Convergence analysis and novel algorithms in multi-objective optimization

Eugénie Marescaux

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAX105

Thèse de doctorat

Convergence analysis and novel algorithms in bi-objective optimization

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques

Thèse présentée et soutenue à Palaiseau, le 21 novembre 2022, par

EUGÉNIE MARESCAUX

Composition du Jury :

Stéphane Gaubert Directeur de recherche, Ecole Polytechnique (Tropical team)	Président du jury
Carlos Fonseca Professeur adjoint, University of Coimbra (Informatics Engineering)	Rapporteur
Kathrin Klamroth Professeure, Wuppertal University (Mathematics and Natural Sciences)	Rapporteure
Matthias Ehrgott Professeur, Lancaster University (Management Science)	Examinateur
Tobias Glasmachers Professeur, Ruhr-Universität Bochum (Institut für Neuroinformatik)	Examinateur
Anne Auger Directrice de recherche, Ecole Polytechnique & Inria Saclay (RandOpt team)	Directrice de thèse

Summary

Optimization is the field of applied mathematics concerned with minimizing (or maximizing) one or more objective functions. It has many applications in science and industry, from scheduling power outages through designing cars to describing physical phenomena. Multi-objective optimization aims at finding a good approximation of the Pareto set, that is the set of feasible solutions that cannot be improved in all objectives simultaneously, and the Pareto front, its image in the objective space. In this Ph.D. thesis, we seek to improve the understanding of the convergence speed of bi-objective optimization algorithms towards the entire Pareto front. We rely on the hypervolume, a widely used set quality indicator, to quantify the optimality gap between the Pareto front and an approximation provided by an algorithm. The computational cost is typically measured by the number of evaluations of the objective functions.

First, we derive a theoretical upper bound on the speed of convergence. We prove that for a wide class of Pareto fronts, the smallest optimality gap associated with a set of n points is larger than $1/(n+1)$ multiplied by a constant. The constant depends on the Pareto front and the reference point used in the definition of the hypervolume. This result shows that the speed of convergence is sublinear at best, which is worse than the convergence rates that can be achieved by single-objective optimization algorithms.

Then, we introduce an algorithmic framework, HyperVolume-Based Incremental Single-Objective Optimization for Multi-Objective Optimization (HV-ISOOMOO). A meta-iteration of HV-ISOOMOO corresponds to the solving of a single-objective optimization subproblem. We provide lower bounds on the convergence speed of the Pareto front approximation formed by the solutions returned by the single-objective solver (which we call the *final incumbents Pareto front approximation*) under the assumption that the solver returns a global optimum with infinite precision. For convex Pareto fronts, this ideal algorithm reaches the best achievable convergence rate: $\Theta(1/n)$.

Finally, we provide an implementation of HV-ISOOMOO coupled with CMA-ES, a state-of-the-art single-objective optimization algorithm. We call it Multi-Objective Covariance Matrix Adaptation 2 (MO-CMA-ES-2). The MO-CMA-ES-2 algorithm presents state-of-the-art performance. On a simple convex problem, we empirically analyze the evolution of the optimality gap of its final incumbents Pareto front approximation with respect to the number of meta-iterations of HV-ISOOMOO and iterations of CMA-ES.

Résumé

L'optimisation est le domaine des mathématiques appliquées qui s'intéresse à la minimisation (ou la maximisation) d'une ou plusieurs fonctions objectif. Elle a de nombreuses applications industrielles et scientifiques, de la planification des pannes de courant à la conception de voitures en passant par la description de phénomènes physiques. L'optimisation multi-objectifs s'intéresse à l'approximation de l'ensemble de Pareto, c'est-à-dire l'ensemble des solutions admissibles qui ne peuvent être améliorées suivant tous les objectifs simultanément, et du front de Pareto, son image dans l'espace des objectifs. Dans cette thèse de doctorat, nous cherchons à approfondir les connaissances sur la vitesse de convergence d'algorithmes d'optimisation multi-objectifs vers la totalité du front de Pareto. Nous nous appuyons sur l'hypervolume, un indicateur de qualité d'ensemble largement utilisé, pour quantifier l'écart entre le front de Pareto et une approximation fournie par un algorithme, autrement dit l'erreur. Le coût est mesuré par le nombre de fois où les fonctions objectif ont été évaluées.

