization distances arrange the embeddings so that, at level $\ell$, there are $2^{\ell-3}$ possible valid embeddings $x_v$ for the vertex $v$ with rank $\ell$. Furthermore, when $P = \{x_v, x'_v\}$, because the discretization distances to $v$ only involve the three immediate predecessors of $v$, we have that $x'_v = R_x^v(x_v)$ [260], the reflection of $x_v$ w.r.t. the plane through $x_{v-3}, x_{v-2}, x_{v-1}$. This also implies that the partial embeddings encoded in two BP subtrees rooted at reflected nodes $v, v'$ are reflections of each other [260]. In other words, for any BP subtree rooted at a node at level $u$, the discretization distances ensure that there are $2^{v-u-3}$ possible BP tree nodes at level $v > u + 3$ and that the corresponding positions have a reflection symmetry through the plane defined by the embeddings of vertices $u, u + 1, u + 2$. This is what is shown in the theorem below.

**18.2.2 Theorem ([260])**
*With probability 1, for each $v > 3$ and $u < v - 3$ there is a finite set $H^{uv}$ of nonnegative real numbers with cardinality $2^{v-u-3}$ such that for each $x \in X$ we have $\|x_v - x_u\| \in H^{uv}$. Furthermore, $\|x_v - x_u\| = \|R_x^{u+3}(x_v) - x_u\|$ and $\forall x' \in X\ (x'_v \notin \{x_v, R_x^{u+3}(x_v)\} \rightarrow \|x_v - x_u\| \neq \|x'_v - x_u\|)$.*

*Proof sketch.* The argument is shown graphically for embeddings in $\mathbb{R}^2$ (replace 3 with 2 in the theorem statement above) in Fig. 18.4; the circles mark distances to vertex 1. □

We also prove (Thm. 4.8 in [260]) that the result above holds even in the case where pruning distances are present. This allows us to show that the number of BP tree nodes at level $|V|$ is a power of two. We

employ a "probability 1" argument which we sketch below.

**18.2.3 Theorem ([260])**
*With probability 1, $|X|$ is a power of two.*

*Proof sketch.* Let $n = |V|$. For all $v > 3$ and a valid embedding $x$ we define a partial reflection operation $g_v(x) = (x_1, \ldots, x_{v-1}, R^v_x(x_v), \ldots, R^v_x(x_n))$: this acts like the identity on the first $v-1$ components of $x$ and as the reflection through the plane defined by $x_{v-1}, x_{v-2}, x_{v-3}$ on the remaining components. With probability 1, for $v > 3$ the $g_v$'s are idempotent injective maps $X_D \to X_D$ which commute. Thus, the set $\mathcal{G}_D = \langle g_v \mid v > 3 \rangle$, generated by all compositions of the $g_v$ operators, is an abelian group (called the *discretization group*) acting on $X_D$ which preserves all discretization distances; it is isomorphic to $C_2^{n-3}$ (the direct product of $n-3$ copies of the cyclic group of order 2). We now consider the subgroup $\mathcal{G}_P \leq \mathcal{G}_D$ (called the *pruned group*) of those partial reflections that also fix pruning distances. Since all subgroups of $C_2^{n-3}$ have cardinality power of two, the pruned group also has this property. Furthermore, its action on $X$ is transitive: thus there is only one orbit, and for all $x \in X$ we have $\mathcal{G}_P x = X$. This shows that $X$ also has cardinality power of two. □

### 18.2.5 Overcoming practical limitations

Computational experiments (see for example [238, 240, 256, 242]) showed that the BP algorithm, when employing the pruning device DDF only, is very efficient in finding the whole set of solutions for DMDGPs. In at most a few seconds of user CPU time on a standard computer all the possible valid embeddings for $G$ can be identified. The DMDGP, however, is an inaccurate model of the PSNMR, which is our main target application.

#### Interval distances

NMR experiments cannot provide exact distances, but only a lower and an upper bound to these distances. As a consequence, for each distance, an interval is generally available in which the actual distance value is contained. This makes the discretization process much more complex. While the pruning device DDF, for example, can be trivially adapted for interval data [283], the generation of the binary tree of solutions may require the computation of the intersection of three spherical shells [286]. In other words, interval distances cannot natively be used to satisfy the Discretization axiom, but they can be used effectively to prune the BP search tree.

Suppose we need to find the possible positions for the vertex $v$. If $d_{v-3,v}$, $d_{v-2,v}$ and $d_{v-1,v}$ are represented by the intervals $[d^L_{v-3,v}, d^U_{v-3,v}]$, $[d^L_{v-2,v}, d^U_{v-2,v}]$ and $[d^L_{v-1,v}, d^U_{v-1,v}]$, three spherical shells can be defined, which are centered in $x_{v-3}$, $x_{v-2}$ and $x_{v-1}$, have inner radii $d^L_{v-3,v}$, $d^L_{v-2,v}$ and $d^L_{v-1,v}$, and outer radii $d^U_{v-3,v}$, $d^U_{v-2,v}$ and $d^U_{v-1,v}$, respectively. Algorithms that allow to compute a finitary representation of arbitrary spherical shell intersections in function of distance intervals are, to the best of our knowledge, hereto unknown. In [241] we propose a strategy to deal with this problem.

#### Distances between hydrogens

Another important issue is related to the enforcement of the Discretization requirement in DMDGP instances arising from proteins. Discretization requires the availability of a certain number of distances, whereas NMR experiments can usually only estimate short range distances (no larger than 4Å or 5Å, depending on the NMR machinery). Moreover, generally, only distances between hydrogen atoms are available from NMR experiments [325].

We address this problem from two points of view. In Sect. 18.4 we describe an automatic method to find best vertex orders to satisfy Discretization. A second strategy, addressing the limitation in hydrogen-related distances posed by the NMR, is discussed in Sect. 18.5: a hand-crafted vertex order satisfying Discretization is defined for the hydrogen atoms of the protein backbones, which are placed first; the other backbone atoms (mainly carbons and nitrogens) are placed in a second stage using an auxiliary DMDGP instance.

This, however, disregards much of the distance information obtained from chemical bonds. Since these distances are known precisely [325], they are likely to be useful in order to satisfy Discretization. In Sect. 18.6, we describe a third strategy, which consists in reordering the atoms in such a way that a maximum number of precise distances are exploited to satisfy Discretization, leaving only a minimum of (arbitrarily discretized) interval distances to the same purpose. The bulk of the (non-discretized) interval distances are relegated to DDF pruning.

Chronologically, we addressed most of the limitations posed by real protein instance incrementally, starting from the basic DMDGP with exact distances and going through a sequence of problems where the crucial Discretization requirement was maintained through clever modifications of the vertex order. We believe that the BP variant discussed in Sect. 18.6 solves instances which provide a realistic model of the PSNMR.

## 18.3 The Discretizable Distance Geometry Problem

Although our driving application is to embed proteins in 3D, other applications of graph embedding (wireless sensor networks, graph drawing) require embeddings in Euclidean spaces of varying dimensions. Since the finite sphere intersection property also holds in Euclidean spaces of arbitrary dimensions, we discuss two vartiants of the DMDGP requiring embeddings in $\mathbb{R}^K$.

The DGP, which generalizes the MDGP to a Euclidean space of arbitrary dimension $K$, asks for a valid embedding of $G$ in $\mathbb{R}^K$. The generalization of the DMDGP to $\mathbb{R}^K$ replaces triplets of immediate adjacent predecessors with $K$-uples of adjacent (but not necessarily immediate) predecessors. Furthermore, strict triangle inequalities are replaced with strict simplex inequalities [71]. Strict triangle inequalities ensure that the three predecessors in the DMDGP statement are not collinear; in other words, they ensure that the 2-simplex defined by the predecessors has nonzero volume. Strict simplex inequalities generalize this idea. For a set $U = \{x_i \in \mathbb{R}^{K-1} \mid i \leq K\}$ of points in $\mathbb{R}^{K-1}$, let $D$ be the symmetric matrix whose $(i, j)$-th component is $\|x_i - x_j\|^2$ for all $i, j \leq K$ and let $D'$ be $D$ bordered by a left $(0, 1, \ldots, 1)^\top$ column and a top $(0, 1, \ldots, 1)$ row (both of size $K + 1$). Then the Cayley-Menger formula [71, 207] states that the volume $\Delta_{K-1}(U)$ of the $(K-1)$-simplex on $U$ is given by

$$\Delta_{K-1}(U) = \sqrt{\frac{(-1)^K}{2^{K-1}((K-1)!)^2}|D'|}. \tag{18.7}$$

The strict simplex inequalities are given by $\Delta_{K-1}(U) > 0$. For $K = 3$, these reduce to strict triangle inequalities. We remark that only the distances of the simplex edges are necessary to compute $\Delta_{K-1}(U)$, rather than the actual points in $U$; the needed information can be encoded as a complete graph $\mathbf{K}_K$ on $K$ vertices with edge weights as the distances. This implies that $\Delta_{K-1}(U)$ is well defined also if $U$ is a set of *vertices* of $V$ (instead of points in $\mathbb{R}^{K-1}$) as long as $G[U] = \mathbf{K}_K$. We also let $[K] = \{1, \ldots, K\}$.

For each $v > K$ we denote the set of $K$ consecutive predecessors of $v$ by $\gamma_K(v)$.

Generalized Discretizable Molecular Distance Geometry Problem (GDMDGP). Given a positive integer $K$, a nonnegatively weighted simple undirected graph $G = (V, E, d)$, an order $<$ on $V$ and a mapping $x' : [K] \to \mathbb{R}^K$ such that:

1. $x'$ is a partial embedding of $G[[K]]$ (Start)

2. $\forall v \in V \setminus [K]$ ($|\gamma_K(v) \cup \{v\}| = K + 1$) (Discretization)
3. $\forall v \in V \setminus [K]$ ($\Delta_{K-1}(\gamma_K(v)) > 0$) (Strict Simplex Inequalities),

is there a valid embedding $x$ of $G$ in $\mathbb{R}^K$ extending $x'$?

We denote GDMDGP instances where $K$ is a fixed constant by GDMDGP$_K$. The results of Sect. 18.2.4 apply to the GDMDGP.

### 18.3.1 From immediate to adjacent predecessors

For intersections of $K$ appropriately defined spheres to yield at most 2 points, the centers need not necessarily be *immediate* predecessors, as the DMDGP require. To embed a vertex $v$ using the embedding of vertices before $v$ in the order, it suffices that there are at least $K$ adjacent predecessors of $v$. Because of this, we can relax Discretization and define a larger class of discretizable instances [284]. For $v \in V$, if $V$ is ordered let $\gamma(v)$ be the set of predecessors of $v$.

Discretizable Distance Geometry Problem (DDGP). Given a positive integer $K$, a nonnegatively weighted simple undirected graph $G = (V, E, d)$, an order $<$ on $V$ and a mapping $x' : [K] \to \mathbb{R}^K$ such that:

1. $x'$ is a valid embedding of $G[[K]]$ (Start)
2. $\forall v \in V \setminus [K]$ ($|N(v) \cap \gamma(v)| \geq K$) (Discretization)
3. $\forall v \in V \setminus [K]\ \exists U_v \subset N(v) \cup \gamma(v)$ ($G[U_v] = \mathbf{K}_K \wedge \Delta_{K-1}(U_v) > 0$) (Strict Simplex Inequalities),

is there a valid embedding $x$ of $G$ in $\mathbb{R}^K$ extending $x'$?

Again, we denote DDGP instances with fixed $K$ by DDGP$_K$. By Discretization and Strict Simplex Inequalities, the DDGP can be solved using BP (Alg. 10) — just replace Step 9 with "let $U' = U_{v+1}$". We remark that the results of Sect. 18.2.4 do not apply to the DDGP.

Requiring $G[U_v] = \mathbf{K}_K$ is a strong condition. In practice we usually relax $G[U_v] = \mathbf{K}_K \wedge \Delta_{K-1}(U_v) > 0$ to simply $|U_v| = K$. This does not necessarily ensure that the instance can be discretized. However, because BP is an iterative algorithm on the order of $V$, the positions of all vertices in $U_v$ are known before embedding $v$, which implies that the Strict Simplex Inequalities condition can be verified by the BP algorithm itself.

### 18.3.2 Problem complexity

Because the DDGP contains all DMDGP instances as a subproblem, it is **NP**-hard by restriction. We remark that the Discretization condition makes this problem the "smallest" **NP**-hard problem with respect to $K$: replacing $K$ by $K + 1$ would yield instances having a $K$-trilateration order [139], for which the embedding problem is in **P**. This can be seen by restricting the set $P$ in Alg. 10 such that $|P| \leq 1$: the BP search tree width would then be bounded by 1, which means that the BP would have a worst-case running time $O(L|V|)$, where $L$ is the complexity of finding $P$.

## 18.4 Discretization orders

In the family of problems that the BP can solve, i.e. DMDGP and DDGP, an order $<$ on the vertex set $V$ is always given, guaranteeing that the edges in $E$ satisfy the Discretization requirement. In practice,

DMDGP instances coming from proteins are endowed with their natural *backbone order*, which may not satisfy Discretization. In this section we discuss the problem of finding a good order or determining that one such order does not exist.

> Discretization Vertex Order Problem (DVOP). Given a simple undirected graph $G = (V, E)$ and a positive integer $K$, establish whether there is an order $<$ on $V$ such that: (a) $\{v \in V \mid \rho(v) \leq K\}$ is a $K$-clique in $G$ and (b) for each $v \in V$ with rank $\rho(v) > K$, we have $|N(v) \cap \gamma(v)| \geq K$.

We note that the DVOP does not verify whether the order satisfies the Strict Simplex Inequalities requirement. This is because the set of distance matrices yielding a Cayley-Menger determinant (see Eq. (18.7)) having value exactly zero has Lebesgue measure zero within the set of all possible (real) distance matrices. **NP**-completeness of the DVOP follows trivially from **NP**-completeness of the $K$-clique problem, for finding a DVOP order implies finding $K$ vertices forming a clique in $G$.

Intuitively, the larger the sets $N(v) \cap \gamma(v)$ (for $v$ of rank exceeding $K$), the smaller the sets $P$ in Alg. 10 for early ranks will be, and the better the BP will perform. Sets of adjacent predecessors of size exactly $K$ ensure that $|P| \leq 2$, but more pruning distances to $v$ might make the current position for $v$ infeasible, thereby pruning the current branch and speeding up the search. We therefore also consider the optimization version of the DVOP:

> Optimal Discretization Vertex Ordering Problem (ODVOP). Given a simple undirected graph $G = (V, E)$ and a positive integer $K$, establish whether there is an order $<$ on $V$ such that: (a) $\{v \in V \mid \rho(v) \leq K\}$ is a $K$-clique in $G$ and (b) for each $v \in V$ with rank $\rho(v) > K$, $|N(v) \cap \gamma(v)|$ is maximum and exceeds $K$.

The ODVOP is a multi-objective maximization problem, whose objective function vector is $(|N(v) \cap \gamma(v)| \mid v \in V\ (\rho(v) > K))$. We prove in [236] that all DVOP solutions are in the Pareto set of the ODVOP. In practice, however, we use the ODVOP maximality requirements to influence the choice of the next vertex in the order in case of a draw. In other words, if there exist two or more candidate next vertices whose set of adjacent predecessors is greater than $K$, we choose one among the vertices yielding the largest such set.

**NP**-completeness of the DVOP notwithstanding, when $K$ is fixed the DVOP is in **P**: for each possible $K$-clique of $G$, we greedily build the order on $V$ by choosing large sets of adjacent predecessors earliest. Because $K$ is typically much smaller than $|V|$, and in practical instances arising from proteins $K$ is really fixed to 3, this algorithm performs fast enough to be able to determine useful orders as a pre-processing step to the BP.

As a testbed for DVOP-based techniques, we considered a subset of $\text{DDGP}_3$ instances from the PDB [55] where we kept all inter-atomic distances up to 5.5Å. With such a low threshold, the backbone order is not valid w.r.t. Discretization. Using the DVOP, we were able to embed all 18 protein graphs (from 90 to 2259 backbone atoms) in around 21 seconds of user CPU time for the whole test set (this includes solving the DVOP, which took 1/40th of the DDGP solution time on average), with average accuracy $10^{-10}$ measured in LDE (see Eq. (18.6)); this confirms the reliability of the BP. By comparison, DGSOL [282] in its standard configuration took 800s and yielded an average accuracy of $5 \times 10^{-1}$.

## 18.5 An artificial backbone of hydrogens

Our first attempt to consider NMR data, which usually provide distances between hydrogen atoms only if closer than a given threshold, has been presented in [243, 244, 285]. We defined an order for the hydrogens related to protein backbones which allows us to satisfy Discretization. Figure 18.5 shows the proposed vertex order, indicated by the black arrows in the picture and by their labels (showing

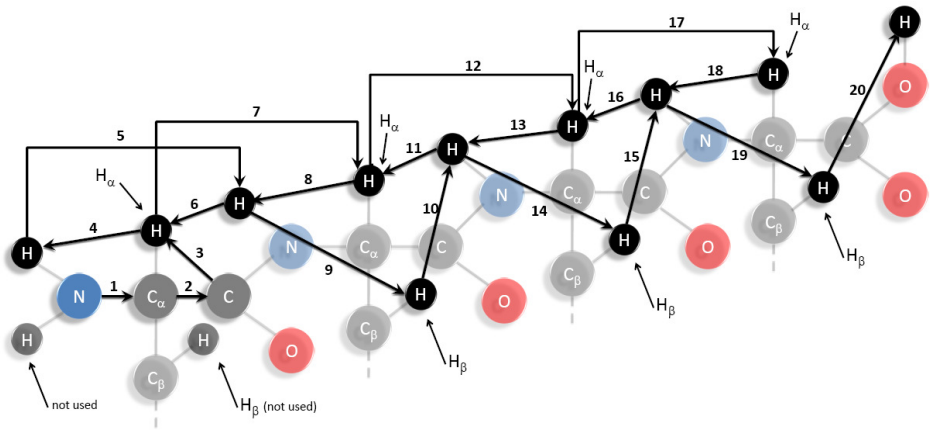



Figure 18.5: The artificial backbone of hydrogens satisfying the DMDGP requirements; the order is given in the arrow labels. Four amino acids are shown.

a progressive index), which we named *artificial backbone of hydrogens*. Note that this particular order considers the same atom more than once. Because of this, the relative distances between atoms farther in sequence are reduced, and a new kind of distance is introduced: the distance equal to zero existing between two copies of the same atom (obviously placed in the same spatial point).

Considering the same atom more than once does not change the problem complexity. In practice, the complexity of computing $P$ in Alg. 10 does not change, because the second copy of an atom can only be placed in the same place as the first copy. Thus, no branching occurs in correspondence with duplicated atoms, and the worst-case complexity of the BP variant exploiting orders with repetitions in still exponential in $|V|$. In [245], we showed that, because of steric constraints due to the particular structure of protein backbones, all distances necessary to guarantee DISCRETIZATION can be obtained by NMR. Once the problem is discretized and solved by the BP algorithm limited to hydrogen atoms, the remaining backbone atoms, and in particular the sequence of atoms $N$, $C_\alpha$, $C$ can be obtained by solving another MDGP. We proved that this MDGP is easy to solve, because assumptions stronger than the ones needed for the DMDGP are satisfied [243]. In particular, each other backbone atom $N$, $C_\alpha$, $C$ has at least 4 adjacent predecessors. As a consequence, the order is a trilateration order and the instance can be embedded in polynomial time [139].

Even though we showed that this approach works on a set of artificially generated instances (see for example the experiments in [245]), we remarked its limitations when we tried to apply it to real NMR data. The main issue is that the distances obtained by NMR are not precise (Sect. 18.1.5). Moreover, even though the second MDGP needed for finding the coordinates of the atoms $N$, $C_\alpha$ and $C$ is solvable in polynomial time, its solution relies on a sequence of quadratic systems to be solved, which can easily cause the propagation of round-off errors and numerical instabilities. The recently proposed order discussed in Sect. 18.6 overcomes both these issues.