Tout d'abord, nous prouvons une borne supérieure théorique sur la vitesse de convergence. Nous démontrons que pour une vaste catégorie de fronts de Pareto, la plus petite erreur associée à une approximation du front de Pareto composée de n solutions est supérieure à $1/(n+1)$ multiplié par une constante. Cette constante dépend du front de Pareto et du point de référence utilisé pour définir l'hypervolume. Cela garantit que la vitesse de convergence est au mieux sous-linéaire, soit plus lente que les vitesses de convergence qui peuvent être atteintes par des algorithmes d'optimisation mono-objectifs.

Ensuite, nous définissons une nouvelle classe d'algorithmes, HV-ISOOMOO. Une méta-itération de HV-ISOOMOO correspond à la résolution d'un sous-problème d'optimisation mono-objectif. Les solutions renvoyées par le solveur mono-objectif fournissent une approximation du front de Pareto. Nous démontrons des bornes inférieures sur la vitesse de convergence de cette approximation vers le front de Pareto en supposant que le solveur mono-objectif fournisse toujours un optimum global avec exactitude. Pour les fronts de Pareto convexes, cet algorithme idéal a une vitesse de convergence optimale: $\Theta(1/n)$.

Finalement, nous détaillons une implémentation de HV-ISOOMOO qui utilise le solveur mono-objectif CMA-ES, MO-CMA-ES-2. Cet algorithme se compare favorablement aux meilleurs algorithmes multi-objectifs connus à ce jour. Sur un problème convexe simple, nous analysons empiriquement l'évolution de l'erreur en fonction du nombre de méta-itérations de MO-CMA-ES-2 et du nombre d'itérations de CMA-ES.

Remerciements

Ces travaux s'inscrivent dans ceux d'une communauté scientifique, présente et passée.

Tout d'abord, je remercie Anne, qui a accepté d'encadrer ma thèse. Merci de m'avoir fait découvrir des travaux de recherche qui ont nourri ma réflexion et ma sensibilité mathématique et de m'avoir proposé des problématiques porteuses. Ce manuscrit doit aussi beaucoup aux deux autres membres permanents de l'équipe Randopt: Dimo et Niko. *Many ideas from this manuscript owe to our frequent discussions. I am especially grateful to Niko for his regular scientific support, which inspired me to be more rigorous.* L'équipe Randopt, ce sont aussi Paul, Cheikh, Kostas, Marie-Ange et Alann, et ceux qui se sont joint à nous vers la fin, Armand et Mohamed.

Les discussions scientifiques que j'ai eu l'occasion d'avoir avec d'autres doctorants, que ce soit au CMAP ou sur le discord des doctorants m'ont beaucoup apporté. Le CMAP est très accueillant et m'a abondamment fournie en joie et en camaraderie. L'équipe administrative nous fournit aussi un soutien remarquable, réalisant l'exploit de rendre les démarches administratives agréables. Un merci tout particulier à Nassera. Merci à Clément M., Louis, Constantin, Jessie, Manon, Apolline, et plus généralement à tous les membres du CMAP avec qui j'ai eu le plaisir de discuter.

Le bureau 2015 a vraiment été mon foyer pendant ces trois ans. A droite, Dominik, toujours prêt à écouter mes dernières idées et à présenter les concepts les plus abstraits qui lui passent sous la main sans me tenir rigueur de décrocher. Fournisseur culinaire officiel en toffifee et fin connaisseur en confit de canard. En face, l'ex-toulousaine Solange, un rayon de soleil plein de générosité et de curiosité toujours prête à se battre pour ce qui lui tient à coeur. Sur sa rangée, de nouvelles têtes se sont rajoutées au fil des années: Ignacio et Michael. En poursuivant à la droite de Dominik, on trouve Corentin, source inépuisable de projets éclectiques et mine d'informations en tout genre et d'aide pratique. Au bout, Claire, cuisinière hors pair et partenaire de discussion enflammée. J'oubliais le porte-manteau et son fidèle destrier, Morfale. Digne mascotte d'un tel bureau.