## 18.6 *i*BP: discrete search with interval distances

The *interval* BP (*i*BP) is an extension of the BP algorithm which is able to manage interval data [241]. It is supposed that the distances $d_{v,v-1}$ and $d_{v,v-2}$, needed for the discretization, are always exact, whereas only the distances $d_{v,v-3}$ can be represented by an interval $[d^L_{v,v-3}, d^U_{v,v-3}]$. This way, the discretization process does not imply the definition of three spherical shells (see Sect. 18.1.5), whose intersection gives the possible positions for the current atom. If only the distance $d_{v,v-3}$ can be represented by an interval, the hardest subproblem to be solved is the one of finding the intersection of two spheres (related to

exact distances) and a spherical shell (related to the interval). This procedure would define a curve in the three-dimensional space in which the possible positions for the current atom can be searched. However, the equation of the curve would provide information on the atomic positions with a precision which is actually not needed for the purposes of the computation **LEO: cite**. Therefore, we can discretize the interval related to the distance $d_{v,v-3}$ and apply the standard discretization process for a subset of sample distances extracted from the interval $[d^L_{v,v-3}, d^U_{v,v-3}]$. Fig. 18.6 shows our hand-crafted order for a small protein backbone containing 3 amino acids. It shows the ordering for the first amino acid, for

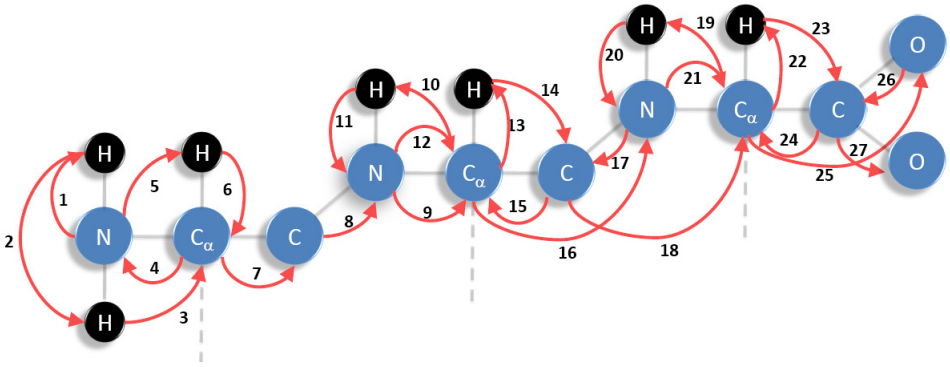



Figure 18.6: The order used for for discretizing MDGPs with interval data.

the second one, and for a generic amino acid, where the last three atoms concerns the last amino acid of the sequence. The ordering is specified by the red arrows in the picture and by their labels. It can be verified that $d_{v,v-1}$ and $d_{v,v-2}$ are always exact, because they can be computed from information known a priori on bond lengths and bond angles. Only $d_{v,v-3}$ may be represented by intervals. When this is the case, sample distances are taken from the interval and more than two branches are added to the tree representing the discrete search domain. As in the order shown in Sect. 18.5, we allow for repeated atoms.

The *i*BP algorithm is potentially able to solve MDGPs containing real data from NMR experiments. An important point is that no discretization distance is expected to be provided by NMR: all NMR distances (which are subject to imprecisions and errors) are pruning distances. This would make it possible to compute a finite over-approximation of the solution set (which only depends on the number of amino acids in the protein) where, potentially, the solutions to different DMDGP instances can be searched using the pruning distances. In other words, one might pre-compute exponential-sized sets of "potential embeddings" for proteins of given length, and then only traverse these trees by pruning when each DMDGP instance of conformant length is provided.

## 18.7 Implementation and parallelization

`MD-jeep` is an implementation of the BP algorithm in the C programming language [289]. It is distributed under the GNU General Public License (v.2) and it can be downloaded from `http://www.antoniomucherino.it/en/mdjeep.php`. `MD-jeep` accepts as input a list of distances in a text file with a predefined format, and returns PDB files containing the solutions to the problem as output. The PDB is a standard format for storing molecular conformations [55], which is compatible with many other software packages for molecular management and visualization. For example, two views, obtained using `RasMol` (`http://www.rasmol.org/`), of one of the solutions found by `MD-jeep` are given in Fig. 18.7.

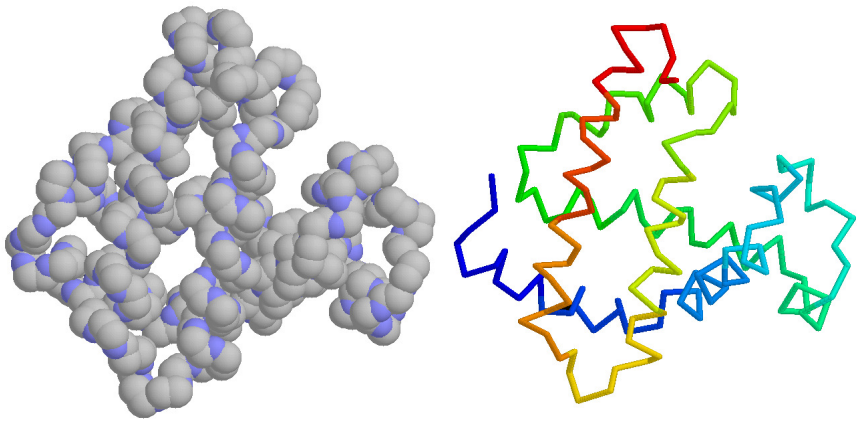


Figure 18.7: Two different graphical representations of the embeddings obtained by `MD-jeep`.

### 18.7.1 Parallel BP

We are also working on parallel implementations of the BP algorithm [287, 288] for the DMDGP. The basic idea is to partition the instance graph into subgraphs whose embeddings can be found independently by separate processes and then recombined; embedding each subgraph requires a call to a sequential BP algorithm, and the recombination is carried out by the master process.

Let us suppose that the number $n$ of vertices related to a given instance graph is divisible by the number $p$ of processes involved in the parallel computation. Then $p$ induced subgraphs $G_i = G[V_i]$ can be defined for all $i \leq p$ by setting:

$$V_i = \left\{1 + \frac{(i-1)n}{p}, \ldots, 3 + \frac{in}{p}\right\}.$$

For all $i \leq p$ we define $E_i = E[V_i] = \{\{u, v\} \in E \mid u, v \in V_i\}$. This partition of $V$ guarantees that each $G_i$ is a DMDGP instance if and only if $G$ is.

We remark that $\bar{E} = \bigcup_{i \leq p} E_i$ does not cover $E$; in particular, edges $\{u, v\} \in E$ with $u \in V_i, v \in V_j$ for $i \neq j$ do not belong to $\bar{E}$. As a consequence, the corresponding distances $d_{uv}$ are not used while the single processes work on the subgraphs $G_i$. However, they can be exploited later after the communication phase, when the local solutions found by the single processes are combined together in order to find the final set of solutions to the original instance.

The communication phase is implemented by following the classical cascade schema, so that only $\log_2 p$ communications are required to make $p$ processes exchange the partial embeddings found by the sequential calls to BP (we suppose that $p$ is a power of 2). Each partial embedding is coded by a sequence of binary variables. In order to reduce the time needed for the communication, each binary variable is stored in a single bit of an array of integer numbers. Each set of partial embeddings can be used for defining the local binary tree of solutions, which can be represented by graphs $T_k = (W_k, H_k)$, where vertices in $W_k$ represent atomic positions, and edges in $H_k$ connect vertices related to consecutive atomic positions. We employ the following procedure for combining the sets of partial embeddings found by two processes $k_1$ and $k_2 = k_1 + 1$. Let $T_{k_1,k_2} = (W_{k_1,k_2}, H_{k_1,k_2})$ be the graph which is the combination between $T_{k_1}$ and $T_{k_2}$. The vertex set $W_{k_1,k_2}$ is defined so that it contains all the vertices in $W_{k_1}$ and $W_{k_2}$, and the vertices in $W_{k_2}$ are also duplicated as many times as the number of leaf vertices in $W_{k_1}$, and new labels are assigned to them. The edge set $H_{k_1,k_2}$ is computed similarly, and, for each leaf vertex $v_l$ of $W_{k_1}$, a new edge is added between $v_l$ and the various copies of the first vertex of $W_{k_2}$. If this procedure is performed recursively considering all the graphs $T_k$, then the final tree of solutions, representing the final set of

embeddings, can be reconstructed. Distances related to atoms previously assigned to different processes can be used for pruning branches of the final tree for removing infeasible solutions.

Computational experiments (refer to [288] for more details) showed the efficiency of the parallel approach; the CPU time gain ratio between successive processor configurations (e.g. 1 against 2, 2 against 4 and so on) decreases as $p$ increases (in a few cases, executions with more processes actually took slightly longer). This is due to the fact that, as $p$ increases, the subgraphs assigned to each process get smaller, whereas the number of edges in $E \setminus \bar{E}$ increases. As a consequence, the calls to the (sequential) BP process on each subgraph tends to be less expensive than the master process that builds the BP tree for the whole graph. Parallel implementations overcoming this issue are currently under study.

## 18.8 Conclusion and future work

This paper gives an overview of the Discretizable Molecular Distance Geometry Problem, which offers a good model for finding protein structures with NMR data. We discussed variants, complexity, solution algorithms and extensions to deal with protein-specific features, such as limitations on the type of atoms that NMR usually provides information on.

On a short term, future work concerns the following topics: treatment of errors in the NMR data; polynomiality of the BP in the average case; exploitation of the BP tree symmetries. Longer term future work includes: the integration of the side chain embeddings; discovering unknown protein structures from real NMR data; synthesizing a BP-based integrated method to solve the PSNMR problem; looking for more applications (notably in embedding whole molecular complexes).

One notable open theoretical question is whether the DGP$_K$ is in **NP** for $K > 1$. The embedding that certifies a YES instance usually involves real numbers even though the instance data is rational (or even integer). As the embeddings solve a system of polynomials of second degree in several variables, it is easy to show that only algebraic numbers, rather than transcendental ones, are needed to express the components of each vector in the embedding. Thus, a finite precise symbolic representation for the embeddings is readily available, for example as the set of minimal polynomials having all the required algebraic numbers as roots. Whether all such numbers can be encoded by means of expressions whose length is polynomial in the instance size is as yet unclear.

# Chapter 19

# Quelles politiques publiques durables pour l'accès aux ressources naturelles?

Alassane Bah, Patrick Daquino, Ivan Lavallée, Diaraff Seck, Ibra Touré

*Report for OSD Project Lavallée-Dakar*

*Territorial laws in Senegal for regulating public access to natural resources are currently being debated. The debate concerns the trade-off between productivity, fairness, equitability and sustainability. This project aims to develop a modelling/simulation tool based on optimization techniques to evaluate the impact of different public policies concerning the assignment of natural resource access to different actors. Theoretically speaking, our proposal is based on a complex assignment problem with several objectives. This report concerns the first phase of the work where data are collected from domain actors (farmers and shepherds on the terrain).*

## 19.1 Introduction: les enjeux de la démarche de modélisation participative des politiques foncières

### 19.1.1 Les politiques foncières et de gestion des ressources naturelles : une cohérence à développer

Le renouvellement ou l'évaluation des politiques à l'ordre du jour Dans la région sahélienne, acteurs publics et organisations de la société civile réfléchissent à de nouvelles politiques, lois et règlementations pour optimiser les modes d'accès, la gestion, l'exploitation et l'appropriation des terres et des ressources naturelles renouvelables. De nombreuses initiatives sont développées en matière de réformes foncières, de décentralisation de la gestion des ressources, d'aménagement du territoire, d'adaptation aux changements climatiques, d'appui à l'investissement privé etc. Autant de politiques sectorielles qui touchent aux problématiques relatives à la terre et aux ressources naturelles et qui ont des effets combinés sur les modes d'accès à la terre et aux ressources renouvelables. Ces différentes politiques forment ainsi ce que nous appelons ci-après un cadre de régulations foncières. Les mesures et outils de régulation dépendent de pôles de décisions variés Ces choix politiques sectoriels, actuels et futurs, engendrent l'adoption de mesures et l'utilisation d'outils dont la gamme des possibles est très large : régulations économiques, mesures environnementales, diffusions de nouvelles pratiques culturales, règles domaniales, modes d'appropriation et de transmission du foncier, formalisation de marchés fonciers, règlementation de la

transhumance, schémas d'aménagement, distribution de l'occupation et de l'affectation des sols, règles d'accès et d'exploitation des ressources, délimitation physique des espaces ou territoires, plan foncier rural, cadastre, formalisation des transactions foncières, conventions locales etc.

Pour être cohérent, le cadre de régulations foncières doit combiner différentes mesures, de manière à traiter de façon complémentaire les différents enjeux posés par les politiques. Or ce cadre de régulations relève au niveau national de la responsabilité de différents départements ministériels et résulte également, notamment sous l'effet des processus d'intégration régionale et de décentralisation, de responsabilités situées à d'autres niveaux territoriaux : international, national, décentralisé, local. Cette multiplicité des lieux de décision représente une contrainte forte à la nécessaire cohérence des mesures adoptées.

Les effets des mesures et outils de régulation sont très difficiles à prévoir Selon leur nature et leur mode d'application, ces mesures et outils peuvent profondément modifier les équilibres socio-économiques et environnementaux (équilibres et complémentarités entre activités, impacts environnementaux, accroissement et répartition des richesses, paix sociale). Or il est très difficile de prévoir a priori les différents effets possibles , imprévisibilité accrue par l'aggravation de l'incertitude climatique.

De plus, un même outil peut servir à atteindre des objectifs opposés, selon la façon dont il est utilisé. Quelques exemples : l'élaboration d'une convention locale peut élargir la gamme des possibles en matière de modalités d'accès à la terre tout comme elle peut tendre à la réduire. Un outil comme le plan foncier rural peut être utilisé pour sécuriser les exploitations familiales ou faciliter l'installation de nouveaux investisseurs privés. La formalisation des transactions foncières peut dans certains cas favoriser les "propriétaires coutumiers", dans d'autres les exploitants.

En outre, sur un sujet aussi sensible, un outil ou une mesure peut engendrer des effets sur le terrain avant même d'être mis en œuvre. Il suffit en effet que les populations locales apprennent qu'un outil va être utilisé sans en maîtriser les conséquences pour qu'elles développent, par anticipation, des stratégies de limitation des risques. De nombreux cas ont par exemple été observés dans le cadre du projet Plan Foncier Rural en Côte d'Ivoire dans les années 90.

**Productivité, équité, durabilité, adaptabilité, ou le défi de la complémentarité des enjeux**

De multiples enjeux motivent la mise en place de ces nouvelles politiques, dont certains peuvent parfois apparaître contradictoires. Entre productivité, équité et durabilité, les responsables des politiques devront pourtant réussir à construire un cadre de régulations foncières qui intègrera complémentairement ces différents enjeux, auxquels s'ajoutent des enjeux émergents, comme faciliter l'adaptabilité des pratiques aux changements climatiques.

Or, ces enjeux sont rarement explicités et les oppositions entre acteurs aux intérêts distincts se cristallisent, dans un contexte d'aléas et de pressions croissantes sur les ressources : chacun des acteurs est porteur d'une partie des enjeux mais a du mal à prendre en compte, dans la définition des solutions, les enjeux des autres parties prenantes.

En conclusion, plusieurs obstacles sont aujourd'hui difficiles à franchir pour améliorer les cadres de régulations des terres et des ressources naturelles au Sahel :

- les mesures et outils de régulation mis en œuvre sont les produits de décisions sectorielles et prises à différentes échelles géographiques ;

- les effets sociaux, économiques ou environnementaux d'un mécanisme donné de régulation ne sont pas déterminables a priori, car ils dépendent de la façon dont la règle sera reçue, interprétée et utilisée par les différents acteurs ciblés et de la combinaison des effets de cette règle avec les autres mesures du cadre de régulation en place ;

- les enjeux prioritaires sont perçus comme contradictoires et donnent lieu à des positions doctrinales ;
- les méthodes de concertation mises en œuvre ont du mal à aboutir à des solutions pertinentes, où chaque acteur réussirait à intégrer de façon constructive les enjeux portés par les autres acteurs.

Des améliorations méthodologiques sont donc nécessaires, ce qui a abouti à la démarche suivante.

### 19.1.2 Les fondements de la démarche

La démarche proposée a comme objectifs :

1. une meilleure intégration par chacun des groupes ciblés (voir ci-dessous) des enjeux portés par les autres parties prenantes. Sur cette base :
2. le développement chez les groupes ciblés de capacités d'évaluation collective des pratiques et politiques actuelles ou possibles ;
3. la production collective entre ces différents groupes de propositions opérationnelles de mécanismes de régulation qui prennent en compte de façon complémentaire la diversité des enjeux identifiés comme prioritaires par ces groupes.

Cela passe par l'organisation d'une réflexion où les groupes ciblés mettent en commun leurs savoirs, leurs expériences et leurs perceptions pour co-évaluer des *scenarii* de régulation et co-construire des éléments pragmatiques de régulation. Ce type de co-construction peut aboutir à des propositions originales, car il permet de dégager des articulations positives entre des pratiques locales, parfois anciennes, et des outils et mesures appropriables par les différents pôles de décisions.

Pour ce faire, la démarche s'appuie sur des innovations méthodologiques :

1. des outils de transferts de compétences à portée opérationnelle (construction et évaluation de *scenarii* de régulation), qui mobilisent des expertises scientifiques pluridisciplinaires sous une forme très éloignée des canons académiques de la production scientifique ;
2. un support spatial de simulation collective , à l'image d'un jeu de société (sous forme de jeu de table ou informatisée), avec différentes cartes de terroirs (reprenant la diversité agro écologique existante), où les participants jouent des usagers de différents types, avec leurs besoins diversifiés (différents types d'agriculteur et d'éleveur, femme, jeune, autre type d'usager) ;
3. avec ce même support, le développement d'une réflexion prospective : les acteurs/joueurs construisent des *scenarii* de régulation (combinaison de mesures et d'outils) et d'aléas, puis simulent sur plusieurs années ces *scenarii* et les réactions/adaptations qu'il peut y avoir d'après eux chez les usagers ; ils analysent enfin les effets de ces *scenarii* sur les différents enjeux qu'ils ont identifiés précédemment ;
4. une réflexion "multi niveaux", car d'une part le support de simulation présente la caractéristique de pouvoir être utilisé avec des acteurs de profils très différents (du décideur à l'usager local, du scientifique au membre d'une ONG) et d'autre part car le support spatial permet d'observer à différentes échelles (exploitation familiale, lignage, terroir, région, pays) les effets de mécanismes de régulation. Son utilisation en parallèle avec les différents groupes ciblés permet alors une analyse collective équilibrée et constructive.

Les groupes ciblés sont :

- les populations rurales usagères des ressources renouvelables, y compris les groupes dits vulnérables (femmes, jeunes, exploitations familiales modestes, éleveurs transhumants). La méthodologie permet de réduire considérablement les asymétries observées dans les démarches participatives classiques au dépend de ces groupes d'acteurs.
- les organisations faîtières paysannes, qui pourront s'engager dans une dynamique de dénouement de situations tendues et de cohabitation améliorée entre acteurs aux intérêts variés.
- les collectivités locales, qui pourront recourir à cette démarche pour définir des mesures de régulation relevant de leurs responsabilités et en cohérence avec les politiques nationales.
- les praticiens du développement (projets, ONG), qui seront en situation d'adapter et de diffuser la démarche de prospective participative en fonction de leurs besoins, pour dépasser les blocages, de plus en plus décriés, relatifs aux démarches participatives classiques.
- les départements ministériels (Élevage, Agriculture, Environnement, Hydraulique etc.) qui pourront optimiser leurs compétences et leurs ressources dans un processus de formulation ou d'évaluation consensuelle de règles qui soient appropriables par tous.

### 19.1.3 La méthodologie "Prospective Participative Multi Niveaux" (2PMN)

La démarche repose sur une méthodologie originale de Prospective Participative Multi Niveaux (2PMN), qui relie des ateliers menés en parallèle au niveau local, régional et national, de telle sorte qu'ils puissent s'alimenter de leurs conclusions respectives.

Les ateliers sont conçus de manière à transférer progressivement et à un même rythme aux différents groupes ciblés les savoirs et les capacités leur permettant de formaliser et d'évaluer eux-mêmes, ensemble, des combinaisons de règles nationales, régionales et locales, aboutissant ainsi à des propositions validées par tous.

La 2PMN est constituée d'une série d'ateliers de simulation prospective réalisés séparément (en parallèle) auprès de chacun des différents groupes ciblés, avec le même support spatial de simulation prospective et selon les mêmes étapes. Le support de simulation prospective présente la caractéristique principale de pouvoir évoluer de façon itérative suivant les apports des différents groupes. Ces apports (sur les enjeux possibles d'une régulation, sur l'enrichissement des indicateurs de suivi, la modification du modèle) sont intégrés au fur et à mesure : chaque groupe s'alimente ainsi à chaque étape des conclusions respectives des ateliers des autres groupes.

La 2PMN se décompose schématiquement en 4 étapes, qui peuvent donner lieu à un nombre variable d'ateliers, pour s'adapter aux contraintes opérationnelles des groupes ciblés :

1. accord collectif sur les enjeux et indicateurs
    - identification collective des différents enjeux sociaux, économiques ou environnementaux possibles des cadres de régulations foncières, tels qu'ils sont perçus par les différentes parties prenantes ,
    - de là, définition des indicateurs d'évaluation (des mécanismes de régulation qui seront simulés) pour suivre les effets spécifiques sur ces différents enjeux;
2. test puis correction collective du jeu de simulation reprenant ces indicateurs
    - un support spatial, sous forme de différentes cartes de terroir englobant la diversité agro écologique du Sahel, sur lequel agissent des usagers de différents types, avec des besoins diversifiés et spécifiques (agriculteur, éleveur, femme, jeune),
    - chaque participant joue le rôle d'un de ces usagers ; tous les participants définissent et évaluent (selon les différents indicateurs) les mécanismes de régulation qui sont testés dans le jeu ,

- introduction dans le jeu de simulation de *scenarii* d'aléas ; puis introduction de premières combinaisons simples de mécanismes de régulation,
- simulations de ces différents éléments et suivi de leurs effets sur les indicateurs,
- analyses collectives et améliorations des *scenarii* à explorer;

3. Transfert de compétence sur les mécanismes complexes de régulation
   - quels sont les objectifs spécifiques de chacun des m écanismes de régulation possibles (du cadastre à la convention locale) ?
   - quels sont les effets possibles de chacun de ces différents mécanismes sur les différents enjeux sociaux, économiques et environnementaux ?
4. Construction et test collectif de *scenarii* complexes
   - simulation, sur plusieurs années, dans le modèle (*via* le jeu de table ou informatisé) des effets possibles de ces propositions sur les différents types d'usagers, sur les ressources et sur les différents enjeux sociaux, économiques et environnementaux répertoriés ,
   - de là, co-construction de propositions opérationnelles de régulation, appropriées et validées par les différents types d'acteurs participant au processus.