Une thèse comporte souvent des moments difficiles, et celle-ci n'a pas été épargnée. J'ai eu la chance de pouvoir m'appuyer sur mes amis et ma famille pour alléger mon

fardeau et voir d'un œil nouveau les problèmes rencontrés. Merci à vous. Je tiens à remercier tout particulièrement Yann, pour sa bienveillance et ses retournements de situation, pour être la preuve vivante qu'on peut vivre autrement ; Antonin, pour m'avoir fait découvrir la guitare et pour sa présence paisible ; Morgane, Marie-Claire, Isabelle et Eléonore, pour être mes amies depuis plus de dix ans, malgré la distance ; Lesly, parce que les quelques fois où on se voit chaque année m'apportent beaucoup de joie. Et enfin, le fin des fins et la fin des faims, Dara. Je pourrais te remercier de tant de choses et tu les connais déjà toutes.

À mon grand-père Philippe

Contents

List of Symbols and Abbreviations

A-PFA Archive Pareto Front Approximation

`bbob-biobj` The Bi-objective Black-Box Optimization Benchmarking test suite

COMO-CMA-ES COmma-selection Multi-Objective CMA-ES

FI-PFA Final Incumbents Pareto Front Approximation

FM-PFA Final Means Pareto Front Approximation

HV-ISOOMOO HyperVolume based Incremental Single-Objective Optimization for Multi-Objective Optimization

HMO-CMA-ES Hybrid Multi-Objective CMA-ES

MO-CMA-ES Multi-Objective CMA-ES

MO-CMA-ES-2 Multi-Objective Optimization CMA-ES 2

PFA Pareto Front Approximation

UP-MO-CMA-ES Unbounded Population Multi-Objective CMA-ES

CMA-ES Covariance Matrix Adaptation Evolutionary Strategy

COCO COmparing Continuous Optimizers (a benchmarking platform)

EA Evolutionary algorithm

ECDF Empirical Cumulative Distribution Function

MOEA Multi-Objective Evolutionary Algorithm

CHAPTER 1

Introduction

Optimization is the domain of applied mathematics that addresses the problem of finding the minimum (or maximum) of one or more objective functions. We focus on problems that can be formulated as follows:

$$\min_{X \in \Omega} F_1(X), \ldots, F_m(X) \ , \tag{1.1}$$

where Ω is typically a subset of a vector space of finite dimension and the objective functions F_i are functions from Ω to $\mathbb{R}$. The letter m denotes the number of objectives. Such a problem is called a *single-objective optimization* problem when $m = 1$ and a *multi-objective optimization* problem otherwise.

In this Ph.D. thesis, the main focus is on bi-objective optimization but single-objective optimization problems and algorithms will also be discussed. To make the distinction clearer, the unique objective function of a single-objective problem is later denoted by h.

1.1 Single-objective optimization

The idea of looking for a minimum or a maximum is quite old and dates back to antiquity. At that time, calculus did not exist yet, and this problem was studied from a geometrical perspective. A well-known example is the elementary theorem of Euclid's geometry stating that the smallest path between two points is a straight line. Another classic example took place a few centuries later when Heron used a geometrical argument to prove that the path of light when reflected on a surface is the shortest one [25, p261-265]. The geometrical approach is limited to problems with a geometric interpretation and lacks generality.

It was only in the 17th century that the invention of calculus by Leibniz and Newton gave rise to systematic methods for solving optimization problems. The potential of

calculus for optimization was apparent from the very beginning. It is meaningful that the title of the first article about calculus is *Nova Methodus pro Maximis et Minimis*, which can be translated as "New method for maxima and minima" [60]. In this article, Leibniz discusses the relationship between derivatives and optima. Though it was not formulated this way at the time, this pioneer work already contained in spirit the basic result that the derivatives of a function at its minima and maxima are null. For simple differentiable functions, this necessary condition and a basic analysis suffice to obtain an analytical expression of a global optimum.