## 19.2 Mise en œuvre de la méthode de modélisation participative

**Lieu :** Les ateliers ont eu lieu dans un village agropastoral du centre Ouest du Sénégal, Widou, (communauté rurale de Tessekere).

Figure 19.1: Le milieu agropastoral, village de Widou

### 19.2.1 1ère étape : présentation de la démarche prospective participative multi niveaux sur les politiques de sécurisation foncière

Figure 19.2: Le milieu agropastoral et ses acteurs

**Questions abordées**[1]

- Présentation des enjeux de notre présence : leur donner l'occasion de produire eux-mêmes à partir de leurs savoirs des propositions originales de règles, se basant sur leur double connaissance : les savoirs traditionnels et leur expérience des projets de gestion des ressources naturelles qu'ils ont subis.
- Expliquez que l'on souhaite simplement aujourd'hui qu'ils comprennent bien les enjeux de notre proposition, pour qu'ils puissent s'organiser comme bon leur semble pour le véritable atelier, la prochaine fois, qui se tiendra sur trois jours.

**Méthodologie :**

**Présentation des enjeux de l'approche multi-niveaux participative prospective**

- L'enjeu de notre présence :
  - Les faire directement participer à la définition des politiques publiques foncières (= réfléchir à des règles à l'échelle du pays).
  - Pour obtenir de meilleures règles, besoin de mettre en commun les différents savoirs et points de vue, de les faire réfléchir ensemble ;

[1] Période des 8 et 9 février 2011

  - Faire co construire entre les différentes parties prenantes des propositions de politiques de régulation, via un transfert de compétences auprès de chacune d'entre elles sur la construction et l'évaluation des règles/lois
  - par aussi un échange soutenu entre les différents niveaux. Faire remonter leurs points de vue ("vers Dakar") sur ce qu'il faut conserver dans les pratiques foncières actuelles et ceux qu'il faut améliorer. Et d'un autre côté faire redescendre vers eux les points de vue des décideurs de Dakar sur le sujet : "chacun dans le costume de l'autre".
  - Pas simplement obtenir un consensus sur des propositions, mais créer des propositions de règles originales, à la croisée des différents savoirs

  *En résumé, les aider à produire des idées nouvelles de règles sous une forme compatible avec une politique publique, en mettant en commun les différents savoirs et points de vue, du local au national, du paysans au scientifique.*

- Notre proposition:
  - Notre objectif : leur fournir méthodes et outils pour qu'ils aient la capacité de faire cela
    * Faire mutuellement comprendre les contraintes et les enjeux de chacun,
    * puis proposer des supports pour qu'ils puissent trouver ensemble des solutions qui satisfassent les différents points de vue,
    * Le but : une amélioration des compétences des différentes parties (multi niveau) sur la construction, le suivi et l'évaluation de politiques publiques foncières
  - Notre position :
    * être un media entre les discussions à distance (le principe de diversité des points de vue, les aller retour, les ateliers et les supports identiques)
    * leur donner les compétences (à tous) pour qu'ils définissent mieux ensemble les solutions d'une situation complexe sans solution simple
    * la place du chercheur : n'est pas là pour leur donner une expertise, mais comme :
      1. "écrivain public", pour aider à mettre sur cartes et autres ce qu'ils pensent (cf. enjeux) ;
      2. "facteur", pour l'aller-retour entre les différents lieux d'ateliers, par exemple entre Dakar et ici, pour communiquer ce qui se dira dans chacun des lieux sur le même sujet ; il est en quelque sorte "la table sur laquelle aura lieu le dialogue entre des personnes qui ne sont pas au même; ce qui signifie de continuels aller-retour entre ateliers, pour que chacun des groupes avance au même rythme, à partir des mêmes cartes et autres, et puissent mieux échanger.

- Présentation en détails de notre démarche (les différents ateliers, leurs objectifs)
  - 1ère étape (entre "haut" et "bas", et avec ailleurs) : que chacun exprime ses points de vue et ses enjeux puis comprenne les points de vue et les enjeux des autres, sur les enjeux d'une politique publique foncière dans le pays ("prendre le costume des autres")
  - 2ème étape (entre "haut" et "bas", et avec ailleurs) : qu'ils imaginent ensemble et se communiquent les différentes possibilités de solution et qu'ils les testent dans un jeu au regard des différents points de vue déjà exprimés.

    *"Essayer sur la table avant d'essayer dans la réalité"* (simulation prospective) : personne ne peut avoir idée de l'effet de différentes règles, et de leur combinaison. Autant pouvoir avoir une idée un peu plus claire de leurs conséquences avant de les mettre en place...et du coup pouvoir essayer des choses différentes sans devoir trancher, prendre en compte les différentes idées sans être obligé de trancher.
  - 3ème étape (entre "haut" et "bas", et avec "ailleurs) : transfert des connaissances qui leur manquent pour affiner leurs idées de solution et surtout accès à l'ordinateur; nouveaux tests dans le jeu et l'ordinateur

### 19.2.2 2ème étape : Atelier d'appropriation de la démarche prospective participative multi niveaux sur les politiques de sécurisation foncière (2 au 4 Mars 2011)

**Questions abordées**

- La présentation de la démarche ;
- Une première discussion sur les enjeux fonciers ;
- Une première séance de jeu pour découverte (jusqu'à la SSF et avec l'autosuffisance alimentaire comme seul indicateur quantifié) ;
- La présentation des supports pour définir les scénarios de règles et la discussion y attenant, sur les règles possibles ;
- Un rendez vous pour une journée (courant juin) consacrée au test d'un premier scénario de règles, qu'ils vont réfléchir d'ici là.

**Méthodologie :**

**Analyse collective sur les enjeux potentiels du foncier**

1. Faire sentir la complexité de la question.
   Pour éviter les réponses qu'ils ont déjà en tête : affecter, zonage, etc.
   - L'animateur rappelle la logique des ateliers : d'abord on aide les décideurs des politiques de régulation à se mettre à la place ("dans le costume") des paysans, pour comprendre leurs besoins et leurs pratiques, puis on aide les paysans à se mettre à la place des décideurs, pour comprendre de la même façon leurs objectifs et leurs raisonnements ("pour réussir à être entendu de l'autre, il faut d'abord comprendre ce qu'il a dans la tête, en se mettant à sa place").
   - Jusqu'à présent les ateliers ont servi à construire le "costume de paysan" dans lequel devra se mettre le décideur pour comprendre les paysans : le "jeu", où les décideurs vont "jouer" les rôles des paysans et essayer de s'en sortir comme le font les paysans. Maintenant, c'est la deuxième étape : que les paysans comprennent les objectifs et le raisonnement des décideurs, pour qu'ils puissent ensuite "mettre leur costume" et réfléchir à des mécanismes de régulation.
     *"Après avoir construit un jeu pour représenter leurs perceptions et leurs besoins, il faut maintenant que ce soit eux qui se mettent à la place des décideurs". "Il faut comprendre d'abord comment ils pensent pour ensuite pouvoir se faire entendreet comprendre"*. Dans les deux sens : que les décideurs comprennent les paysans (construire le jeu), que les paysans comprennent les décideurs (choisir les règles collectives à appliquer au jeu).
   - L'animateur explique aux participants les enjeux des politiques publiques de régulation au Sahel, d'après les décideurs :
     (a) Il faut accroître la productivité des différentes activités, pour nourrir et développer le pays
     (b) Pour y arriver, il faut sécuriser l'investissement (qu'ils soient des paysans ou de nouveaux arrivants) : "on n'investi tpas sur l'agriculture parce que ce que on n'est pas sûr de garder suffisamment longtemps la terre une fois que l'on aura investi"
     (c) Pour sécuriser cet investissement, il faut une assurance que l'investissement que l'on met sur une terre ne nous sera pas retiré : sécurité du foncier, par un "papier"

(d) Pour aider à l'investissement, il faut aussi que les banques acceptent de prêter pour investir, et elles ne le feront que si elles ont une garantie de récuérer quelque chose de valeur si on ne les rembourse pas : un droit d'accès au foncier, via ces "papiers"

(e) Il faut donc introduire des papiers, que ce soit pour une propriété privée (à l'occidentale) ou pour des droits d'accès plus officiels et sur une durée suffisamment longue

(f) Et il faut que ces papiers soient accessibles à ceux qui ne sont pas du village mais qui ont les moyens d'investir sur l'agriculture ("l'argent est ailleurs, il faut permettre à ceux qui l'ont d'investir sur l'agriculture").

- L"animateur insiste bien sur le fait qu'il ne s'agit pas à ce stade de juger de ce raisonnement, mais de se " mettre dans le costume" de ceux qui le tiennent, donc de le comprendre et de le partager
- L'animateur donne ensuite les principales questions qui se posent les décideurs au moment de définir quels papiers seraient les plus pertinents:

  (a) Vaut-il mieux un droit de propriété exclusif, à l'occidentale (un seul propriétaire a tous les droits sur une parcelle et pour toujours) ou des droits d'usage (un" propriétaire" a le droit sur une seule activité et pour une durée déterminée) mais suffisamment sécurisé pour convaincre investisseurs et banquiers ? L'animateur explique que dans le cas de droit d'usage, la durée doit être suffisamment longue pour que l'investisseur (ou le banquier) acceptent de s'engager mais en même temps suffisamment courte pour que le propriétaire ne se sente pas dépossédé de sa terre.

  (b) Vaut-il mieux donner ces papiers à des individus, des chefs de famille, des chefs de village ?

  (c) Vaut-il mieux pouvoir vendre et acheter ces papiers ou bien l'empêcher ? L'animateur explique les dérives possibles d'un système marchand (concentration entre les mains des plus riches) et les dérives possibles d'un système non marchand (marchandisation informelle, non contrôlée).

- Selon les contextes et les réponses des participants, on peut prendre un autre exemple : pour limiter la pression pastorale sans aboutir des couts d'accès à l'eau trop élevés pour les éleveurs, on peut définir au départ le nombre d'éleveurs" pouvant avoir accès au terroir, puis distribuer le nombre de papiers correspondants à des éleveurs, selon des proportions (transhumants/ autochtones, entre chaque campement, etc.) à négocier d'abord. Cela a l'avantage de fixer de façon sûre la pression pastorale sans que cela soit difficile à supporter financièrement.

  *Inconvénient :* même si le "papier" est gratuit au départ, il y aura toujours un marché informel des papiers qui va se développer et cela risque d'être les plus riches qui finissent par récupérer les droits d'accès.

- Il n'y a pas de solution toute faite à ces questions et les décideurs comme les experts n'ont pas de réponse, mais partagent toutes ces questions. Si les participants ont bien compris ces questions, ils en savent donc déjà suffisamment pour réfléchir à quelles réponses ils se verraient donner, s'ils étaient "dans le costume du décideur". l'animateur insiste sur le fait qu'il s'ait d'un exercice, d'un apprentissage, et que, comme pour apprendre le vélo, il faut commencer par essayer, même si l'on sait que les premières fois, on va tomberDonc, un temps pour "essayer" : réfléchir, discuter et imaginer des réponses, puis les tester dans le jeu pour voir leur défauts et en imaginer de nouvelles.
- L'animateur demande aux participants de réfléchir entre eux à un (premier) avis sur les réponses possibles : temps de discussion et de débat;
- après ce temps libre de débat, l'animateur aide les participants à oser de premières réponses, en reprenant uniquement la première question et en leur demandant une réponse. Puis, une fois obtenu une réponse, il explique les avantages et les inconvénients qui peuvent en découler, en "déroulant l'histoire qui pourrait arriver avec cette règle".

  Ensuite il recommence la même approche avec la deuxième question :

(a) Vaut-il mieux un "papier" droit de propriété ou un papier "droit d'usage" sur une activité ?

A cette première question, il y a souvent une réponse en deux temps :

- "Il faut un droit de propriété" (pour sécuriser les paysans sur leur foncier traditionnel)
- L'animateur : "et comment fait on pour répondre au besoin de l'Etat : ouvrir la terre à l'investissement ?"
- "Il faut par-dessus cela des droits d'usages (avec papiers)

De là, l'animateur décortique ("déroule la pelote), les questions qu'entraîne ce choix, pour illustrer dans ce cas concret les questions qu'il a déjà soulevées plus haut, sans émettre de jugement de valeur de l'animateur sur ces différentes "histoires" possibles :

- dans ce système, quelle durée de droit d'usage sera-t-elle suffisante pour sécuriser un investissement lourd (type aménagement irrigué ou arbres fruitiers) tout en ne dépossédant pas le propriétaire ?
- ce type de formule peut aboutir à des évolutions où les propriétaires ont des dividendes mais n'ont plus de terres à cultiver ;
- cela peut aussi aboutir à la concentration des droits d'usage entre les mains de ceux qui auront le plus de moyens d'investissement ;
- il risque de ne plus avoir de terres disponibles pour ceux qui ne sont ni propriétaires fonciers ni investisseurs
- ...

(b) Vaut-il mieux autoriser l'achat/vente de ces papiers ou l'interdire ?
En règle générale, les participants répondent d'abord de l'interdire. L'animateur décortique alors de la même façon les questions qu'entraîne ce choix, pour illustrer les questions qu'il a déjà soulevées plus haut, *toujours sans émettre de jugement de valeur de l'animateur sur ces différentes "histoires" possibles :*

- dans ce système, les banquiers ne voudront pas prêter de l'argent aux propriétaires pour investir, puisqu'ils n'auront pas le droit de récupérer la terre s'ils ne remboursent pas
- Mais si l'on introduit un marché des terres, les propriétaires risquent petit à petit de vendre toute leurs terres lorsqu'ils auront besoin d'argent et les jeunes vont se retrouver sans terres ;
- cela peut aussi aboutir à la concentration des droits fonciers entre les mains de ceux qui auront le plus de moyens;
- ...

- Petit temps de réflexion et de débat.
- Puis l'animateur explique aux participants un deuxième "outil" pour les règles : le zonage. "Le décideur, qui a lui aussi conscience de toutes ces dérives possibles, à aussi accès à d'autres outils, avec lesquels il peut essayer d'empêcher les plus graves dérives". Premier exemple de ces outils, le zonage" :
  - on peut par exemple imaginer un zonage qui ne met dans le système des "papiers" qu'une partie des terres, l'autre restant dans le système traditionnel ;
  - on peut aussi imaginer de réserver une zone pour un type particulier d'usager, qu'il veut protéger (femmes, jeunes, etc.) ou pour une activité qu'il veut préserver (élevage, forêt).

  L'animateur "déroule" alors les "histoires" possible à partir de ces choix, pour illustrer les questions qu'il a déjà soulevées plus haut, toujours sans émettre de jugement de valeur de l'animateur sur ces différentes "histoires" possibles :
  - si l'on partage le terroir en deux zones, une zone "papiers", une zone "terroir", les meilleures terres risquent de se retrouver dans la zone "papiers" et les paysans réduit à utiliser les mauvaises terres de la zone terroir,
  - ou bien, les paysans qui sont dans la zone "papier" peuvent finir par vendre toutes leurs terreset demander ensuite à avoir un accès à la zone "terroir".

  - ...

  Différents exemples de "papiers" et de "zonages" possibles sont donc ainsi donnés, pour illustrer des idées d'utilisation de ces outils pour répondre aux contraintes qu'ils ont, en insistant toujours sur le fait que nous savons que ce sont pas de bonnes solutions, mais que les outils ainsi illustrés pourraient leur être utiles pour qu'ils construisent des solutions.

- Une fois que le zonage a été abordé, l'animateur peut introduire une troisième question, qui est facilement illustrée à partir du cas concret d'un zonage : la distribution des responsabilités sur la gestion foncière.

  "Qui sera responsable de décider, de gérer ces droits ?" On peut imaginer un zonage qui atténuerait les effets des "papiers (les participants ont peut-être déjà émis l'idée dans la discussion) : une partie du terroir en zone "papiers", une partie en zone "terroir". Dans ce cas, selon qui est chargé de faire le découpage, cela va aboutir à des "histories" très différentes:

  - si c'est l'État qui est responsable du découpage, il risque fort de mettre dans la zone "papiers" les meilleures terres, pour améliorer la production agricole du pays,
  - si c'est le chef de village, il risque au contraire d'y situer les terres les plus médiocres, ce que l'Etat ne peut pas accepter puisque l'objectif est d'ouvrir à l'investissement pour augmenter la production.
  - Ou bien le chef de village peut mettre dans la zone "papiers" seulement les volontaires parmi les paysans (car cela permet de faire fructifier monétairement ses terres), mais ces derniers risquent ensuite de se retourner vers la zone "terroir" une fois qu'ils auront tout venduou leurs enfants.
  - ...

- Après avoir fait sentir la complexité des conséquences de tout choix, on récapitule ensuite, pour conclure, les enjeux :
  - "les gens de Dakar et les experts ont conscience de toutes ces questions qu'ont vient de leur poser, mais ils n'ont pas plus de réponses qu'eux, car ils savent que chaque réponse a des avantages et des inconvénients (pour que les participants osent y réfléchir et ne pensent pas qu'ils ne sont pas assez expert pour cela)
  - Chaque réponse, même si elle paraît bonne, est un fil que l'on déroule et on ne sait pas ce que l'on va trouver au bout. Personne ne sait. D'où l'intérêt de l'essayer d'abord dans le jeu.
  - "de la même façon que les gens de Dakar se mettent à leur place en jouant leur rôles, ils vont maintenant se mettre à la place des gens de Dakar et essayer de choisir des règles/papiers du jeu"
  - On ne leur demande pas de donner une position qui serait ensuite leur position "officielle" : "c'est un apprentissage, ils vont avoir de premières idées, ils vont les essayer dans le jeu, ils vont se rendre compte qu'elles ont des effets qu'ils n'avaient pas prévu, ils vont recommencer. comme lorsqu'on apprend à faire du vélo : la première fois qu'on essaie, on sait que l'on va tomber, mais on sait aussi que c'est comme cela que l'on va apprendre : il faut oser essayer même si on sait que "on" n'a pas encore trouver la bonne idée".
  - C'est tout l'intérêt du jeu : essayer des règles qui risquent d'être mauvaises sans avoir les catastrophes que l'on aurait eu si on les avait essayé dans la réalité.

- Puis, on distribue une fiche donnant toute la complexité des questions à se poser pour définir un "papier" ou un "zonage" leur est distribuée (cf. annexe 1), en français et en Pulaar, en leur expliquant qu'il s'agit juste pour l'instant de découvrir (et non de maîtriser) la complexité des questions qu'il faudra (un jour) se poser. Mais que l'on reprendra progressivement tout cela avec eux, petit à petit.

- Puis, on leur demande de réfléchir à tout cela entre eux, jusqu'à la prochaine rencontre, à laquelle ils devront arriver avec de premières propositions collectives de règles foncières à tester dans le jeu de simulation. Cette prochaine rencontre aura lieu à Barkedji, les 5 et 6 Octobre, et rassemblera des "délégués" des différents endroits où ont été réalisés les ateliers : Malene Niani, Belel Tuflé et Dayane Gueloldé..

2. Diagnostic participatif des enjeux fonciers de chaque partie prenante Formuler le plus précisément possibles les différents enjeux que les différentes parties prenantes ressentent concernant les règles d'accès aux ressources naturelles et au foncier.