Closer to us, the development of computers in the 20th century paved the way for a further drastic expansion of the range of optimization problems we are able to solve at least approximatively. The solution of many analytically unsolvable problems can be approximated by relying on a computer to automatically execute an algorithm. Optimization benefitted from both the exponential growth of computing power and the development of a wide variety of algorithms. Objective functions without derivatives or without a known analytical formula are no longer insurmountable obstacles. Handling such problems is the focus of derivative-free and blackbox optimization [4].

Applications of optimization are hardly limited to the few examples in the field of optics that we mentioned earlier. They are abundant in physics. The two examples discussed above are special cases of the more general principle of least action, which was introduced by Maupertuis in the 18th century. This principle allows to formulate many laws of physics, from mechanics to quantum theory, as the minimization of a quantity called *action* [69]. Furthermore, a number of concrete natural phenomena are the product of natural selection of habits or genetics. Such phenomena can be considered as approximate solutions of optimization problems. This is the case, for instance, of the form of honeycombs [44]. Last but not least, optimization is used extensively in industry, particularly in industrial design. Back in the 18th century, MacLaurin, the developer of the first complete theory of minima and maxima depending on the sign of derivatives, exploited optimization to design ships, watermills and windmills [44]. Nowadays, optimization algorithms are used in a wide range of engineering applications, from the design of cars [35] to the scheduling of nuclear reactor outages [56]. This quick overview reveals that the problem of finding the minimum (or the maximum) of a certain quantity appears in many places. Optimization is useful both in science, to understand and describe natural phenomena, and in industry.

1.2 Multi-objective optimization

Multi-objective optimization is generally considered to have been born in the end of the 19th century with the pioneer works of Edgeworth and Pareto: *Mathematical Psychics* (published in 1881) [37] and *Manual of Political Economy* (published in 1906 and translated in English in 1971) [66]. Edgeworth and Pareto were both economists. The field benefited from the previous theoretical and algorithmic developments made in the context of single-objective optimization.

Like its single-objective counterpart, multi-objective optimization has many appli-

cations in a variety of fields: from Economy, its original field, [23] through Ecology [55] to Neuroscience [34]. Multi-objective optimization is also particularly helpful for engineering. Engineering problems often involve more than one interesting quantity whose minimization or maximization is desirable. When designing a new product, a low cost of production, a high quality, a large impact resistance and a reasonably low mass are all frequent goals that conflict with each other. For example, [35] presents a multi-objective car design problem involving the minimization of both the total mass of the car and the intrusion into the passenger cabin during a crash. Multi-objective optimization is also at play in machine learning. Indeed, machine learning seeks to provide the most accurate model possible while avoiding overfitting, an intrinsicly multi-objective problem [52].

Problem solving and decision making are easier with only one objective at stake. It is therefore tempting to turn a multi-objective problem into a single-objective one, focusing either on the most important objective or on a combination of objectives. However, it would come at the cost of erasing the intrinsic complexity of the decision. A multi-objective approach, while more complex, allows an informed choice that takes into account several important aspects of the decision simultaneously.

The core concept of multi-objective optimization is *domination*. We say that one solution dominates another if it is not worse along all objectives and better along at least one. If among two solutions, none dominates the other, they are said to be *incomparable*. Domination can also apply to sets of solutions. We say that a set of solutions dominates another set when every solution of the latter is dominated by at least one solution of the former. Multi-objective optimization is interested in the *Pareto set*, the set of all non-dominated feasible solutions, and the *Pareto front*, its image in the objective space. The Pareto set and front respectively gather the solutions and the compromises that could be chosen by a rational person who only cares about the objectives taken into consideration.

In the early days, the goal of multi-objective optimization was to find a single vector of the Pareto set, an *efficient solution*. Now, the focus has shifted to finding a good approximation of either the Pareto set or the Pareto front. When looking for the Pareto front, we are satisfied with one efficient solution per sound compromise. On the contrary, looking for the Pareto set means that we are concerned with all efficient solutions. Depending on the application, the knowledge of redundant efficient solutions is of varying interest.

The most fundamental problem that arises when the number of objectives exceeds one is the absence of a natural order on the solutions or solution sets. Many solutions and sets of solutions are incomparable with respect to domination. This raises the practical issue of how to guide the algorithms and measure the quality of the solutions or sets of solutions they return.