   - **Quelle situation veut-on atteindre dans 20 ans ?**
   - **Qu'est ce qui est bien, qui marche bien, qu'il faudrait conserver ?**
   - **Qu'est ce qui mérite amélioration, qui n'est pas bon, qu'il faudrait améliorer ?**
   - **Pourquoi faut-il changer quelque chose d'après eux ? Quelles sont les évolutions du contexte, de leur situation, qui peuvent impliquer des évolutions des règles d'accès**

3. Restitution des enjeux décrits dans les autres ateliers, par les autres parties prenantes, à l'issue du même type de discussion :

   - Une diversité, peut-être hétéroclite :
     - Depuis l'ouverture à l'investissement jusqu'à la préservation de la paix sociale, en passant par durabilité environnementale ;
     - Selon les enjeux d'un exploitant, d'un gestionnaire de terroir, d'un responsable régional, puis d'un responsable national
     - Selon un point de vue écologique, social, économique, culturel
   - Reconnaissance de l'existence d'une diversité de points de vue sur les enjeux, selon l'échelle et la région, au niveau régional, national et sous régional, selon le métier, la discipline, la priorité (sociale, environnementale, économique, culturelle)
   - Information sur les enjeux fonciers exprimés à différents niveaux : restitution de la liste d'enjeux en cours de constitution, à partir des ateliers précédents :
   - Validation de la réalité de cette diversité de points de vue, à prendre en compte et à faire dialoguer (sous entendu : il peut y avoir des solutions gagnant-gagnant, à explorer grce à la simulation participative)
     (a) Analyse des enjeux présentés : discussion, amendement, enrichissement, validation (y compris leur ordre de priorité entre ces enjeux et le comparer à ceux des autres, etc.)
     (b) Des enjeux aux outils, un chemin délicat :
   - Une fiche donnant toute la complexité des questions à se poser pour définir un "papier" ou un "zonage" leur est distribuée, en langue locale si possible, en leur expliquant qu'il s'agit juste pour l'instant de découvrir (et non maîtriser) la complexité des questions qu'il faudra se poser. Mais que l'on reprendra progressivement tout cela avec eux, petit à petit.
   - Toujours pour illustrer cette découverte d'outils, on précise alors le type d'avantages et inconvénients que peuvent entraîner deux de ces outils, le zonage règlementaire et l'appropriation personnalisée :

   Exemples :

   (a) Pour limiter la pression pastorale sans aboutir à des couts d'accès à l'eau trop élevés, on peut définir par exemple au départ le nombre d'éleveurs pouvant avoir accès au terroir, puis distribuer le nombre de "papiers" correspondants à des éleveurs, selon des proportions à négocier (p.e. transhumants/autochtones, entre chaque campement). Avantage : fixe de façon sûre la pression pastorale sans que cela soit difficile à supporter financèrement ; Inconvénient : même si le "papier" est gratuit au départ, il y aura toujours un marché informel qui va se développer et cela risque d'être les plus riches qui finissent par récupérer les droits d'accès.
   (b) Un zonage distinguant zone de terroir (sans "papiers ) et zone d'intensification (avec "papiers")
   (c) Un zonage distinguant élevage et agriculture Insister toujours sur le fait que ce ne sont pas de bonnes solutions, mais seulement des exemples des questions à se poser avant de réfléchir "outils", a cause de la complexité des effets induits possibles. La bonne démarche

est de commencer par s'entendre sur les enjeux, puis sur les règles effectives d'accès pouvant répondre à ces enjeux et ensuite, sur les outils permettant de mettre en œuvre ces règles (où, là, une connaissance des différents outils possibles leur sera transmise).

4. Apprentissage par l'action : *mise en jeu du prototype*

   (a) Rappel de l'enjeu de ce "jeu" :

   - "Travailler sur la même table en haut, en bas et ailleurs : un support commun de diagnostic,
   - que l'on construit ensemble (entre "haut" et "bas", et avec "ailleurs") et progressivement, en intégrant au fur et à mesure les points de vue des uns et des autres
   - pour évaluer ensemble les possibilités de règles les plus adaptées
   - Donc, qu'il permette de visualiser, et suivre les évolutions, de tous les enjeux repérés dans les étapes précédentes :
     - Comprendre les différents enjeux possibles d'une politique foncière
     - Comprendre les différents mécanismes concrets possibles pour ces politiques foncières
     - Analyser collectivement les effets des différents types possibles de politique foncière, en fonction des différents points de vue sur les enjeux à atteindre (accroissement mise en valeur, sécurisation des producteurs, préservation environnement...)
     - Réaliser ces analyses avec les concepteurs des politiques

   (b) Utilisation expérimentale du prototype de jeu, pour une simulation participative sur un cycle annuel

## 19.3 Déroulement du jeu de simulation

### 19.3.1 Présentation pédagogique, élément par élément, de la plateforme prototype déjà construite

**Les supports spatiaux des activités**

- des cartes qui ont été faites par des participants comme eux, dans d'autres ateliers;
- plusieurs régions, de type différent (Ferlo, Sine Saloum, Fleuve,), pour prendre en compte les effets des politiques foncières dans des régions différentes ainsi que les déplacements (hommes et animaux) entre zones ;
- des couleurs représentant schématiquement les différents types de milieux agro-pédologiques ;
- un parcellaire schématique, pour distinguer les différents espaces à exploiter.

**La visualisation des activités**

- des pions de couleur différentes pour chaque type d'activité (vert : agriculture, bleu : élevage, .), à positionner sur les parcelles ;
- que l'on déplace selon les saisons, pour exploiter les ressources des parcelles.

**La visualisation des enjeux**

- économiques familiaux (autosuffisance, "cagnotte" de chaque famille) ;
- économiques globaux (productions et productivités aux échelles locales, régionales et nationales)
- sociaux : préservation de l'accès aux ressources naturelles et aux revenus des différents groupes introduits dans le jeu (exploitations de différentes taille, groupes défavorisés), paix sociale
- culturels : interactions entre joueurs (liens sociaux), responsabilités collectives de gestion introduites dans le jeu, types de règles collectives mises en œuvre
- environnementaux : indicateurs visibles d'état des différentes ressources naturelles, sous tendus par des dynamiques écologiques formalisées de façon participative.

**La capacité de tout modifier facilement, d'introduire leurs idées et de les tester**

- sur les logiques des activités ;
- sur les règles collectives d'accès aux ressources et de répartition des responsabilités de gestion
- sur les aléas à introduire dans la simulation (pluviométrie, feu de brousse, modification des cours des intrants ou productions)
- sur l'introduction d'autres types d'usagers (agro business, sylviculture, migrants agricoles) et d'usages (bois de feu, chasse, cueillette de gomme arabique)

**L'intégration de deux dimensions importantes pour la réflexion sur les politiques foncières au Sahel**

- une perception multi niveaux des acteurs impliqués, des enjeux et des indicateurs ;
- la question de la mobilité (des troupeaux et des hommes) et l'évaluation des modalités de sa pertinence.

### 19.3.2 Définition du scénario à simuler

**Construction par les participants d'un scénario de règles (qu'ils souhaitent tester)**

- Présentation des trois fonds de carte pour la spatialisation des règles d'acc ès aux ressources, surtout de leur logique :
  1. Entrée par les règles effectives d'accès plutôt que par les outils (affectation, réserve) ;
  2. poser comme possibilité a priori (ils peuvent s'ils le veulent le supprimer) le multi usages d'un même espace et la distinction de règles différentes par saison ;
  3. Une représentation permettant une souplesse de construction : distinction des différents usages extensifs avec souplesse de suppression de certains d'entre eux; introduction d'usages intensifs et de leurs effets sur les autres accès possibles aux ressources du même espace ; modes d'accès aux points d'eau
- Puis on leur pose les questions suivantes, sur la bibliothèque de droits/règles qu'ils retiennent pour leur premier scénario :
  - *Quels usages ont besoin d'une attention particulière (préserver, favoriser) ?*
  - *Quels usagers ont besoin d'une attention particulière (préserver, favoriser) ? Quelles règles pour y parvenir ?*

  - *Quelles "brousses" (espaces, zones) ont besoin d'une attention particulière ? Quelles règles pour y parvenir ?*
  - *Quelles ressources ont besoin d'une attention particulière ? Quelles règles pour y parvenir ?*
  - *Quelles règles pour faciliter l'augmentation de la production ?*
  - *Quels usages globalement autoriser ?*
  - *Qui est autorisé à pratiquer ces différentes activités à chaque saison* (c'est-à-dire quels usagers globalement autorisés) ?

- Rappel du contenu de la Fiche enjeux fonciers: ont ils pensé à int égrer dans leurs choix ci-dessus les enjeux (des autres) qu'ils trouvent pertinents ?
- l'animateur rempli le recto de la Fiche 3 à partir de leurs réponses
  - type d'usages retenus ;
  - extensif/intensif ;
  - saisons concernés ;
  - groupe concerné (tous, local, femmes, jeunes) ;
  - modalités d'attribution (paiement) ;
  - ...

### 19.3.3 Choix de distribution spatiale de ces règles

Après avoir rappelé la logique retenue de formalisation des règles (cf. point précédent), l'animateur demande aux participants de distribuer ces droits dans l'espace (zonage) pour chaque saison (l'animateur note leurs choix au verso de la Fiche 3) :

1. Les droits extensifs sont inscrits par défaut sur les parcelles de trois fonds de carte (un par saison), sous forme de ronds colorés
2. Des petits post-it carrés de couleur neutre (jaune pale) sont prévus pour oblitérer certains de ces ronds si les participants décident de supprimer les droits qu'ils représentent
3. Les droits intensifs sont sous forme de post-it de couleur plus intense et de la taille d'une parcelle (permettant de supprimer les droits extensifs lorsque le post it est collé par-dessus), de couleur différente par activité. [2]
4. Les droits d'accès sélectifs sont représentables par une gommette colorée de forme ronde, placée au centre des ronds (usages) concernés, la couleur correspondant au type de sélection de l'accès (codes couleurs définis sur la fiche 3)
5. On peut aussi décréter l'interdiction totale d'accès à une parcelle (type mise en défens) avec des post it de couleur vert pale de la taille de la parcelle.
6. D'autres types de post-it sont disponibles, pour spatialiser toute autre forme de droit qu'ils définiraient

### 19.3.4 Simulation du scénario retenu de règles foncières

1. **Lancement du jeu :**

[2] remarquons qu'il s'agit ici d'un droit de mise en en valeur, pas d'un titre de propriété (qui est un outil particulier parmi d'autres pour formaliser ce droit)

- Calcul de la distribution des effectifs en joueurs par terroir (via Fiche 1)
- Présentation des trois supports terroir (différentes régions ; expression mobilité hommes et troupeaux) :
  - présentation de la structure en cases, équivalente de parcelles
  - illustration par la structuration en trois régions aux proportions différentes en types de "brousse"
- Calcul de la distribution des effectifs en joueurs par terroir (via Fiche 1)
- Distribution des gobelets "familles" à chaque joueur, chacun personnalisé avec un logo différent (du type marque de bétail)
- Distribution des cartes de main d'œuvre disponible à chaque famille, en fonction des proportions du calibrage de la Fiche 1
- Présentation des pions activités (élevage : pion bleu ; agriculture : pion vert ;)
- Chaque joueur choisit sa stratégie de répartition de sa force de travail entre agriculture et élevage et au niveau de l'élevage entre bovins et petits ruminants.

2. **Au début de la saison des pluies**

- Chaque joueur réfléchit à sa stratégie élevage/agriculture puis décide d'une localisation spatiale de ses activités en fonction des droits possibles sur les différentes parcelles.
- On insiste sur le fait que les productions obtenues seront fonction du type de brousse, de la pluviométrie, etc.
- Jet de dé : possibilité d'introduction d'un nouvel exploitant (Fiche 1)
- L'animateur marque la pression pastorale : il positionne des marqueurs de page verts (SP) sur les parcelles où il y a eu des troupeaux (1/ troupeau), en expliquant l'intérêt de repérer la pression sur les ressources (état des ressources)

3. **À la fin de la saison des pluies**

- Jet de dé : pluviométrie de la saison (Fiche 1)
- L'animateur distribue dans les gobelets les boules (production) produites par chaque joueur, en fonction (Fiche 2) :
- des règles de productivité selon type de milieu et pluviom étrie,
- des effets négatifs possibles des interactions agriculture élevage (trop grande proximité entre troupeaux et cultures),
- des effets négatifs de la pression pastorale
- des effets négatifs des modalités d'accès à l'eau (accès difficile ou impossible)
- Tableau 1A0. (penser à repérer les terroirs, selon leur position et leur région) On y note les productions obtenues

| | |
|---|---|
| Famille + (3) | |
| Famille X (1) | |
| Total/terroir NE-T1 | |
| Total/Région NE | |
| Agrobusin n1 | |
| Total/Ville (agrobusiness) | |
| Total Etat | |

- Puis retrait des besoins pour auto consommation : l'animateur retire à chacun les boules correspondantes, distribue les "carton rouge" à ceux qui n'ont pas atteint l'autoconsommation.
- Tableau 1A0. On note dans le premier tiers de la case la couleur (rouge/orange/vert) correspondant à l'indice d'auto suffisance alimentaire

- On peut vendre de la force de travail pour améliorer l'auto consommation

4. **Au début de la saison fraîche**
   - Chaque joueur réfléchit à sa stratégie élevage/agriculture puis décide d'une localisation spatiale de ses activités
   - Chaque joueur cherche à poser ses pions "activités" là où il l'a décidé;
   - L'animateur marque la pression pastorale : il positionne des marque page oranges sur les parcelles où il y a eu des troupeaux (1/troupeau).
5. **À la fin de la saison fraîche**
   - L'animateur distribue dans les gobelets les boules (production) produites par chaque joueur, en fonction (Fiche 2) :
     - des règles de productivité selon type de milieu et pluviométrie,
     - des effets négatifs possibles des interactions agriculture élevage (trop grande proximité entre troupeaux et cultures),
     - des effets négatifs de la pression pastorale
     - des effets négatifs des modalités d'accès à l'eau (accès difficile ou impossible)
   - On note les productions obtenues par chacun en couleur (production par activité) les boules obtenues par famille
   - Puis retrait des besoins pour auto consommation : l'animateur retire à chacun les boules correspondantes, distribue les "carton rouge" à ceux qui n'ont pas atteint l'autoconsommation.
   - Tableau 1A0. On note dans le deuxième tiers de la case la couleur (rouge/orange/vert) correspondant à l'indice d'auto suffisance alimentaire et renseigner en parallèle le verso de la Fiche 4 de suivi
6. **Au début de la saison sèche chaude**
   - Chaque joueur réfléchit à sa stratégie élevage/ agriculture puis décide d'une localisation spatiale de ses activités
   - Les feux de brousse apparaissent (cf. Fiche 1) : il faut repositionner ses activités, en fonction des règles d'accès existantes
   - Chaque joueur cherche à poser ses pions "activités" là où il l'a décidé.
   - L'animateur marque la pression pastorale : il positionne des page markers rouges sur les parcelles où il y a eu des troupeaux (1/troupeau).
7. **À la fin de la saison sèche chaude**
   - L'animateur distribue dans les gobelets les boules (production) produites par chaque joueur, en fonction (Fiche 2) :
     - des règles de productivité selon type de milieu et pluviométrie,
     - des effets négatifs possibles des interactions agriculture élevage (trop grande proximité entre troupeaux et cultures),
     - des effets négatifs de la pression pastorale
     - des effets négatifs des modalités d'accès à l'eau (accès difficile ou impossible)
   - Tableau 1A0. On y note les productions obtenues par chacun en couleur (production par activité) les boules obtenues par famille
   - Puis retrait des besoins pour auto consommation : l'animateur retire à chacun les boules correspondantes, distribue les "carton rouge" à ceux qui n'ont pas atteint l'autoconsommation.
   - Tableau 1A0. On note dans le dernier tiers de la case la couleur (rouge/orange/vert) correspondant à l'indice d'auto suffisance alimentaire et renseigner en parallèle le verso de la Fiche 4 de suivi.

8. **Bilans de l'année**

- Chacun des joueurs peut conserver le capital qui lui reste ou l'investir sur plus d'activités

- Bilan quantitatif des activités réalisées et du capital produit par terroir : faire les sommes (tel que ci-dessous) dans Tableau 1 A0 :

  – Faire les totaux de chaque production ;

  – Puis surligner en rouge, orange et vert le niveau d'auto satisfaction de chaque famille (lié à auto suffisance)

| **Famille** | **Année 1** | | | | **Année 2** | | | | **Année 3** | | | | **Bilans sur simulation** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SP | SSF | SSC | autosuf | SP | SSF | SSC | autosuf | SP | SSF | SSC | autosuf | Agr | Elev | Cueil | Total | Autosatisfaction |
| *Famille* + (3) | | | | | | | | | | | | | Σ | Σ | Σ | Σ | *Rouge/orange/vert* |
| *Famille* X (1) | | | | | | | | | | | | | Σ | Σ | Σ | Σ | *Rouge/orange/vert* |
| Total/terroir 1 | | | | | | | | | | | | | Σ | Σ | Σ | Σ | Proportion rouge/orange/vert |
| …… | | | | | | | | | | | | | Σ | Σ | Σ | Σ | *Rouge/orange/vert* |
| Total/Région 1 | | | | | | | | | | | | | Σ | Σ | Σ | Σ | Proportion rouge/orange/vert |
| ……. | | | | | | | | | | | | | Σ | Σ | Σ | Σ | *Rouge/orange/vert* |
| Total/Région 2 | | | | | | | | | | | | | Σ | Σ | Σ | Σ | Proportion rouge/orange/vert |
| Agrobusin n°1 | | | | | | | | | | | | | Σ | Σ | Σ | Σ | |
| Total/Ville (agrobusiness) | | | | | | | | | | | | | Σ | Σ | Σ | Σ | |
| Total Etat | | | | | | | | | | | | | Σ | Σ | Σ | Σ | Proportion rouge/orange/vert |

- Bilan collectif

  – bilan comparatif des différences: *pour chaque activité, pour bien jouer (et gagner),*

    * *Quels types de milieux doit on utiliser en fonction des saisons ?*

    * *Quelles limites dues à la main d'œuvre disponible ?*

    * *Quelles limites dues au capital disponible ?*

  – analyse des résultats: *Dans la simulation quels types d'acteurs (de joueurs), quels types d'usages ont ils été mieux préservés ? Si l'on avait continué la simulation avec ces règles collectives, comment cela risquait de continuer à évoluer ? Vers quelle diversité d'activités se dirigeait-on ? Vers quel type de structuration sociale et économique ? Vers quels niveaux de productivité ? Vers quel type d'environnement naturel ?*

  – selon différents points de vue :*D'après eux, à partir de quels résultats chaque point de vue est il satisfait ? Chef de famille ? Responsable local ? Responsable régional ?Etat ? Scientifique ?. . .*

– revenir aux indicateurs existant sur les tableaux:

| Famille | Année 1 | | | | Année 2 | | | | Bilans sur simulation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SP | SSF | SSC | autosuf | SP | SSF | SSC | autosuf | Agr | Elev | Cueil | Total | Autosatisfaction |
| *Famille* + (3) | | | | | | | | | Σ | Σ | Σ | Σ | *Rouge/orange/vert* |
| *Famille* X (1) | | | | | | | | | Σ | Σ | Σ | Σ | *Rouge/orange/vert* |
| …….. | | | | | | | | | Σ | Σ | Σ | Σ | *Rouge/orange/vert* |
| Total/terroir 1 | | | | | | | | | Σ | Σ | Σ | Σ | Proportion rouge/orange/vert |
| | | | | | | | | | Σ | Σ | Σ | Σ | *Rouge/orange/vert* |
| | | | | | | | | | Σ | Σ | Σ | Σ | *Rouge/orange/vert* |
| Total/Région 1 | | | | | | | | | Σ | Σ | Σ | Σ | Proportion rouge/orange/vert |
| | | | | | | | | | Σ | Σ | Σ | Σ | *Rouge/orange/vert* |
| Total/Région 2 | | | | | | | | | Σ | Σ | Σ | Σ | Proportion rouge/orange/vert |
| Agrobusin n°1 | | | | | | | | | Σ | Σ | Σ | Σ | |
| Total/Ville (agrobusiness) | | | | | | | | | Σ | Σ | Σ | Σ | |
| Total Etat | | | | | | | | | Σ | Σ | Σ | Σ | Proportion rouge/orange/vert |
| | | | | | | | | | | | | | |

*Point de vue des familles*

*Point de vue responsable local*

*Point de vue responsable régional*

*Point de vue responsable national*

9. **Amélioration de la plateforme**

   Expliquer que maintenant que le support de simulation est compris et qu'ils saisissent bien l'enjeu de l'exercice, ils peuvent vérifier si le jeu permet de visualiser et de suivre les différents enjeux identifiés sur le foncier, ce qui manque en précisions pour que les questions qu'ils se posent sur les politiques foncières (enjeux) puissent être traitées dans ce type de plateforme :

   - L'amélioration des indicateurs de la simulation : Les différents enjeux répertoriés peuvent ils être suivis de façon suffisamment subtile pour jauger des différences entre scénarios de simulation ? Pour chacun des enjeux identifiés, première réflexion collective pour repérer quels indicateurs concrets pourraient permettre de (mieux) les suivre et les évaluer dans un scénario de régulation ? Réfléchir à traduire chacun en effets observables sur le terrain : du point de vue d'un producteur local, d'un investisseur, d'un responsable foncier, d'un gestionnaire du terroir, d'un responsable régional ou national, d'une direction de l'environnement,
   - L'amélioration de la plateforme de simulation : Voir dans quelle mesure le jeu pourrait permettre de suivre les enjeux spécifiques qu'ils y ont ajoutés (par exemple les droits d'accès aux puits). Les différents éléments qu'ils jugent pertinents dans le choix d'un scénario de régulation peuvent ils être introduits dans ce type de simulation ? Quelles améliorations peut-on y intégrer ?...
   - Enfin, l'évaluation des profils nécessaires de participants :
     – Pour intégrer les différents types d'usagers : Dans la séance de "jeu", ont-ils cherché à reproduire la diversité des stratégies des différents types d'usagers présents sur un terroir type ou simplement à mettre en œuvre leurs stratégies propres habituelles ? Pensent ils qu'il faut intégrer dans le jeu les stratégies d'autres types d'acteur pour que ce soit plus réaliste ? Préfèrent ils les jouer eux-mêmes ont intégrer ces autres acteurs dans le jeu pour qu'ils les jouent eux-mêmes ?...
     – Pour faire participer aux séances tous les acteurs utiles à une réflexion sur une politique foncière : Pour imaginer et tester de règles foncières, quels acteurs qui ne participent pas pour l'instant à nos discussions et qu'ils pensent intéressants de faire participer ?

10. **Utilisation de la grille de lecture et d'analyse pour une prospective participative multi-niveaux sur les politiques foncières**

- Rappel des enjeux de la grille de lecture et d'analyse :
  - Mobiles et besoins des différents types d'usagers;
  - Diversité des points de vue sur le foncier
  - Représentation multi-niveau des effets
- Présentation de la grille de lecture de "réussite" d'une politique foncière ("indicateurs" visualisables des différents enjeux) :

| | Caractérisation des indicateurs | | | | |
|---|---|---|---|---|---|
| Enjeu | Nouveau terroir | | Nelle région | | Pays |
| Auto suffisance alimentaire | T1 | *prop.rouge/orange/vert* | | | |
| | T2 | | A | *idem* | |
| | T3 | | | | |
| | T4 | | B | *idem* | |
| | T5 | | C | *idem* | *idem* |
| Productions totales | | | | | |
| Production par activité | | | | | |
| État des RN | | | | | |
| Paix sociale | | | | | |
| Inégalités entre types d'exploitants | | | | | |
| ... | | | | | |

- Utilisation de la grille pour comparer les résultats des simulations de *scenarii* de politiques foncières:
  - Rappeler les scénarios de règles mises en œuvre dans chaque simulation (compulser les Fiches 3 de chaque simulation) ;
  - Revenir à la liste des enjeux et identifier les indicateurs qui manquent encore et en discuter
- Après usage, faire réfléchir les participants au contexte foncier initial à poser dans la simulation
  - Quelle distribution initiale des terres pourrait-on intégrer pour une simulation ensuite plus intéressante ?
  - Quels types de transaction (y compris collectives) ont-ils spontanément utilisés ? Faut-il en intégrer formellement dans les règles du jeu ?
  - Quels types de droits et de règles ont-ils spontanément utilisés ? Faut-il en intégrer formellement dans les règles du jeu ?
  - Quels types d'organisation pour les négociations foncières ont-ils spontanément utilisés ? Par famille ? Par lignage ?
  - Faut-il intégrer formellement dans les règles du jeu des types de responsabilité et d'autorité foncière ? Quels rôles supplémentaires devraient être introduits (PCR par exemple) ?
- Transférer si possible à des correspondants locaux (agents techniques, membre d'ONG) la capacité de présenter la plateforme, pour plus grande diffusion et même utilisation autonome. Si possible, leur demander pour la prochaine fois d'écrire une explication en langue locale sur comment jouer le jeu (règles du jeu), que l'on reproduira ensuite (une fois corrigé à la séance suivante) pour qu'ils puissent le diffuser localement aux autres personnes qu'ils souhaitent associer : **constitue une phase supplémentaire d'appropriation collective des règles du jeu, par diffusion entre ceux qui ont mieux compris et les autres (et d'évaluation de leur niveau d'appropriation)**.

## 19.4 Premiers résultats

Il s'agit de résultats provisoires, issus du premier atelier, qui seront confirmés et développés à l'atelier suivant.

### 19.4.1 Des idées originales de régulation collective, à développer

a) Pour s'adapter à l'incertitude climatique et l'extrême variabilité spatiale et temporelle des ressources naturelles, les paysans proposent, en puisant dans leurs principes traditionnels de gestion, de construire un système complexe de régulation se basant sur les principes suivants : deux types de régulation :

⇒ l'une, souple, pour les combinaisons espaces-temps où la situation de compétition est moyenne (p.e. la plupart des espaces en année à pluviométrie bonne). Dans ces contextes, toutes les activités sont autorisées partout, mais l'espace est divisé en zones "de priorité", qui définissent les responsabilités : par exemple, dans les zones "à priorité agricole", les éleveurs ont le droit de venir mais ils sont responsables des dégts qu'ils pourraient occasionner aux champs (fourrière, amendes), tandis que dans les zones "à priorité élevage" les agriculteurs ont le droit d'installer des champs mais sont responsables d'empêcher leur dégradation par els troupeaux (surveillance du champ, clôture).

⇒ Un autre type de régulation pour les "couples espace-temps" critiques (les bas fonds lorsque la pluviométrie est mauvaise, certaines ressources en saison sèche, les points d'eau). Dans ces contextes-là, la gestion doit être collective et selon des règles très strictes.

b) La garantie bancaire (pour les investissements productifs) est habituellement assurée par des biens immobiles (immobilier, propriété foncière) alors que vu les conditions climatiques du Sahel, ce sont les biens mobiles (troupeaux par exemple) qui sont les plus sûrs, puisqu'ils sont moins sensibles aux variations climatiques car on peut les déplacer vers des zones moins touchées au moment des crises de sécheresse. Il serait ainsi utile d'imaginer un système de garantie bancaire reposant sur des biens mobiles ou leur production.

### 19.4.2 Une première caractérisation économique des différents types de productions

Au Sénégal, l'affectation foncière est conditionnée pour l'instant à une mise en valeur de la parcelle, qu'il faut justifier. Or, au Sahel, les zones les plus productives sont différentes d'une saison à l'autre et d'une année à l'autre (selon le type de pluviométrie) : les systèmes les plus productifs sont donc ceux qui n'investissent pas sur une seule parcelle mais qui utilisent de façon la plus stratégique possible des espaces répartis dans le pays. Ceci est particulièrement vrai pour les productions animales, qui peuvent avoir une productivité doublée selon le type de déplacements qui est mis en œuvre. Les conditions d'intensification de la production ne sont donc pas qualifiables en franc à l'hectare mais en type de pratique technique, comme montré ci après pour l'élevage.

#### Élevage

Bovin:

- si pratique moyenne (peu de transhumance, pas d'accès aux zones les plus riches du sud du pays durant la saison ou les années les plus sèches) : 75 000 FCFA/an ou 90 000 F CFA/an (dépend du sexe du veau)
- si pratique performante (déplacements nombreux pour être sur les zones les plus riches à chaque saison) : 125 000 ou 150 000 F CFA/an (dépend du sexe du veau)

Ovin

- si pratique moyenne (peu de transhumance, pas d'accès aux zones les plus riches du sud du pays durant la saison ou les années les plus sèches) : 35 000 ou 15 000 F CFA/an (dépend du sexe de

l'agneau)

- si pratique performante (déplacements nombreux pour être sur les zones les plus riches à chaque saison) : 70 000 F 30 000 F CFA/an (dépend du sexe de l'agneau)

Caprin

- si pratique moyenne (peu de transhumance, pas d'accès aux zones les plus riches du sud du pays durant la saison ou les années les plus sèches) : 20 000 FCFA/an ou 10 000 F CFA/an (dépend du sexe du chevreau)
- si pratique performante (déplacements nombreux pour être sur les zones les plus riches à chaque saison) : 40 000 ou 20 000 F CFA/an (dépend du sexe du chevreau)

**Agriculture**

Mil:

- si bonne pluviométrie : 1 T/ha
- si pluviométrie moyenne : 0,8 T/ha
- si pluviométrie mauvaise : 0,4 T/ha

Arachide :

- si bonne pluviométrie : 2 T/ha
- si pluviométrie moyenne : 1,5 T/ha
- si pluviométrie mauvaise : 1 T/ha

Figure 19.3: Les intéressés en action

Figure 19.4: Le jeu lui même

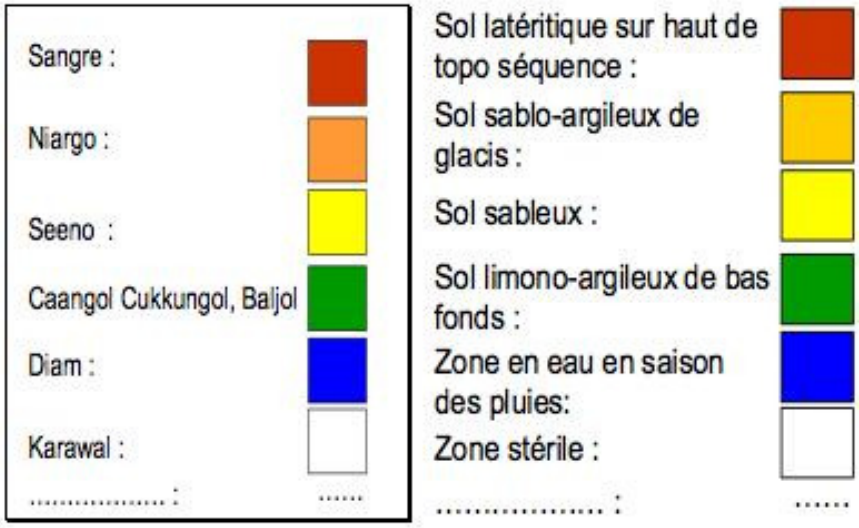



Figure 19.5: Types de brousses

Figure 19.6: Les grands troupeaux et l'accès à l'eau

Figure 19.7: Une agriculture très pauvre

Figure 19.8:

# Bibliography


[1] Subject, instances and results of the challenge 2010. http://challenge.roadef.org/2010/en/index.html.

[2] M. Abdullah and B. Dwolatzky. Smart demand-side energy management based on cellular technology - a way towards smart grid technologies in africa and low budget economies. *IEEE AFRICON 2009*, pages 1–5, 2009.

[3] M. Abkowitz and P.D. Cheng. Developing a risk-cost framework for routing truck movements of hazardous materials. *Accident Analysis & Prevention*, 20(1):39–51, 1988.

[4] W.P. Adams and H.D. Sherali. A tight linearization and an algorithm for 0-1 quadratic programming problems. *Management Science*, 32(10):1274–1290, 1986.

[5] C.S. Adjiman, S. Dallwig, C.A. Floudas, and A. Neumaier. A global optimization method, $\alpha$BB, for general twice-differentiable constrained NLPs: I. Theoretical advances. *Computers & Chemical Engineering*, 22(9):1137–1158, 1998.

[6] V. Akgun, E. Erkut, and R. Batta. On finding dissimilar paths. *European Journal of Operational Research*, 121(2):232–246, 2000.

[7] V. Akgun, E. Erkut, and R. Batta. On finding dissimilar paths. *European Journal of Operational Research*, 121(2):232–246, 2000.

[8] A. Alfaikh, A. Khandani, and H. Wolkowicz. Solving Euclidean distance matrix completion problems via semidefinite programming. *Computational Optimization and Applications*, 12:13–30, 1999.

[9] S. Amin amd B. Wollenberg. Toward a smart grid. *IEEE power and energy magazine*, 3:34–41, 2005.

[10] American Society of Heating, Refrigerating and Air-Conditioning Engineers. *ASHRAE Handbook of Fundamentals*. ASHRAE, 2009.

[11] S. Amin. For the good of the grid. *IEEE power and energy magazine*, 6:48–59, 2008.

[12] S. Amin. Securing the electricity grid. *The Bridge*, 40:13–20, 2010.

[13] S. Amin and P. Schewe. Preventing blackouts. *Scientific American Magazine*, May:60–67, 2007.

[14] S. Amin and J. Stringer. The electrical power grid: Today and tomorrow. *MRS BULLETIN*, 33:309–407, 2008.

[15] D. Anciaux, D. Roy, and S. Mirdamadi. Un modèle de simulation d'une chaîne logistique avec prise en compte des moyens de transports intermodaux. In *Proceedings of 5ème Conf??rence Internationale sur la Conception et Production Intégrées, Casablanca, Maroc*, 2005.

[16] E. Angelelli and M.G. Speranza. The periodic vehicle routing problem with intermediate facilities. *European Journal of Operational Research*, 137:233–247, 2002.

[17] Plamen P. Angelov, Y. Zhang, Jonathan A. Wright, V. I. Hanby, and R. A. Buswell. Automatic design synthesis and optimization of component-based systems by evolutionary algorithms. In Erick Cantú-Paz et al, editor, *Proc. GECCO*, volume 2724 of *LNCS*, pages 1938–1950. Springer, 2003.

[18] R. Aringhieri, M. Bruglieri, F. Malucelli, and M. Nonato. An asymmetric vehicle routing problem arising in the collection and disposal of special waste. In L. Liberti and F. Maffioli, editors, *CTW04 Workshop on Graphs and Combinatorial Optimization*, volume 17 of *Electronic Notes in Discrete Mathematics*, pages 41–46, Amsterdam, 2004. Elsevier.

[19] D. V. Arnold and R. Salomon. Evolutionary gradient search revisited. *IEEE Transactions on Evolutionary Computation*, 11(4):480–495, 2007.

[20] D.V. Arnold. Optimal weighted recombination. *Foundations of Genetic Algorithms*, pages 215–237, 2005.

[21] D.V. Arnold. Weighted multirecombination evolution strategies. *Theoretical computer science*, 361(1):18–37, 2006.

[22] A. Artmeier, J. Haselmayr, M. Leucker, and M. Sachenbacher. The optimal routing problem in the context of battery-powered electric vehicles. In *Workshop: CROCS at CPAIOR-10, Second International Workshop on Constraint Reasoning and Optimization for Computational Sustainability, Bologna, Italy*, 2010.

[23] A. Auger. Convergence results for the $(1, \lambda)$-SA-ES using the theory of $\varphi$-irreducible Markov chains. *Theoretical Computer Science*, 334(1–3):35–69, 2005.

[24] A. Auger, J. Bader, D. Brockhoff, and E. Zitzler. Investigating and Exploiting the Bias of the Weighted Hypervolume to Articulate User Preferences. In G. Raidl et al., editors, *Genetic and Evolutionary Computation Conference (GECCO 2009)*, pages 563–570, New York, NY, USA, 2009. ACM.

[25] A. Auger, J. Bader, D. Brockhoff, and E. Zitzler. Theory of the Hypervolume Indicator: Optimal $\mu$-Distributions and the Choice of the Reference Point. In *Foundations of Genetic Algorithms (FOGA 2009)*, pages 87–102, New York, NY, USA, 2009. ACM.

[26] A. Auger, J. Bader, D. Brockhoff, and E. Zitzler. Hypervolume-based Multiobjective Optimization: Theoretical Foundations and Practical Implications. *Theoretical Computer Science*, 2010. submitted.

[27] A. Auger, D. Brockhoff, and N. Hansen. Benchmarking the (1,4)-CMA-ES With Mirrored Sampling and Sequential Selection on the Noiseless BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1617–1624. ACM, 2010.

[28] A. Auger, D. Brockhoff, and N. Hansen. Benchmarking the (1,4)-CMA-ES With Mirrored Sampling and Sequential Selection on the Noisy BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1625–1632. ACM, 2010.

[29] A. Auger, D. Brockhoff, and N. Hansen. Comparing the (1+1)-CMA-ES with a Mirrored (1+2)-CMA-ES with Sequential Selection on the Noiseless BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1543–1550. ACM, 2010.

[30] A. Auger, D. Brockhoff, and N. Hansen. Investigating the Impact of Sequential Selection in the (1,2)-CMA-ES on the Noiseless BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1591–1596. ACM, 2010.

[31] A. Auger, D. Brockhoff, and N. Hansen. Investigating the Impact of Sequential Selection in the (1,2)-CMA-ES on the Noisy BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1605–1610. ACM, 2010.

[32] A. Auger, D. Brockhoff, and N. Hansen. Investigating the Impact of Sequential Selection in the (1,4)-CMA-ES on the Noiseless BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1597–1604. ACM, 2010.

[33] A. Auger, D. Brockhoff, and N. Hansen. Investigating the Impact of Sequential Selection in the (1,4)-CMA-ES on the Noisy BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1611–1616. ACM, 2010.

[34] A. Auger, D. Brockhoff, and N. Hansen. Mirrored Sampling and Sequential Selection for Evolution Strategies. Rapport de Recherche RR-7249, INRIA Saclay—Île-de-France, June 2010.

[35] A. Auger, D. Brockhoff, and N. Hansen. Mirrored Variants of the (1,2)-CMA-ES Compared on the Noiseless BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1551–1558. ACM, 2010.

[36] A. Auger, D. Brockhoff, and N. Hansen. Mirrored Variants of the (1,2)-CMA-ES Compared on the Noisy BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1575–1582. ACM, 2010.

[37] A. Auger, D. Brockhoff, and N. Hansen. Mirrored Variants of the (1,4)-CMA-ES Compared on the Noiseless BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1559–1566. ACM, 2010.

[38] A. Auger, D. Brockhoff, and N. Hansen. Mirrored Variants of the (1,4)-CMA-ES Compared on the Noisy BBOB-2010 Testbed. In *GECCO workshop on Black-Box Optimization Benchmarking (BBOB'2010)*, pages 1583–1590. ACM, 2010.

[39] A. Auger, D. Brockhoff, and N. Hansen. Analyzing the Impact of Mirrored Sampling and Sequential Selection in Elitist Evolution Strategies. In *Foundations of Genetic Algorithms (FOGA 2011)*. ACM, 2011. to appear.

[40] A. Auger and N. Hansen. A Restart CMA Evolution Strategy With Increasing Population Size. In *Congress on Evolutionary Computation (CEC 2005)*, volume 2, pages 1769–1776, Piscataway, NJ, USA, 2005. IEEE Press.

[41] A. Auger and N. Hansen. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In *Genetic and Evolutionary Computation Conference (GECCO 2006)*, pages 445–452, 2006.

[42] A. Auger and N. Hansen. Theory of evolution strategies: a new perspective. In A. Auger and B. Doerr, editors, *Theory of Randomized Search Heuristics: Foundations and Recent Developments*. World Scientific Publishing, 2010. In press.

[43] J. Bader. *Hypervolume-Based Search For Multiobjective Optimization: Theory and Methods*. PhD thesis, ETH Zurich, 2010.

[44] J. Bader and E. Zitzler. Hype: An algorithm for fast hypervolume-based many-objective optimization. *Evolutionary Computation*, 19(1):45–76, 2011.

[45] Johannes Bader and Eckart Zitzler. Hype: An algorithm for fast hypervolume-based many-objective optimization. *Evolutionary Computation*, 19(1):45–76, 2011.

[46] J. Bailleul and C. Byrnes. Geometric critical point analysis of lossless power system models. *IEEE Transactions on Circuits and Systems*, CAS-29(11):724–737, 1982.

[47] R. Baldacci, L. Bodin, and A. Mingozzi. The multiple disposal facilities and multiple inventory locations rollon???rolloff vehicle routing problem. *Computers & Operations Research*, 33:2667–2702, 2006.

[48] R. Bapna, L.S. Thakur, and S.K. Nair. Infrastructure development for conversion to environment friendly fuel. *European Journal of Operational Research*, 142(3):480–490, 2002.

[49] A. Bar-Noy, Y. Feng, M. Johnson, and O. Liu. When to reap and when to sow — lowering peak usage with realistic batteries. In C.C. McGeoch, editor, *Workshop on Experimental Algorithms — WEA*, volume 5038 of *LNCS*, pages 194–207, New York, 2008. Springer.

[50] Cynthia Barnhart, Christopher A. Hane, and Pamela H. Vance. Using branch-and-price-and-cut to solve origin-destination integer multicommodity flow problems. *Oper. Res.*, 48(2):318–326, 2000.

[51] P. Belotti, J. Lee, L. Liberti, F. Margot, and A. Wächter. Branching and bounds tightening techniques for non-convex MINLP. *Optimization Methods and Software*, 24(4):597–634, 2009.

[52] P. Belotti and G. Sun. Branching rules for integer multi-commodity network flow problems. Technical report, Lehigh University, 2009.

[53] T. Benoist, B. Estellon, F. Gardi, R. Megel, and K. Nouioua. Local solver: A black-box local-search solver for combinatorial optimization.

[54] B. Berger, J. Kleinberg, and T. Leighton. Reconstructing a three-dimensional model with arbitrary errors. *Journal of the ACM*, 46(2):212–235, 1999.

[55] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acid Research*, 28:235–242, 2000.

[56] P. Bernus, K. Mertins, and G. Schmidt. *Handbook on Architectures of Information Systems*. Springer, Berlin, 2006.

[57] D. P. Bertsekas, B. Gendron, and W. K. Tsai. Implementation of an optimal multicommodity network flow algorithm based on gradient projection and a path flow formulation. Technical Report LIDS-P-1364, Massachusetts Institute of Technology, 1984.

[58] D. Bertsimas and S.S. Patterson. The traffic flow management rerouting problem in air traffic control: a dynamic network flow approach. *Transportation Science*, 34(3):239–255, 2000.

[59] P. Beullens, L. Muyldermans, D. Cattrysse, , and D.V. Oudheusden. A guided local search heuristic for the capacitated arc routing problem. *European Journal of Operational Research*, 147(3):629–643, 2003.

[60] N. Beume, C. M. Fonseca, M. Lopez-Ibanez, L. Paquete, and J. Vahrenhold. On the Complexity of Computing the Hypervolume Indicator. Technical Report CI-235/07, University of Dortmund, December 2007.