An option is to rely on a *set quality indicator* that assigns a real value to any set of solutions. Dozens of quality indicators have been developed [80, 2]. Except for those based on cardinality, which are not very informative, they are typically built to favor sets of solutions whose image is close to the Pareto front, or offer a better coverage of the Pareto front [2]. While indicators often assign the same value to many different sets, it is

rare that a non-discrete indicator assigns the same value to two distinct sets of solutions generated by one or more algorithms. Among the most popular indicators are the Inverted Generational Distance (IGD), the averaged Hausdorff distance and the family of ϵ-indicators, all of which measure closeness between the assessed set and the Pareto front [2]. In this manuscript, we focus on the hypervolume, another popular set quality indicator that favors sets whose image is close to the Pareto front and well-distributed over the Pareto front. Unlike the indicators mentioned above, the hypervolume does not rely on the knowledge of the Pareto front. We will cover this indicator in more details later in the introduction.

We develop an algorithm that is in line with the tradition of *Multi-Objective Evolutionary Algorithms* (MOEAs). *Evolutionary algorithms* (EAs) originally form a class of single-objective algorithms that provides a good foundation for developing multi-objective optimization algorithms. These are stochastic algorithms inspired by the biological process of natural selection, that evolve a population of solutions. The evolution of the population is driven by a *fitness function*, which attributes a value to any sampled solution. MOEAs were developed in the 90s. They follow the same framework as EAs but their *fitness function*, which guides the evolution of the population, is designed to handle multi-objective problems. With the proper guidance, the population can evolve to form a good Pareto set approximation. The algorithm we develop is not a multi-objective EA per se. However, it is a multi-objective optimization algorithm that relies on an EA, CMA-ES, to solve single-objective subproblems.

The MOEAs are diverse. Emmerich et al. separate them into three categories: Pareto, decomposition, and indicator-based, which differ in the way the fitness is built [39]. Pareto-based algorithms first iteratively rank the solutions using domination. Solutions that have the same domination rank are then ranked according to some criterion that captures their contribution to diversity. It is the principle behind one of the best-known MOEAs, NSGA-II [31]. Decomposition-based algorithms decompose the multi-objective problem into subproblems based on *scalarization*. Scalarization transforms a multi-objective problem in a parametrized class of single-objective problems. With a well-chosen scalarization, the optimal solutions of the scalarized problems are efficient and every vector of the Pareto set is an optimal solution of a scalarized problem. Solutions are attributed dynamically to these subproblems. The NSGA-III [33] and the MOEA/D [68] are two famous decomposition-based algorithms. Finally, indicator-based algorithms simply have a fitness built on set-quality indicators. It is quite natural to base the fitness function of an MOEA algorithm on them. The algorithm we develop is close to this last category of MOEAs.

1.3 Convergence speed

The focus of this manuscript is to investigate the convergence speed of multi-objective optimization algorithms. An algorithm's convergence speed determines the amount of computational resources that must be spent in order to get a set of solutions with a given quality. The knowledge of the convergence speed guides the practical choices of

the algorithm used to solve the problem and the allocated budget. In this section, we lay out the vocabulary and notations necessary to describe the convergence speed and detail the kind of convergence speed we are interested in this manuscript.

1.3.1 Notations and vocabulary

We introduce here the standard notations and vocabulary that we use to describe the convergence speed of algorithms. We define them formally in the context of the convergence of a sequence of real numbers to a limit.

The Bachmann-Landau notations o, O, Θ and Ω allow to compare the convergence speed of sequences converging either to $+\infty$ or to 0. Here, we are exclusively concerned with the latter case. We define them rigorously below.