[61] N. Beume, B. Naujoks, and M. Emmerich. SMS-EMOA: Multiobjective Selection Based on Dominated Hypervolume. *European Journal of Operational Research*, 181(3):1653–1669, 2007.

[62] N. Beume, B. Naujoks, M. Preuss, G. Rudolph, and T. Wagner. Effects of 1-Greedy $\mathcal{S}$-Metric-Selection on Innumerably Large Pareto Fronts. In M. Ehrgott et al., editors, *Conference on Evolutionary Multi-Criterion Optimization (EMO 2009)*, volume 5467 of *LNCS*, pages 21–35. Springer, 2009.

[63] L. Bianco, M. Caramia, and S. Giordani. A bilevel flow model for hazmat transportation network design. *Transportation Research Part C: Emerging Technologies*, 17(2):175–196, 2009.

[64] D. Bienstock and S. Mattia. Using mixed-integer programming to solve power grid blackout problems. *Discrete Optimization*, 4:115–141, 2007.

[65] D. Bienstock and A. Verma. The *n-k* problem in power grids: new models, formulations and computation. *SIAM Journal on Optimization*, to appear.

[66] M. Bisceglie, C. Galdi, A. Vaccaro, and D. Villacci. Cooperative sensor networks for voltage quality monitoring in smart grids. *IEEE Bucharest Power Tech Conference*, pages 1–6, 2009.

[67] S. Bleuler, M. Laumanns, L. Thiele, and E. Zitzler. Pisa—a platform and programming language independent interface for search algorithms. In C. M. Fonseca et al., editors, *Conference on Evolutionary Multi-Criterion Optimization (EMO 2003)*, volume 2632 of *LNCS*, pages 494–508, Berlin, 2003. Springer.

[68] S. Bliudze and D. Krob. Towards a functional formalism for modelling complex industrial systems. In P. Bourgine, F. Kepes, and M. Schoenauer, editors, *European Conference on Complex Systems*, 2005.

[69] S. Bliudze and D. Krob. Towards a functional formalism for modelling complex industrial systems. *ComPlexUs, special Issue : Complex Systems - European Conference 2005*, 2(3-4):163–176, 2006.

[70] J.M. Bloemhof-Ruwaard, P. Van Beek, L. Hordijk, and L.N. Van Wassenhove. Interactions between operational research and environmental management. *European journal of operational research*, 85(2):229–243, 1995.

[71] L. Blumenthal. *Theory and Applications of Distance Geometry*. Oxford University Press, Oxford, 1953.

[72] B.S. Boardman, E.M. Malstrom, D.P. Butler, and M.H. Cole. Computer assisted routing of intermodal shipments. *Computers & Industrial Engineering*, 33(1-2):311–314, 1997.

[73] L. Bodin, A. Mingozzi, R. Baldacci, and M. Ball. The rollon-rollof vehicle routing problem. *Transportation Science*, 34:271–288, 2000.

[74] M. Bojić and S. Dragičević. MILP Optimization of energy supply by using a boiler, a condensing turbine and a heat pump. *Energy Conversion and Management*, 43:591–608, 2002.

[75] M. Bojić and B. Stojanović. MILP Optimization of a chp energy system. *Energy Conversion Management*, 39(7):637–642, 1998.

[76] M. Bojić, B. Stojanović, and P. Mourdoukoutas. MILP Optimization of energy systems with a condensing turbine. *Energy*, 23(3):231–238, 1998.

[77] J.F. Bonnans. Mathematical study of very high voltage power networks i: The optimal dc power flow problem. *SIAM Journal on Optimization*, 7:979–990, 1997.

[78] J.F. Bonnans. Mathematical study of very high voltage power networks ii: The ac power flow problem. *SIAM Journal on Applied Mathematics*, 58:1547–1567, 1998.

[79] J.F. Bonnans. Mathematical study of very high voltage power networks iii: The optimal ac power flow problem. *Computational Optimization and Applications*, 16:83–101, 2000.

[80] Y.M. Bontekoning, C. Macharis, and J.J. Trip. Is a new applied transportation research field emerging? a review of intermodal rail-truck freight transport literature. *Transportation research, Part A, Policy and practice*, 38(1):1–34, 2004.

[81] S. Braithwait. Behavior modification. *IEEE Power & Energy Magazine*, may/june:36–45, 2010.

[82] J. Branke, K. Deb, H. Dierolf, and M. Osswald. Finding Knees in Multi-objective Optimization. In X. Yao et al., editors, *Conference on Parallel Problem Solving from Nature (PPSN VIII)*, volume 3242 of *LNCS*, pages 722–731. Springer, 2004.

[83] J. Brest, S. Greiner, B. Boskovic, M. Mernik, and V. Zumer. Self-adapting control parameters in differential evolution: A comparative study on numerical benchmark problems. *IEEE Trans. Evol. Comput.*, 10:646–657, 2006.

[84] K. Bringmann and T. Friedrich. The Maximum Hypervolume Set Yields Near-optimal Approximation. In *Genetic and Evolutionary Computation Conference (GECCO 2010)*. ACM, 2010. to appear.

[85] D. Brockhoff, A. Auger, N. Hansen, D. V. Arnold, and T. Hohm. Mirrored Sampling and Sequential Selection for Evolution Strategies. In R. Schaefer et al., editors, *Conference on Parallel Problem Solving from Nature (PPSN XI)*, pages 11–21. Springer, 2010.

[86] A. Brook, D. Kendrick, and A. Meeraus. GAMS, a user's guide. *ACM SIGNUM Newsletter*, 23(3-4):10–11, 1988.

[87] A. Brooks, E. Lu, D. Reicher, C. Spirakis, and B. Weihl. Demand dispatch: Using real-time control of demand to help balance generation and load. *IEEE Power & Energy Magazine*, may/june:20–29, 2010.

[88] G.G. Brown, C.J. Ellis, G.W. Graves, and D. Ronen. Wide area dispatching of mobil tank trucks. *Interfaces*, 17:107–120, 1987.

[89] R. Brown. Impact of smart grid on distribution system design. In *Conversion and Delivery of Electrical Energy in the 21st Century*, pages 1–4. IEEE, 2008.

[90] F. Butler. A call to order, a regulatory perspective on smart grid. In *The Business Scene*, 2009.

[91] Luisa Caldas. Generation of energy-efficient architecture solutions applying gene_arch: An evolution-based generative design system. *Advanced Engineering Informatics*, 22(1):59 – 70, 2008. Intelligent computing in engineering and architecture.

[92] M. Caramia and P. Dell'Olmo. Hazardous material transportation problems. In *Multi-objective Management in Freight Logistics: Increasing Capacity, Service Level and Safety with Optimization Algorithms*, pages 65–101. Springer-Verlag London, 2008.

[93] R.S. Carvalho, C. Lavor, and F. Protti. Extending the geometric build-up algorithm for the molecular distance geometry problem. *Information Processing Letters*, 108:234–237, 2008.

[94] Y. Caseau. *Performance du système d'information – Analyse de la valeur, organisation et management (in French)*. Dunod, Paris, 2007.

[95] G. Centrone, R. Pesenti, and W. Ukovich. Hazardous materials transportation: a literature review and an annotated bibliography. In C. Bersani, A. Boulmakoul, E. Garbolino, and R. Sacile, editors, *Advanced Technologies and Methodologies for Risk Management in the Global Transport of Dangerous Goods*. IOS NATO Science Series Book, Amsterdam, 2008.

[96] I. Chabini. A new shortest paths algorithm for discrete dynamic networks. In *Proceeding of 8th IFAC Symposium on Transport Systems*, pages 551–556, 1997.

[97] T. Chang, L. Nozick, and M. Turnquist. Multiobjective path finding in stochastic dynamic networks, with application to routing hazardous materials shipments. *Transportation Science*, 39(3):383–399, 2005.

[98] T.-S. Chang. Best routes selection in international intermodal networks. *Computers and Operations Research*, 35(9):2877–2891, 2008.

[99] F. Chauvet, N. Hafez, and J.-M. Proth. Electric vehicles: Effect of the availability threshold on the transportation cost. *Applied stochastic models in business and industry*, 15(3):169–181, 1999.

[100] D.M. Chitty and M.L. Hernandez. A hybrid ant colony optimisation technique for dynamic vehicle routing. *Lecture Notes in Computer Science*, 3102:48–59, 2004.

[101] J-P. Clarke, M. Lowther, L. Ren, W. Singhose, S. Solak, A. Vela, and L. Wong. En route traffic optimization to reduce environmental impact. Technical report, PARTNER Project 5 report, 2008.

[102] S. Coe, A. Ott, and D. Pratt. Demanding standards: Developing uniformity in wholesale demand response communications to enhance industry growth. *IEEE Power & Energy Magazine*, may/june:55–59, 2010.

[103] C. A. Coello Coello, G. B. Lamont, and D. A. Van Veldhuizen. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Springer, Berlin, Germany, 2007.

[104] I.D. Coope. Reliable computation of the points of intersection of $n$ spheres in $\mathbb{R}^n$. *Australian and New Zealand Industrial and Applied Mathematics Journal*, 42:C461–C477, 2000.

[105] J.-F. Cordeau. A branch-and-cut algorithm for the dial-a-ride problem. *Operations Research*, 54(3):573–586, 2006.

[106] J.-F. Cordeau and G. Laporte. The dial-a-ride problem (darp): variants, modeling issues and algorithms. *A Quarterly Journal of Operations Research*, 1:89–101, 2003.

[107] J. R. Current, C. ReVelle, and J. L Cohon. The maximum covering/shortest path problem: a multiobjective network design and routing formulation. *European Journal of Operational Research*, 21(2):189–199, 1985.

[108] J. R. Current, C. ReVelle, and J. L Cohon. The minimum-covering/shortest-path problem. *Decision Sciences*, 19(3):490–503, 1988.

[109] E.G. Koukios D. Voivontas, D. Assimacopoulos. Assessment of biomass potential for power production: a gis based method. *Biomass and Bionergy*, 20:101–112, 2001.

[110] S.E. Daniel, D.C. Diakoulaki, and C.P. Pappis. Operations research and environmental planning. *European journal of operational research*, 102:248–263, 1997.

[111] I. Das. On Characterizing the "Knee" of the Pareto Curve Based on Normal-Boundary Intersection. *Structural and Multidisciplinary Optimization*, 18(2–3):107–115, 1999.

[112] S. Das and P. N. Suganthan. Differential evolution – a survey of the state-of-the-art. *IEEE Trans. Evol. Comput.*, in press.

[113] L. Davis. Adapting operator probabilities in genetic algorithms. In *Proc. ICGA*, pages 61–69, 1989.

[114] R. Davis, C. Ernst, and D. Wu. Protein structure determination via an efficient geometric build-up algorithm. *BMC Structural Biology*, 10(Suppl 1):S7, 2010.

[115] K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley, Chichester, UK, 2001.

[116] K. Deb, M. Mohan, and S. Mishra. Evaluating the $\epsilon$-Domination Based Multi-Objective Evolutionary Algorithm for a Quick Computation of Pareto-Optimal Solutions. *Evolutionary Computation*, 13(4):501–525, Winter 2005.

[117] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, 6:182–197, 2002.

[118] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, 6:182–197, 2002.

[119] K. Deb, L. Thiele, M. Laumanns, and E. Zitzler. Scalable Test Problems for Evolutionary Multi-Objective Optimization. In A. Abraham, R. Jain, and R. Goldberg, editors, *Evolutionary Multiobjective Optimization: Theoretical Advances and Applications*, chapter 6, pages 105–145. Springer, 2005.

[120] K. Deb, L. Thiele, M. Laummans, and E. Zitzler. Scalable test problems for evolutionary multiobjective optimization. In A. Abraham et al., editor, *Evolutionary Multiobjective Optimization*, pages 105–145. Springer, 2005.

[121] G. Desaulniers and D. Villeneuve. The shortest path problem with time windows and linear waiting costs. *Transportation science*, 34(3):312–319, 2000.

[122] J. Desrosiers, Y. Dumas, and F. Soumis. A dynamic programming solution of the largescale single-vehicle dial-a-ride problem with time windows. *American Journal of Mathematical and Management Sciences*, 6:301–325, 1986.

[123] Jacques Desrosiers and Marco E. Lübbecke. A primer in column generation. In *Column Generation*, pages 1–32. Springer US, 2006.

[124] Christina Diakaki, Evangelos Grigoroudis, and Dionyssia Kolokotsa. Towards a multi-objective optimization approach for improving energy efficiency in buildings. *Energy and Buildings*, 40(9):1747 – 1754, 2008.

[125] I. Dobson, B. Carreras, V. Lynch, and D. Newman. Complex systems analysis of series of blackouts: Casacading failure, critical points and self-organization. *Chaos*, 17:026103, 2007.

[126] A.V. Donati, R. Montemanni, N. Casagrande, A.E. Rizzoli, and L.M. Gambardella. Time dependent vehicle routing problem with a multi ant colony system. *European Journal of Operational Research*, 185(3):1174–1191, 2008.

[127] Q. Dong and Z. Wu. A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances. *Journal of Global Optimization*, 22:365–375, 2002.

[128] Q. Dong and Z. Wu. A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. *Journal of Global Optimization*, 26:321–333, 2003.

[129] M. Dror. *Arc routing, theory, solutions, and applications*. Boston, MA: Kluwer, 2000.

[130] M. Dror, D. Fortin, and C. Roucairol. Redistribution of self-service electric cars: A case of pickup and delivery. Technical report, Technical Report W.P. 3543, INRIA-Rocquencourt, Rocquencourt, France, 1998.

[131] Olivier du Merle, Daniel Villeneuve, Jacques Desrosiers, and Pierre Hansen. Stabilized column generation. *Discrete Mathematics*, 194(1-3):229 – 237, 1999.

[132] N. Eén and N. Sörensson. Translating pseudo-boolean constraints into sat. *Journal on Satisfability, Boolean Modeling and Computation*, 2:1–25, 2006.

[133] M. Ehrgott. *Multicriteria optimization*. Lecture Notes in Economics and Mathematical Systems. Springer-Verlag, 2000.

[134] M. Ehrgott and X. Gandibleux. *Multiple Criteria Optimization. State of the art annotated bibliographic surveys*. Kluwer Academic, Dordrecht, 2002.

[135] A. E. Eiben, R. Hinterding, and Z. Michalewicz. Parameter control in evolutionary algorithms. *IEEE Trans. Evol. Comput.*, 3:124–141, 1999.

[136] M. Emmerich, N. Beume, and B. Naujoks. An EMO Algorithm Using the Hypervolume Measure as Selection Criterion. In *Conference on Evolutionary Multi-Criterion Optimization (EMO 2005)*, volume 3410 of *LNCS*, pages 62–76. Springer, 2005.

[137] M. Emmerich, A. Deutz, and N. Beume. Gradient-Based/Evolutionary Relay Hybrid for Computing Pareto Front Approximations Maximizing the S-Metric. In *Hybrid Metaheuristics*, volume 4771 of *LNCS*, pages 140–156. Springer, 2007.

[138] EnergyPlus. Getting started with energyplus. Technical report, Lawrence Berkeley National Laboratory, 2010.

[139] T. Eren, D.K. Goldenberg, W. Whiteley, Y.R. Yang, A.S. Morse, B.D.O. Anderson, and P.N. Belhumeur. Rigidity, computation, and randomization in network localization. *IEEE Infocom Proceedings*, pages 2673–2684, 2004.

[140] G. Ericsson. Cyber security and power system communication essential parts of a smart grid infrastructure. IEEE TRANSACTIONS ON POWER DELIVERY This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[141] E. Erkut and A. Ingolfsson. Transport risk models for hazardous materials: revisited. *Operations Research Letters*, 33:81–89, 2005.

[142] E. Erkut, S. Tjandra, and V. Verter. Hazardous material transportation problems. In *Handbooks in Operations Research & Management Science: Transportation*, pages 539–621. Elsevier, Editors. Cynthia Barnhart and Gilbert Laporte, 2007.

[143] E. Erkut and V. Verter. Modeling of transport risk for hazardous materials. *Operations Research*, 46(5):625–642, 1998.

[144] E. Erkut and V. Verter. Modeling of transport risk for hazardous materials. *Operations Research*, 46(5):625–642, 1998.

[145] N. Evmorfopoulos, D. Karampatzakis, and G. Stamoulis. Voltage-drop-constrained optimization of power distribution network based on reliable maximum current estimates. In *International Conference on Computer Aided Design*, pages 479–484, Washington D.C., 2004. IEEE.

[146] H. Farhangi. The path of the smart grid. *IEEE power and energy magazine*, 8:18–28, 2010.

[147] Á. Fialho, R. Ros, M. Schoenauer, and M. Sebag. Comparison-based adaptive strategy selection in differential evolution. In R. Schaefer et al., editor, *Proc. PPSN*, pages 194–203. Springer, 2010.

[148] Álvaro Fialho, Luis Da Costa, Marc Schoenauer, and Michele Sebag. Extreme value based adaptive operator selection. In G. Rudolph, editor, *Proc. PPSN*, pages 175–184. Springer, 2008.

[149] G. Fiorese, G. Marino, and G. Guariso. Environmental and economic evaluation of biomass as an energy resource: application to a farming provincial area. Technical report, DEI, Politecnico di Milano, 2005.

[150] M. Fleischer. The Measure of Pareto Optima. Applications to Multi-Objective Metaheuristics. In C. M. Fonseca et al., editors, *Conference on Evolutionary Multi-Criterion Optimization (EMO 2003)*, volume 2632 of *LNCS*, pages 519–533, Faro, Portugal, 2003. Springer.

[151] M. Fleischmann, J.M. Bloemhof-Ruwaard, R. Dekker, E. van der Laan, J.A.E.E. van Nunen, and L.N. van Wassenhove. Quantitative models for reverse logistics: A review. *European journal of operational research*, 103:1–17, 1997.

[152] C. Floudas and P. Pardalos, editors. *Encyclopedia of Optimization*. Springer, New York, second edition, 2009.

[153] R. Fortet. Applications de l'algèbre de Boole en recherche opérationelle. *Revue Française de Recherche Opérationelle*, 4:17–26, 1960.

[154] R. Fourer and D. Gay. *The AMPL Book*. Duxbury Press, Pacific Grove, 2002.

[155] D. Freppaz, R. Minciardi, M. Robba, M. Rovatti, R. Sacile, and A. Taramasso. Optimizing forest biomass exploitation for energy supply at a regional level. *Biomass and Bionergy*, 26:15–25, 2004.

[156] T. Friedrich, C. Horoba, and F. Neumann. Multiplicative Approximations and the Hypervolume Indicator. In G. Raidl et al., editors, *Genetic and Evolutionary Computation Conference (GECCO 2009)*, pages 571–578. ACM, 2009.

[157] M.R. Garey and D.S. Johnson. *Computers and Intractability: a Guide to the Theory of NP-Completeness*. Freeman and Company, New York, 1979.

[158] A. Geoffrion. Proper efficiency and the theory of vector maximization. *Journal of Mathematical Analysis and Applications*, 22:618–630, 1968.

[159] M.A. Turnquist G.F. List, P.B. Mirchandani and K.G. Zografos. Modeling and analysis for hazardous materials transportation: Risk analysis, routing/scheduling and facility location. *Transportation Science*, 25(2):100–114, 1991.

[160] G. Ghiani, F. Guerriero, G. Laporte, and R. Musmanno. Real-time vehicle routing: solution concepts, algorithms and parallel computing strategies. *European Journal of Operational Research*, 151(1):1–11, 2003.

[161] A. Ghosh, S. Boyd, and A. Saberi. Minimizing effective resistance of a graph. *SIAM Review*, 50(1):37–66, 2008.

[162] V. Giakoumakis, D. Krob, L. Liberti, and F. Roda. Optimal technological architecture evolutions of information systems. In M. Aiguier et al., editor, *Complex Systems Design and Management*, Berlin, 2010. Springer.

[163] I. Giannikos. A multiobjective programming model for locating treatment sites and routing hazardous wastes. *European Journal of Operational Research*, 104:333–342, 1998.

[164] M. Goldberg. Measure twice, cut once. *IEEE Power & Energy Magazine*, may/june:46–54, 2010.

[165] Wenyin Gong, Álvaro Fialho, and Zhihua Cai. Adaptive strategy selection in differential evolution. In J. Branke et al., editor, *Proc. GECCO*. ACM, July 2010.

[166] R. Gopalan, R. Batta, and M.H. Karwan. The equity constrained shortest path problem. *Computers and Operations Research*, 17:297–307, 1990.

[167] R. Gopalan, K.S. Kolluri, R. Batta, and M.H. Karwan. Modeling equity of risk in the transportation of hazardous materials. *Operations Research*, 38(6):961–973, 1990.

[168] L. Grandinetti, F. Guerriero, D. Laganà, and O. Pisacane. An approximate epsilon-constraint method for the multiobjective undirected capacitated arc routing problem. *Lecture Notes in Computer Science*, pages 214–225, 2010.

[169] P. Greistorfer. A tabu scatter search metaheuristic for the arc routing problem. *Computers & Industrial Engineering*, 44(2):249–266, 2003.

[170] H. Gunther. *NMR Spectroscopy: Basic Principles, Concepts, and Applications in Chemistry*. Wiley, New York, 1995.

[171] A. Hajimiraga, C. Ca nizares, M. Fowler, and A. Elkamel. Optimal transition to plug-in hybrid electric vehicles in ontario, canada, considering the electricity-grid limitations. *IEEE Transactions on Industrial Electronics*, 57(2):690–701, 2010.

[172] K. Hamilton and N. Gulhar. Taking demand response to the next level. *IEEE Power & Energy Magazine*, may/june:60–65, 2010.

[173] J. M. Hammersley and D.C. Handscomb. *Monte Carlo methods*. Methuen's monographs on applied probability and statistics. Methuen, 1967.

[174] E. Handschin, F. Neise, H. Neumann, and R. Schultz. Optimal operation of dispersed generation under uncertainty using mathematical programming. *Electrical Power and Energy Systems*, 28:618–626, 2006.

[175] N. Hansen. An analysis of mutative $\sigma$-self-adaptation on linear fitness functions. *Evolutionary Computation*, 14(3):255–275, 2006.

[176] N. Hansen, A. Auger, S. Finck, and R. Ros. Real-Parameter Black-Box Optimization Benchmarking 2009: Experimental Setup. INRIA Research Report RR-6828, INRIA Saclay—Ile-de-France, May 2009.

[177] N. Hansen, A. Auger, S. Finck, and R. Ros. Real-parameter black-box optimization benchmarking 2010: Experimental setup. Technical Report RR-7215, INRIA, 2010.

[178] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. Technical Report RR-6829, INRIA, 2009. Updated February 2010.

[179] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Noisy functions definitions. Technical Report RR-6869, INRIA, 2009. Updated February 2010.

[180] N. Hansen and S. Kern. Evaluating the CMA Evolution Strategy on Multimodal Test Functions. In X. Yao et al., editors, *Conference on Parallel Problem Solving from Nature (PPSN VIII)*, volume 3242 of *LNCS*, pages 282–291, Berlin, Germany, 2004. Springer.

[181] N. Hansen and A. Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[182] Nikolaus Hansen, André S. P. Niederberger, Lino Guzzella, and Petros Koumoutsakos. A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion. *IEEE Trans. Evolutionary Computation*, 13(1):180–197, 2009.

[183] P. Hansen and N. Mladenović. Variable neighbourhood search: Principles and applications. *European Journal of Operations Research*, 130:449–467, 2001.

[184] H. Hashimoto, M. Yagiura, and T. Ibaraki. An iterated local search algorithm for the time-dependent vehicle routing problem with time windows. *Discrete Optimization*, 5(2):434–456, 2008.

[185] T. Havel, I. Kuntz, and G. Crippen. The theory and practice of distance geometry. *Bulletin of Mathematical Biology*, 45(5):665–720, 1983.

[186] P. Van Hentenryck and R. Bent. *Online Stochastic Combinatorial Optimization*. MIT Press, 2006.

[187] A. Hertz, G. Laporte, and M. Mittaz. A tabu search heuristic for the capacitated arc routing problem. *Operations Research*, 48(1):129–135, 2000.

[188] A.V. Hill and W.C. Benton. Modeling intra-city time-dependent travel speeds for vehicle scheduling problems. *Journal of the Opertions Research Society*, 43(4):343–351, 1992.

[189] M.J. Hodgson. A flow capturing location-allocation model. *Geographical Analysis*, 22(3):270–279, 1990.

[190] Å. Holmgren. Using graph models to analyze the vulnerability of electric power networks. *Risk Analysis*, 26(4):955–969, 2006.

[191] C. Horoba. Analysis of a simple evolutionary algorithm for the multiobjective shortest path problem. In *Foundations of Genetic Algorithms (FOGA 2009)*, pages 113–120. ACM, 2009.

[192] V. L. Huang, S. Z. Zhao, R. Mallipeddi, and P. N. Suganthan. Multi-objective optimization using self-adaptive differential evolution algorithm. In *Proc. CEC*, pages 190–194. IEEE, 2009.

[193] S. Huband, P. Hingston, L. Barone, and L. While. A Review of Multiobjective Test Problems and a Scalable Test Problem Toolkit. *IEEE Transactions on Evolutionary Computation*, 10(5):477–506, 2006.

[194] S. Huband, P. Hingston, L. White, and L. Barone. An Evolution Strategy with Probabilistic Mutation for Multi-Objective Optimisation. In *Congress on Evolutionary Computation (CEC 2003)*, volume 3, pages 2284–2291, Canberra, Australia, 2003. IEEE Press.

[195] E. J. Hughes. Evolutionary Many-Objective Optimisation: Many Once or One Many? In *Congress on Evolutionary Computation (CEC 2005)*, pages 222–227. IEEE Press, 2005.

[196] M. Huneault. A survey of the optimal power flow literature. *IEEE Transactions on Power Systems*, 6(2):762–770, 1991.

[197] S. Ichoua, M. Gendreau, and J-Y. Potvin. Vehicle dispatching with time-dependent travel times. *European Journal of Operational Research*, 144(2):379–396, 2003.

[198] C. Igel, N. Hansen, and S. Roth. Covariance Matrix Adaptation for Multi-objective Optimization. *Evolutionary Computation*, 15(1):1–28, 2007.

[199] M. Ilić. From hierarchical to open access power systems. *Proceedings of the IEEE*, 95(5):1060–1084, 2007.

[200] ILOG. *ILOG CPLEX 10.0 User's Manual*. ILOG S.A., Gentilly, France, 2005.

[201] H.I. Inyang, editor. *Bridging the gaps for global sustainable development*. International Conference on Energy, Environment and Disasters, Charlotte, NC, July 2005.

[202] A. Ipakchi and F. Albuyeh. Grid of the future. *IEEE power and energy magazine*, 7:52–62, 2009.

[203] S. Irnich and G. Desaulniers. Shortest path problems with resource constraints. In G. Desaulniers, J. Desrosiers, and M.M. Solomon, editors, *Column Generation*, pages 33–65. Springer, 2005.

[204] M. Jebalia, A. Auger, and P. Liardet. Log-linear convergence and optimal bounds for the (1+1)-ES. In N. Monmarché and al., editors, *Proceedings of Evolution Artificielle (EA'07)*, volume 4926 of *LNCS*, pages 207–218. Springer, 2008.

[205] Mohamed Jebalia and Anne Auger. Log-linear Convergence of the Scale-invariant $(\mu/\mu_w, \lambda)$-ES and Optimal mu for Intermediate Recombination for Large Population Sizes. Research Report RR-7275, INRIA, 06 2010.

[206] Liyuan Jia, Wenyin Gong, and Hongbin Wu. An improved self-adaptive control parameter of differential evolution for global optimization. In Zhihua Cai, Zhenhua Li, Zhuo Kang, and Yong Liu, editors, *Computational Intelligence and Intelligent Systems*, volume 51 of *Communications in Computer and Information Science*, pages 215–224. Springer Berlin Heidelberg, 2009.

[207] Y. Jiao, F.H. Stillinger, and S. Torquato. Geometrical ambiguity of pair statistics I. point configurations. Technical Report 0908.1366v1, arXiv, 2009.

[208] P.E. Johnson, D.S. Joy, D.B. Clarke, and J.M. Jacobi. Higway 3.01, an enhanced highway routing model: Program, description, methodology and revised user's manual. Technical report, Oak Ridge National Laboratory, ORLN/TM-12124, Oak Ridge, TN, 1992.

[209] N. Jozefowiez, F. Semet, and E-G. Talbi. Multiobjective vehicle routing problems. *European Journal of Operational Research*, 189(2):293–309, 2008.

[210] N. Jozefowiez, F. Semet, and El-Ghazali Talbi. Parallel and hybrid models for multi-objective optimization: application to the vehicle routing problem. In J.J. Merelo Guervos et al. (Eds.), editor, *Parallel Problem Solving from Nature VII*, pages 271–280. Lecture Notes in Computer Science, 2002.

[211] Q. Jun, J. Wang, and B. Zheng. A hybrid multi-objective algorithm for dynamic vehicle routing problems. *Lecture Notes in Computer Science*, 5103:674–681, 2008.

[212] S. Jung and A. Haghani. Genetic algorithm for a pickup and delivery problem with time windows. *Transportation research record*, 1733:1–7, 2000.

[213] P. Kádár. Multi objective optimalisation of smart grid structure. In *International Conference on Intelligent System Applications to Power Systems*. IEEE-PES, 2009.

[214] B. Kallehauge, J. Larsen, O.B.G. Madsen, and M.M. Solomon. Vehicle routing problem with time windows. In G. Desaulniers, J. Desrosiers, and M.M. Solomon, editors, *Column Generation*, pages 67–98. Springer, 2005.

[215] J. Karkazis and T.B. Boffey. Optimal location of routes for vehicles transporting hazardous materials. *European journal of operational research*, 86(2):201–215, 1995.

[216] A.S. Kenyon and D.P. Morton. Stochastic vehicle routing with random travel times. *Transportation Science*, 37(1):69–82, 2003.

[217] B.I. Kim, S. Kim, and S. Sahoo. Waste collection vehicle routing problem with time windows. *Computers & Operations Research*, 33:3624–2642, 2006.

[218] R. King. Information services for smart grids. In *Power and Energy Society General Meeting*, volume 2008, Pittsburgh, 1-5. IEEE.

[219] J. Knowles. ParEGO: A Hybrid Algorithm With On-Line Landscape Approximation for Expensive Multiobjective Optimization Problems. *IEEE Transactions on Evolutionary Computation*, 10(1):50–66, 2005.

[220] J. Knowles and D. Corne. On Metrics for Comparing Non-Dominated Sets. In *Congress on Evolutionary Computation (CEC 2002)*, pages 711–716, Piscataway, NJ, 2002. IEEE Press.

[221] J. Knowles and D. Corne. Properties of an Adaptive Archiving Algorithm for Storing Nondominated Vectors. *IEEE Transactions on Evolutionary Computation*, 7(2):100–116, 2003.

[222] J. D. Knowles, D. W. Corne, and M. Fleischer. Bounded Archiving using the Lebesgue Measure. In *Congress on Evolutionary Computation (CEC 2003*, pages 2490–2497, Canberra, Australia, 2006. IEEE Press.

[223] H. Konno. A cutting plane algorithm for solving bilinear programs. *Mathematical Programming*, 11:14–27, 1976.

[224] MM. Kostreva and MM. Wiecek. Time dependency in multiple objective dynamic programming. *Journal of Mathematical Analysis and Applications*, 173(1):289–307, 1993.

[225] Daniel Krob. Modelling of complex software systems: A reasoned overview. In Elie Najm, Jean-François Pradat-Peyre, and Véronique Donzeau-Gouge, editors, *FORTE*, volume 4229 of *Lecture Notes in Computer Science*, pages 1–22. Springer, 2006.

[226] M. Kuby and S. Lim. The flow-refueling location problem for alternative-fuel vehicles. *Socio-Economic Planning Sciences*, 39(2):125–145, 2005.

[227] M. Kuby, X. Zhongyi, and X. Xiaodong. A minimax method for finding the *k* best differentiated paths. *Geographical Analysis*, 29(4):298–313, 1997.

[228] J. Kuchar and L. Yang. A review of conflict detection and resolution modeling methods. *IEEE Transactions on Intelligent Transportation Systems*, 1, 2000.

[229] S. Kucherenko and Yu. Sytsko. Application of deterministic low-discrepancy sequences in global optimization. *Computational Optimization and Applications*, 30(3):297–318, 2004.

[230] S. Kukkonen and J. Lampinen. GDE3: The third evolution step of generalized differential evolution. In *Proc. CEC*, pages 443–450. IEEE, 2005.

[231] P. Lacomme, C. Prins, and W. Ramdane-Cherif. Competitive memetic algorithms for arc routing problems. *Annals of Operations Research*, 131:159–185, 2004.

[232] P. Lacomme, C. Prins, and M. Sevaux. Multiobjective capacitated arc routing problem. *Lecture Notes in Computer Science*, 2632:550–564, 2003.

[233] P. Lacomme, C. Prins, and A. Tanguy. First competitive ant colony scheme for the carp. *Lecture Notes in Computer Science*, 3005:426–427, 2004.

[234] G. Laporte, F. Louveaux, and H. Mercure. The vehicle routing problem with stochastic travel times. *Transportation Science*, 26(3):161–170, 1992.

[235] M. Laurent. Matrix completion problems. In Floudas and Pardalos [152], pages 1967–1975.

[236] C. Lavor, J. Lee, A. Lee-St. John, L. Liberti, A. Mucherino, and M. Sviridenko. Discretization orders for distance geometry problems. *Optimization Letters*, accepted for publication.

[237] C. Lavor, L. Liberti, and N. Maculan. Computational experience with the molecular distance geometry problem. In J. Pintér, editor, *Global Optimization: Scientific and Engineering Case Studies*, pages 213–225. Springer, Berlin, 2006.

[238] C. Lavor, L. Liberti, and N. Maculan. The discretizable molecular distance geometry problem. Technical Report q-bio/0608012, arXiv, 2006.

[239] C. Lavor, L. Liberti, and N. Maculan. Molecular distance geometry problem. In Floudas and Pardalos [152], pages 2305–2311.

[240] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino. The discretizable molecular distance geometry problem. *Computational Optimization and Applications*, in revision.

[241] C. Lavor, L. Liberti, and A. Mucherino. The *i*BP algorithm for the discretizable molecular distance geometry problem with interval data. Technical report, IMECC-UNICAMP, working paper.

[242] C. Lavor, L. Liberti, A. Mucherino, and N. Maculan. On a discretizable subclass of instances of the molecular distance geometry problem. In D. Shin, editor, *Proceedings of the 24th Annual ACM Symposium on Applied Computing*, pages 804–805. ACM, 2009.

[243] C. Lavor, A. Mucherino, L. Liberti, and N. Maculan. An artificial backbone of hydrogens for finding the conformation of protein molecules. In *Proceedings of the Computational Structural Bioinformatics Workshop*, pages 152–155, Washington D.C., USA, 2009. IEEE.

[244] C. Lavor, A. Mucherino, L. Liberti, and N. Maculan. Computing artificial backbones of hydrogen atoms in order to discover protein backbones. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, pages 751–756, Mragowo, Poland, 2009. IEEE.

[245] C. Lavor, A. Mucherino, L. Liberti, and N. Maculan. On the computation of protein backbones by using artificial backbones of hydrogens. *Journal of Global Optimization*, accepted.

[246] H. Levy. Stochastic dominance and expected utility: Survey and analysis. *Management Science*, 38(4):555–593, 1992.

[247] S. Leyffer. User manual for MINLP_BB. Technical report, University of Dundee, UK, March 1999.

[248] M. Li, J. Zheng, and G. Xiao. Uniformity assessment for evolutionary multi-objective optimization. In *Proc. CEC*, pages 625–632. IEEE, 2008.

[249] L. Liberti. Automatic reformulation of bilinear MINLPs. Technical Report 2004.24, DEI, Politecnico di Milano, July 2004.

[250] L. Liberti. Reduction constraints for the global optimization of NLPs. *International Transactions in Operational Research*, 11(1):34–41, 2004.

[251] L. Liberti. Linearity embedded in nonconvex programs. *Journal of Global Optimization*, 33(2):157–196, 2005.

[252] L. Liberti. Reformulations in mathematical programming: Definitions and systematics. *RAIRO-RO*, 43(1):55–86, 2009.

[253] L. Liberti, S. Cafieri, and F. Tarissan. Reformulations in mathematical programming: A computational approach. In A. Abraham, A.-E. Hassanien, P. Siarry, and A. Engelbrecht, editors, *Foundations of Computational Intelligence Vol. 3*, number 203 in Studies in Computational Intelligence, pages 153–234. Springer, Berlin, 2009.

[254] L. Liberti and M. Dražic. Variable neighbourhood search for the global optimization of constrained NLPs. In *Proceedings of GO Workshop, Almeria, Spain*, 2005.

[255] L. Liberti and S. Kucherenko. Comparison of deterministic and stochastic approaches to global optimization. *International Transactions in Operational Research*, 12:263–285, 2005.

[256] L. Liberti, C. Lavor, and N. Maculan. A branch-and-prune algorithm for the molecular distance geometry problem. *International Transactions in Operational Research*, 15:1–17, 2008.

[257] L. Liberti, C. Lavor, N. Maculan, and F. Marinelli. Double variable neighbourhood search with smoothing for the molecular distance geometry problem. *Journal of Global Optimization*, 43:207–218, 2009.

[258] L. Liberti, C. Lavor, A. Mucherino, and N. Maculan. Molecular distance geometry methods: from continuous to discrete. *International Transactions in Operational Research*, 18:33–51, 2010.

[259] L. Liberti and N. Maculan, editors. *Global Optimization: from Theory to Implementation*. Springer, Berlin, 2006.

[260] L. Liberti, B. Masson, C. Lavor, J. Lee, and A. Mucherino. On the number of solutions of the discretizable molecular distance geometry problem. Technical Report 1010.1834v1[cs.DM], arXiv, 2010.

[261] L. Liberti, B. Masson, C. Lavor, and A. Mucherino. Branch-and-prune trees with bounded width. In *Proceedings of LAGOS 2010*, submitted.

[262] G. Lizarraga-Lizarraga, A. Hernandez-Aguirre, and S. Botello-Rionda. G-Metric: an M-ary quality indicator for the evaluation of non-dominated sets. In *Genetic And Evolutionary Computation Conference (GECCO 2008)*, pages 665–672, New York, NY, USA, 2008. ACM.

[263] K. Lombard and R.L. Church. The gateway shortest path problem: Generating alternative routes for a corridor location problem. *Geographical Systems*, 1:25–45, 1993.

[264] J.N. Luftman. *Competing in the Information Age*. Oxford University Press, Oxford, 2003.

[265] H. Lund and A.N. Andersen. Optimal designs of small chp plants in a market with fluctuating electricity prices. *Energy Conversion and Management*, 46(6):893–904, 2005.

[266] H. Lund and E. Münster. Integrated energy systems and local energy markets. *Energy Policy*, 34:1152–1160, 2006.

[267] O.B.G. Madsen, H.F. Ravn, and J.M. Rygaard. A heuristic algorithm for a dial-a-ride problem with time windows, multiple capacities, and multiple objectives. *Annals of Operations Research*, 60:193–208, 1995.

[268] C. Malandraki and M.S. Daskin. Time dependent vehicle routing problems: formulations, properties and heuristic algorithms. *Transportation Science*, 26:185–200, 1992.

[269] J. Maturana, F. Lardeux, and F. Saubion. Autonomous operator management for evolutionary algorithms. *J. Heuristics*, 2010.

[270] B.A. McCarl, D.M. Adams, R.J. Alig, and J.T. Chmelik. Competitiveness of biomass-fueled electrical power plants. *Annals of Operations Research*, 94:37–55, 2000.

[271] G.P. McCormick. Computability of global solutions to factorable nonconvex programs: Part I — Convex underestimating problems. *Mathematical Programming*, 10:146–175, 1976.

[272] Y. Mei, K. Tang, and X. Yao. Decomposition-based memetic algorithm for multiobjective capacitated arc routing problem. *IEEE Transactions on Evolutionary Computation, to appear*, 2010.

[273] M.W. Melaina. Initiating hydrogen infrastructure: preliminary analysis of a sufficient number of initial hydrogen stations in the us. *International Journal of Hydrogen Energy*, 28(7):743–755, 2003.

[274] L. De Meulemeester, G. Laporte, FV. Louveaux, and F. Semet. Optimal sequencing of skip collections and deliveries. *Journal of the Operational Research Society*, 48:57–64, 1997.

[275] K. Miettinen. *Nonlinear multiobjective optimization*. Kluwer Academic Publishers, Boston, 1999.

[276] E.D. Miller-Hooks. Adaptive least-expected time paths in stochastic, time-varying transportation and data networks. *Networks*, 37(1):35–52, 2001.

[277] E.D. Miller-Hooks and H.S. Mahamassani. Optimal routing of hazardous materials in stochastic, time-varying transportation networks. *Transportation Research Record*, 1645:143–151, 1998.

[278] H. Min. International intermodal choices via chance-constrained goal programming. *Transportation Research Part A: General*, 25(6):351–362, 1991.

[279] L. Moccia, J. F. Cordeau, G. Laporte, S. Ropke, and M.P. Valentini. Modeling and solving a multimodal routing problem with timetables and time windows. Technical report, 2008.

[280] B. Möller. The use of gis in planning biomass industries. In *Biomass production conference "Energy from forestry and agriculture"*, Elgin, November 2005.

[281] J. Momoh. Smart grid design for efficient and flexible power networks operation and control.

[282] J.J. Moré and Z. Wu. Global continuation for distance geometry problems. *SIAM Journal of Optimization*, 7(3):814–846, 1997.

[283] A. Mucherino and C. Lavor. The branch and prune algorithm for the molecular distance geometry problem with inexact distances. In *Proceedings of the International Conference on Computational Biology*, volume 58, pages 349–353. World Academy of Science, Engineering and Technology, 2009.

[284] A. Mucherino, C. Lavor, and L. Liberti. The discretizable distance geometry problem. *Optimization Letters*, in revision.