Definition 1 (Bachmann-Landau notation) *Let f and g be two functions from $\mathbb{N}^*$ to $\mathbb{R}$.*

We say that:

- $f(n) = o(g(n))$ *if for all $\epsilon > 0$, there exists $N \in \mathbb{N}^*$ such that for all $n \geq N$, $f(n) \leq \epsilon \times g(n)$,*
- $f(n) = O(g(n))$ *if there exist $N \in \mathbb{N}^*$ and $C > 0$ such that for all $n \geq N$, $f(n) \leq C \times g(n)$,*
- $f(n) = \Omega(g(n))$ *if there exist $N \in \mathbb{N}^*$ and $C > 0$ such that for all $n \geq N$, $f(n) \geq C \times g(n)$ and*
- $f(n) = \Theta(g(n))$ *if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.*

In this thesis, we regularly refer to $\Omega(1/n)$, $O(1/n)$ and $\Theta(1/n)$ convergence speeds. A typical sequence with a $\Theta(1/n)$ convergence speed is simply $(1/n)_{n\in\mathbb{N}^*}$. For this particular sequence, improving the accuracy by a factor of ten requires to multiply the iteration index by ten. This becomes increasingly expensive as the iteration index n grows.

Convergence in $\Theta(1/n^\alpha)$ with $0 < \alpha$ is a subcategory of *sublinear convergence*. A sequence (u_n) of real numbers is said to *converge linearly* to a limit L when the ratio between $|u_{n+1} - L|$ and $|u_n - L|$ is smaller than a constant $C < 1$ for n large enough. This ensures that, after a while, the accuracy always improves by at least a factor of 10 after $1/log_{10}(1/C)$ iterations. For example, for $C = 0.99$ and $C = 0.999$, it takes respectively 229 and 2301 iterations to ensure an improvement by a factor 10. The smaller the constant C, the faster the convergence. An example of sequence that converges linearly to 0 is $(1/10^n)_{n\in\mathbb{N}^*}$. For this particular sequence, the accuracy improves by a factor of ten with each iteration. Linear convergence is achieved by many single-objective algorithms on easy-to-solve functions. It was recently proven that evolutionary strategies (a subcategory of evolutionary algorithms) with a step-size control mechanism achieve linear convergence speed on a broad class of functions including monotonic transformations of strongly convex functions with Lipschitz continuous gradient [1]. While it has not been

proven theoretically yet, the popular evolutionary strategy CMA-ES that we exploit is also empirically known to achieve linear convergence on convex-quadratic functions. Evolutionary algorithms, as all comparison-based optimization algorithms, cannot converge faster than linearly [72]. However, some non-comparison-based single-objective algorithms are able to reach superlinear convergence speed. For example, Newton's method's rate of local convergence is known to be quadratic on twice differentiable functions whose Hessian is Lipschitz continuous [65, Theorem 3.7]. A sequence (u_n) of real numbers is said to *converge quadratically* to a limit L when the ratio between $|u_{n+1} - L|$ and $|u_n - L|^2$ is smaller than a constant C for n large enough. An example of sequence with quadratic convergence is $(1/10^{2^n})_{n\in\mathbb{N}^*}$.

This categorization of asymptotic convergence speeds informs us on the convergence behavior after some unspecified amount of resources has been spent. In practice, this unspecified amount matters, together with the delay before reaching the asymptotic regime and the constants involved in the definitions. For example, converging linearly as $10^{10} \times 1/2^n$ is worse in the very short term (for $n \leq 55$) than converging sublinearly as $10^{-5} \times 1/n$. From $n = 56$, which is a small number compared to the gap of 10^{15} between the constants involved, the first sequence has caught up.

1.3.2 Local definition of the convergence speed

In this section, we detail the kind of convergence speed we chose to investigate. A convergence speed definition is composed of a progress measure and a cost measure. There are many ways to measure cost and even more to measure progress.

Cost measure: number of function evaluations, iterations or meta-iterations

The cost should reflect in some way the computational resources spent by an algorithm. These resources not only depend on the problem and the algorithm, but they can also vary a lot depending on the hardware, the algorithm implementation, and the computing cost of the objective functions. They can be difficult to analyze theoretically.

Here, we chose to define the cost as the number of objective function evaluations. The number of function evaluations does not depend on the hardware and is well suited to theoretical analysis. The evaluation of an objective function always has a cost, which can be of varying degrees of significance. It is especially important when the computation of the objective function requires a costly simulation or experiment. This is the case in the realistic benchmark problem given in [35], where the computation of the objective function requires a crash simulation whose computational time lies between 10 and 20h on 8 CPUs. In the more general case, the number of function evaluations is a cost to consider among others. For example, the significant gain in time that can be obtained by parallelization is not reflected by the number of function evaluations.

We also consider the convergence speed of algorithms with respect to the number of iterations or meta-iterations (in the case of embedded algorithms).

Progress measure : hypervolume indicator

We rely on the hypervolume $\mathrm{HV}_r()$, a popular set-quality indicator, to measure the quality of the sets of solutions provided by algorithms. The *optimality gap* associated to a set S is designed as $\mathrm{HV}_r(\mathrm{PF}_f) - \mathrm{HV}_r(S)$. Progress corresponds to a quality gain, or equivalently to a decrease of the optimality gap.

The hypervolume measures the size of the region of the objective-space dominated by a set of solutions (and dominating a reference point r), see Figure 1.1. This favors

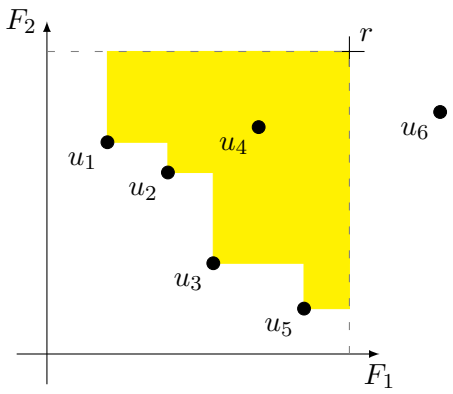


Figure 1.1: Illustration of the hypervolume indicator. The hypervolume with respect to r of the set $\{u_1, \ldots, u_6\}$ is the area of the yellow region.

sets close to the Pareto front and well-distributed. To some extent, theoretical results shed light over the hypervolume preferences. The asymptotic density of the image of n-optimal distributions, that is sets of n solutions with the highest hypervolume, has been proven to be maximal in parts where the slope is -1 [8]. This shows that the hypervolume favors sets with a sparser distribution in parts of the Pareto front where small sacrifice in one objective offers a particularly big reward in the other. Such a distribution can be more informative than an uniform distribution for decision-makers. Indeed, they are likely to choose to make the rewarding sacrifice and select a compromise between the objectives which is not in such parts of the Pareto front.

The hypervolume is a popular indicator. It is part of the fitness of many well-known MOEAs, among which SMS-EMOA [15], HypE [11], the hypervolume-based IBEA [82], POSEA [77] and COMO-CMA-ES [74]. This makes it particularly suitable for studying the performance of these algorithms. It is already used for algorithms benchmark. The quality indicator used in the bi-objective test suites of the benchmarking platform COCO [22] is based on the hypervolume. It is also available in other benchmarking tools such as Nevergrad [12].

Strengths and weaknesses of the hypervolume indicator

As mentioned before, the hypervolume preferences over sets of solutions are interesting for a decision-maker. On top of that, this indicator is easy to apprehend and can be

computed without the knowledge of the Pareto front. The popularity of the hypervolume indicator is largely due to its strict Pareto compliance, meaning that a set of solutions that dominates another always have a strictly higher hypervolume. All known strictly Pareto compliant indicators are built on the hypervolume. This has been considered to be an important desirable property for a long time, but its practical relevance is still under investigation [40].

The hypervolume has two main drawbacks, which are only real drawbacks for more than two objectives. First, it requires setting a reference point. In practice, finding a reference point that does not conceal interesting solutions is relatively easy. Its coordinates can be set to the worst values the decision-maker is interested in or to upper bounds of the objective values of non-dominated solutions. However, the choice of the reference point is not a trivial one and impacts the preference over sets of solutions of the indicator. With more than two objectives, a reference point too far away sometimes lead to a really poor preference, where extremes - which are infinite in number - are over-favored [71]. Secondly, computing the hypervolume is a **#P**-hard problem in m, the number of objectives [17]. It implies that the hypervolume cannot be computed in a time polynomial in m (as long as **P**$\neq$ **NP**). For two and three objectives, the time complexity of computing the hypervolume of a set of n points is $O(n \log(n))$ [45, 13]. For more than 3 objectives, the best algorithm known as a time complexity of $O(n^{m/3}\text{polylog}(n))$ [24, 45]. Fortunately, approximations of the hypervolume can be computed in a more reasonable time. For example, a fully polynomial-time randomized approximation scheme was introduced by Bringmann and Friedrich in 2010 [17]. Developing efficient ways to compute good approximations of the hypervolume for more than three objectives is an active topic of the field [71].