[285] A. Mucherino, C. Lavor, L. Liberti, and N. Maculan. On the definition of artificial backbones for the discretizable molecular distance geometry problem. *Mathematica Balkanica*, 23:289–302, 2009.

[286] A. Mucherino, C. Lavor, L. Liberti, and N. Maculan. Strategies for solving distance geometry problems with inexact distances by discrete approaches. In S. Cafieri, E. Hendrix, L. Liberti, and F. Messine, editors, *Proceedings of the Toulouse Global Optimization workshop*, pages 93–96, Toulouse, 2010.

[287] A. Mucherino, C. Lavor, L. Liberti, and E-G. Talbi. On suitable parallel implementations of the branch & prune algorithm for distance geometry. In *Proceedings of the Grid5000 Spring School*, Lille, France, 2010.

[288] A. Mucherino, C. Lavor, L. Liberti, and E-G. Talbi. A parallel version of the branch & prune algorithm for the molecular distance geometry problem. In *ACS/IEEE International Conference on Computer Systems and Applications (AICCSA10)*, Hammamet, Tunisia, 2010. IEEE conference proceedings.

[289] A. Mucherino, L. Liberti, and C. Lavor. `MD-jeep`: an implementation of a branch-and-prune algorithm for distance geometry problems. In K. Fukuda, J. van der Hoeven, M. Joswig, and N. Takayama, editors, *Mathematical Software*, volume 6327 of *LNCS*, pages 186–197, New York, 2010. Springer.

[290] T. Nagata and H. Sasaki. An efficient algorithm for distribution network restoration. In *Power Engineering Society Summer Meeting*, volume 1, pages 54–59. IEEE, 2001.

[291] T. Nagata, H. Sasaki, and R. Yokoyama. Power system restoration by joint usage of expert system and mathematical programming approach. *IEEE Transactions on Power Systems*, 10(3):1473–1479, 1995.

[292] A.K. Nemaa and S.K. Gupta. Optimization of regional hazardous waste management systems: an improved formulation. *Waste Management*, 19:441–451, 1999.

[293] S. Nguyen and S. Pallottino. Hyperpaths and shortest hyperpaths. *In: Combinatorial Optimization, Lecture Notes in Mathematics, Berlin*, 1403:258–271, 1986.

[294] T. Nuortio, J. Kytojokib, H. Niskaa, and O. Braysy. Improved route planning and scheduling of waste collection and transport. *Expert Systems with Applications*, 30:223–232, 2006.

[295] L. Ochoa, L. Cradden, and G. Harrison. Demonstrating the capacity benefits of dynamic ratings in smarter distribution networks. In *IEEE PES Conference on Innovative Smart Grid Technologies (ISGT)*, 2010.

[296] I. Okhrin and K. Richter. Vehicle routing problem with real-time travel times. *International journal of vehicle information and communication systems*, 2(1-2):59–77, 2009.

[297] S. Orlowski, C. Raack, A. M. C. A. Koster, G. Baier, T. Engel, and P. Belotti. Branch-and-cut techniques for solving realistic two-layer network design problems. In A. M. C. A. Koster and Xavier Muñoz, editors, *Graphs and Algorithms in Communication Networks*, chapter 3, pages 95–117. Springer-Verlag, 2009.

[298] Andreas Ostermeier, Andreas Gawelczyk, and Nikolaus Hansen. Step-size adaption based on non-local use of selection information. In *Problem Solving From Nature (PPSN III)*, pages 189–198, 1994.

[299] L. Pallottino, E.M. Feron, and A. Bicchi. Conflict resolution problems for air traffic management systems solved with mixed integer programming. *IEEE Transactions on Intelligent Transportation Systems*, 3(1):3–11, 2002.

[300] A. Palmer. The development of an integrated routing and carbon dioxide emissions model for goods vehicles. Technical report, Cranfield, PhD thesis, 2007.

[301] S.N. Parragh, K.F. Doerner, and R.F. Hartl. A survey on pickup and delivery problems. *part i: Transportation between customers and depot, Journal für Betriebswirtschaft*, 58(1):21–51, 2008.

[302] S.N. Parragh, K.F. Doerner, R.F. Hartl, and X. Gandibleux. A heuristic two-phase solution approach for the multi-objective dial-a-ride problem. *Networks*, 54(4):227–242, 2009.

[303] Mark Pitman and Andrew King. Engineering solutions to optimise the design of carbon-neutral tall office buildings. In *Proc. International Conference on Solutions for a Sustainable Planet*, 2009.

[304] J-Y. Potvin, Y. Xua, and I. Benyahia. Vehicle routing and scheduling with dynamic travel times. *Computers and Operations Research*, 33(4):1129–1137, 2006.

[305] K. V. Price. An introduction to differential evolution. In D. Corne et al., editor, *New Ideas in Optimization*, pages 79–108. McGraw-Hill, 1999.

[306] Regional Wood Energy Development Programme. *Proceedings of the regional expert consultation on modern applications of biomass energy*. Food and Agricultural Organization of the UN, Bangkok, January 1998.

[307] H.N. Psaraftis. A dynamic programming solution to the single vehicle many-to-many immediate request dial-a-ride problem. *Transportation Science*, 14(2):130–154, 1980.

[308] H.N. Psaraftis. Dynamic vehicle routing: status and prospects. *Annals of Operations Research*, 61:143–164, 1995.

[309] R. C. Purshouse. *On the Evolutionary Optimisation of Many Objectives*. PhD thesis, The University of Sheffield, 2003.

[310] R. C. Purshouse and P. J. Fleming. An Adaptive Divide-and-Conquer Methodology for Evolutionary Multi-criterion Optimisation. In C. Fonseca et al., editors, *Conference on Evolutionary Multi-Criterion Optimization (EMO 2003)*, number 2632 in LNCS, pages 133–147. Springer, 2003.

[311] A. K. Qin, V. L. Huang, and P. N. Suganthan. Differential evolution algorithm with strategy adaptation for global numerical optimization. *IEEE Trans. Evol. Comput.*, 13:398–417, 2009.

[312] John Rawls. *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, Massachussets, 1 edition, 1971.

[313] John Rawls. *Political Liberalism*. Harvard University Press, Cambridge, Massachussets, 1993.

[314] John Rawls. *The Laws of Peoples*. Harvard University Press, Cambridge, Massachussets, 1999.

[315] I. Rechenberg. Optimierung technischer Systeme nach Prinzipien der biologischen Evolution Dr.-Ing. Dissertation. Technical report, Verlag Frommann-Holzboog, Stuttgart-Bad Cannstatt, 1973.

[316] J.C. Robbins. Routing hazardous materials shipments. Technical report, Thesis/Dissertation, Indiana Univ.,Bloomington, IN, 1981.

[317] T. Robic and B. Filipic. Demo: Differential evolution for multiobjective optimization. In *Proc. EMO*, pages 520–533, 2005.

[318] B. Roth. Rigid and flexible frameworks. *American Mathematical Monthly*, 88(1):6–21, 1981.

[319] A. Rudkevich. On the supply function equilibrium and its applications in electricity markets. *Decision Support Systems*, 40:409–425, 2005.

[320] N.V. Sahinidis and M. Tawarmalani. *BARON 7.2.5: Global Optimization of Mixed-Integer Nonlinear Programs,* User's Manual, 2005.

[321] S. Sahoo, S. Kim, B-I. Kim, B. Kraas, and A. Popov. Routing optimization for waste management. *Interfaces*, 35(1):24–36, 2005.

[322] R. Santana, P. Larrañaga, and J.A. Lozano. Combining variable neighbourhood search and estimation of distribution algorithms in the protein side chain placement problem. In *Proc. of Mini Euro Conference on Variable Neighbourhood Search, Tenerife, Spain*, 2005.

[323] J.B. Saxe. Embeddability of weighted graphs in *k*-space is strongly NP-hard. *Proceedings of 17th Allerton Conference in Communications, Control and Computing*, pages 480–489, 1979.

[324] A. Sbihi and R.W. Eglese. Combinatorial optimization and green logistics. *A Quarterly Journal of Operations Research*, 5:99–116, 2007.

[325] T. Schlick. *Molecular modelling and simulation: an interdisciplinary guide*. Springer, New York, 2002.

[326] H.-P. Schwefel. *Evolution and Optimum Seeking*. Sixth-Generation Computer Technology Series. John Wiley & Sons, Inc., New York, 1995.

[327] C. Shannon and D. Hagelbarger. Concavity of resistance functions. *Journal of Applied Physics*, 27(1):42–43, 1956.

[328] H.D. Sherali and W.P. Adams. *A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems*. Kluwer Academic Publishers, Dodrecht, 1999.

[329] H.D. Sherali and A. Alameddine. A new reformulation-linearization technique for bilinear programming problems. *Journal of Global Optimization*, 2:379–410, 1992.

[330] D. Shobrys. A model for the selection of shipping routes and storage locations for a hazardous substance. Technical report, Ph.D Dissertation, John Hopkins University, Baltimore, 1981.

[331] A. Sit, Z. Wu, and Y. Yuan. A geometric build-up algorithm for the solution of the distance geometry problem using least-squares approximation. *Bulletin of Mathematical Biology*, 71:1914–1933, 2009.

[332] R.A. Sivakumar, R. Batta, and M.H. Karwan. A multiple route conditional risk model for transporting hazardous materials. *INFOR*, 33:20–33, 1995.

[333] E. Smith and C. Pantelides. A symbolic reformulation/spatial branch-and-bound algorithm for the global optimisation of nonconvex MINLPs. *Computers & Chemical Engineering*, 23:457–478, 1999.

[334] J. Söderman and F. Pettersson. Structural and operational optimisation of distributed energy systems. *Applied Thermal Engineering*, 26:1400–1408, 2006.

[335] A. Soylu, C. Oruç, M. Turkay, K. Fujita, and T. Asakura. Synergy analysis of collaborative supply chain management in energy systems using multi-period milp. *European Journal of Operational Research*, 174:387–403, 2006.

[336] Richard K. Strand, Drury B. Crawley, Curtis O. Pedersen, Richard J. Liesen, Linda K. Lawrie, F. C. Winkelmann, W. F. Buhl, Y. J. Huang, and Daniel E. Fisher. EnergyPlus: A new-generation energy analysis and load calculation engine for building design. In *Proc. Association of Collegiate Schools of Architecture Technology Conference*, 2000.

[337] K.C. Tan, Y.H. Chew, and L.H. Lee. A hybrid multi-objective evolutionary algorithm for solving truck and trailer vehicle routing problems. *European Journal of Operational Research*, 172:855–885, 2006.

[338] E. Taniguchi and H. Shimamoto. Intelligent transportation system based dynamic vehicle routing and scheduling with variable travel times. *Transportation Research Part C: Emerging Technologies*, 12(3-4):235–250, 2004.

[339] Z. Tarapata. Selected multicriteria shortest path problems : An analysis of complexity, models and adaptation of standard algorithms. *International Journal Of Applied Mathematics And Computer Science*, 17:269–287, 2007.

[340] M. Tawarmalani and N.V. Sahinidis. Exact algorithms for global optimization of mixed-integer nonlinear programs. In P.M. Pardalos and H.E. Romeijn, editors, *Handbook of Global Optimization*, volume 2, pages 65–86. Kluwer Academic Publishers, Dordrecht, 2002.

[341] M. Tawarmalani and N.V. Sahinidis. Global optimization of mixed integer nonlinear programs: A theoretical and computational study. *Mathematical Programming*, 99:563–591, 2004.

[342] O. Teytaud and S. Gelly. General lower bounds for evolutionary algorithms. In *Conference on Parallel Problem Solving from Nature (PPSN 2006)*, volume 4193, pages 21–31. Springer, 2006.

[343] O. Teytaud and S. Gelly. DCMA, yet another derandomization in covariance-matrix-adaptation. In D. Thierens et al., editors, *Genetic and Evolutionary Computation Conference (GECCO)*, pages 955–922. ACM Press, 2007.

[344] O. Teytaud, S. Gelly, and J. Mary. On the Ultimate Convergence Rates for Isotropic Algorithms and the Best Choices Among Various Forms of Isotropy . In T. P. Runarsson et al., editors, *Conference on Parallel Problem Solving from Nature (PPSN IX)*, volume 4193 of *LNCS*, pages 32–41. Springer, 2006.

[345] D. Thierens. An adaptive pursuit strategy for allocating operator probabilities. In H.-G. Beyer et al., editor, *Proc. GECCO*, pages 1539–1546. ACM, 2005.

[346] W. Tian and M. Hu. Study of air traffic flow management optimization model and algorithm based on multi-objective programming. *Second International Conference on Computer Modeling and Simulation*, 2:210–214, 2010.

[347] P. Toth and D. Vigo. The vehicle routing problem. *Philadelphia: Society for Industrial and Applied Mathematics*, 2002.

[348] N. Touati-Moungla, P. Belotti, V. Jost, and L. Liberti. On green routing and scheduling problem. Technical report, LIX, École Polytechnique, 2010.

[349] DV. Tung and A. Pinnoi. Vehicle routing???scheduling for waste collection in hanoi. *European Journal of Operational Research*, 125(3):449–468, 2000.

[350] M. Turnquist. Multiple objectives, uncertainty and routing decisions for hazardous materials shipments. In *Proceedings of the 5th International Conference on Computing in Civil and Building Engineering ASCE, New York*, pages 357–364, 1993.

[351] Jose M. Valerio de Carvalho. Using extra dual cuts to accelerate column generation. *Informs Journal on Computing*, 17(2):175–182, 2005.

[352] P. Vontobel and H. Loeliger. On factor graphs and electrical networks. In J. Rosenthal and D.S. Gilliam, editors, *Mathematical Systems Theory in Biology, Communication, Computation, and Finance*, volume IMA, pages 469–492. Springer, New York, 2003.

[353] T. Wagner, N. Beume, and B. Naujoks. Pareto-, Aggregation-, and Indicator-based Methods in Many-objective Optimization. In S. Obayashi et al., editors, *Conference on Evolutionary Multi-Criterion Optimization (EMO 2007)*, volume 4403 of *LNCS*, pages 742–756, Berlin Heidelberg, Germany, 2007. Springer. extended version published as internal report of Sonderforschungsbereich 531 Computational Intelligence CI-217/06, Universität Dortmund, September 2006.

[354] Tobias Wagner, Heike Trautmann, and Boris Naujoks. Ocd: Online convergence detection for evolutionary multi-objective algorithms based on statistical testing. In Matthias Ehrgott, Carlos Fonseca, Xavier Gandibleux, Jin-Kao Hao, and Marc Sevaux, editors, *Evolutionary Multi-Criterion Optimization*, volume 5467 of *Lecture Notes in Computer Science*, pages 198–215. Springer Berlin / Heidelberg, 2009.

[355] H. Wang, C. Murillo-Sánchez, R. Zimmermann, and R. Thomas. On computational issues of market-based optimal power flow. *IEEE Transactions on Power Systems*, 22(3):1185–1193, 2007.

[356] Y.W Wang. An optimal location choice model for recreation-oriented scooter recharge stations. *Transportation Research Part D: Transport and Environment*, 12(3):231–237, 2007.

[357] Y.W Wang. Locating battery exchange stations to serve tourism transport. *Transportation Research Part D: Transport and Environment*, 13(3):193–197, 2008.

[358] Y.W Wang and C.C Lin. Locating road-vehicle refueling stations. *Transportation Research Part E: Logistics and Transportation Review*, 45(5):821–829, 2009.

[359] Y.W. Wang and C.R. Wang. Locating passenger vehicle refueling stations. *Transportation Research Part E: Logistics and Transportation Review*, 46(5):791–801, 2010.

[360] WBCSD. Transforming the market: energy efficiency in buildings. Technical report, World Business Council for Sustainable Development, 2009.

[361] J. Whitacre, T. Pham, and R. Sarker. Use of statistical outlier detection method in adaptive evolutionary algorithms. In *Proc. GECCO*, pages 1345–1352. ACM, 2006.

[362] A.B. Wijeratne, M.A. Turnquist, and P.B. Mirchandani. Multiobjective routing of hazardous materials in stochastic networks. *European Journal of Operational Research*, 65:33–43, 1993.

[363] H.P. Williams. *Model Building in Mathematical Programming*. Wiley, Chichester, 4th edition, 1999.

[364] T. Van Woensel, L. Kerbache, H. Peremans, and N. Vandaele. A queueing framework for routing problems with time-dependent travel times. *Journal of Mathematical Models and Algorithms, Special Issue "Quantitative aspects of transportation and logistics"*, 6(1):151–173, 2007.

[365] T. Van Woensel, L. Kerbache, H. Peremans, and N. Vandaele. Vehicle routing with dynamic travel times : A queueing approach. *European journal of operational research*, 186(3):990–1007, 2008.

[366] L.A. Wolsey. *Integer Programming*. Wiley, New York, 1998.

[367] Jonathan A. Wright, Heather A. Loosemore, and Raziyeh Farmani. Optimization of building thermal design and control by multi-criterion genetic algorithm. *Energy and Buildings*, 34(9):959 – 972, 2002.

[368] D. Wu and Z. Wu. An updated geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. *Journal of Global Optimization*, 37:661–673, 2007.

[369] D. Wu, Z. Wu, and Y. Yuan. Rigid versus unique determination of protein structures with geometric buildup. *Optimization Letters*, 2(3):319–331, 2008.

[370] F. Wu and A. Monticelli. Analytical tools for power system restoration — conceptual design. *IEEE Transactions on Power Systems*, 3(1):10–16, 1988.

[371] V. Zavala, E. Constantinescu, and M. Anitescu. Economic impacts of advanced weather forecasting on energy system operations. In *Power and Energy Society: Conference on Innovative Smart Grid Technology*, pages 1–8. IEEE, 2010.

[372] V. Zavala, J. Wang, S. Leyffer, E. Constantinescu, M. Anitescu, and G. Conzelmann. Proactive energy management for next-generation building systems. In *SimBuild*, 2010.

[373] J. Zhang and A. C. Sanderson. Self-adaptive multi-objective differential evolution with direction information provided by archived inferior solutions. In *Proc. CEC*, pages 2806–2815. IEEE, 2008.

[374] J. Zhang and A. C. Sanderson. JADE: Adaptive differential evolution with optional external archive. *IEEE Trans. Evol. Comput.*, 13:945–958, 2009.

[375] K.Q. Zhu and K.-L. Ong. A reactive method for real time dynamic vehicle routing problems. In *In Proceedings of 12th IEEE Internationals Conference on Tools with Artificial Intelligence*, pages 176–180. IEEE Computing Society, Los Alamitos, USA, 2000.

[376] E. Zitzler, D. Brockhoff, and L. Thiele. The Hypervolume Indicator Revisited: On the Design of Pareto-compliant Indicators Via Weighted Integration. In S. Obayashi et al., editors, *Conference on Evolutionary Multi-Criterion Optimization (EMO 2007)*, volume 4403 of *LNCS*, pages 862–876, Berlin, 2007. Springer.

[377] E. Zitzler, K. Deb, and L. Thiele. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evol. Comput.*, 8:173–195, 2000.

[378] E. Zitzler, K. Deb, and L. Thiele. Comparison of Multiobjective Evolutionary Algorithms: Empirical Results. *Evolutionary Computation*, 8(2):173–195, 2000.

[379] E. Zitzler and S. Künzli. Indicator-Based Selection in Multiobjective Search. In X. Yao et al., editors, *Conference on Parallel Problem Solving from Nature (PPSN VIII)*, volume 3242 of *LNCS*, pages 832–842. Springer, 2004.

[380] E. Zitzler, M. Laumanns, and L. Thiele. SPEA2: Improving the strength pareto evolutionary algorithm for multiobjective optimization. In K.C. Giannakoglou et al., editor, *Evolutionary Methods for Design, Optimisation and Control with Application to Industrial Problems*, pages 95–100. CIMNE, 2002.

[381] E. Zitzler and L. Thiele. Multiobjective Optimization Using Evolutionary Algorithms - A Comparative Case Study. In *Conference on Parallel Problem Solving from Nature (PPSN V)*, volume 1498 of *LNCS*, pages 292–301, Amsterdam, 1998.

[382] E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE Trans. Evol. Comput.*, 3:257–271, 1999.

[383] E. Zitzler, L. Thiele, and J. Bader. On Set-Based Multiobjective Optimization. *IEEE Transactions on Evolutionary Computation*, 14(1):58–79, 2009.

[384] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. Grunert da Fonseca. Performance Assessment of Multiobjective Optimizers: An Analysis and Review. *IEEE Transactions on Evolutionary Computation*, 7(2):117–132, 2003.

[385] Eckart Zitzler and Lothar Thiele. Multiobjective optimization using evolutionary algorithms - a comparative case study. In *Proceedings of the 5th International Conference on Parallel Problem Solving from Nature*, PPSN V, pages 292–304, London, UK, 1998. Springer-Verlag.

[386] K.G. Zografos and K.N. Androutsopoulos. A heuristic algorithm for solving hazardous materials distribution problems. *European Journal of Operational Research*, 152(2):507–519, 2004.

[387] K.G. Zografos and C.F. Davis. Multiobjective programming approach for routing hazardous materials. *Journal of Transportation Engeneering*, 115(6):661–673, 1989.