1.4 Results

The results presented in the present manuscript provide answers to the following question:

> How fast can a bi-objective optimization algorithm converge to the entire Pareto front?

We use this question as a common thread.

1.4.1 A theoretical upper bound on the convergence speed

First, we address this question from a purely theoretical perspective in Chapter 3:

> Consider a fixed bi-objective optimization problem. How does the smallest optimality gap of a set of n solutions evolve with n?

Sets of n feasible solutions with the largest hypervolume are called *n-optimal distributions*. In Chapter 3, we prove for a wide class of Pareto fronts that the convergence speed of the optimality gap of n-optimal distributions is $\Omega(1/n)$. A direct consequence of the upper bound is that the convergence speed of bi-objective optimization algorithms

is also $\Omega(1/l)$ on problems for which the Pareto front satisfies the relevant assumptions, with l being the number of solutions evaluated so far by the algorithm. In particular, this statement holds true when both objective functions are convex quadratic [73], a class of problems that is among the easiest to solve. An $\Omega(1/l)$ convergence speed is sub-linear and radically slower than the convergence speeds achievable on convex quadratic objective functions by single-objective optimizers.

We also study empirically the convergence speed of the optimality gap of n-optimal distributions. For six benchmark Pareto fronts, we compute a n-optimal distribution up to $n = 1000$. For the values of n higher than 10, the optimality gap of the computed n-optimal distribution proves to be close to a constant times $1/n$.

1.4.2 An algorithm framework with the potential to reach optimal convergence speed

In Chapter 4, we introduce the algorithm framework HV-ISOOMOO that relies on a greedy approach to build an approximation of the entire Pareto front. The acronym HV-ISOOMOO stands for HyperVolume based Incremental Single-Objective Optimization for Multi-Objective Optimization. A *meta-iteration* of HV-ISOOMOO corresponds to a full run of a single-objective optimization solver, which itself can be decomposed into iterations. We define the *final incumbents Pareto front approximation* (FI-PFA) as the non-dominated subset of the set of the final incumbents of all the single-objective solver runs done so far. At each meta-iteration, the goal is to find a solution with the largest hypervolume contribution to the final incumbents Pareto front approximation.

The coupling of the HV-ISOOMOO framework with a perfect single-objective solver which always return an exact global optimum is called *ideal HV-ISOOMOO* for short. Halfway between theory and practice, we investigate the following question:

> Consider a fixed bi-objective optimization problem. How fast does the sequence of the final incumbents Pareto front approximation of the ideal HV-ISOOMOO algorithm converge to the entire Pareto front ?

We prove three main results regarding the convergence speed of the ideal HV-ISOOMOO. We prove that when the Pareto front is convex, the convergence speed of the optimality gap of the FI-PFA of the ideal HV-ISOOMOO is $O(1/n)$, with n being the number of meta-iterations. Combining this result to the symmetrical upper bound provided in Chapter 3, it follows that the convergence speed is then $\Theta(1/n)$. A looser lower bound on the ideal HV-ISOOMOO convergence speed is also found for bilipschitz Pareto fronts. Finally, we prove that for a certain class of Pareto fronts, the asymptotic impact of doubling the number of meta-iterations is to halve the optimality gap.

On top of this theoretical analysis, we examine empirically the optimality gaps of the FI-PFA of the ideal HV-ISOOMOO for seven benchmark Pareto fronts, for n up to 1000. The optimality gap of the FI-PFA after n meta-iterations proves to be close to that of n-optimal distributions for all benchmark Pareto fronts. In our experiments, it is always less than 24% larger.

1.4.3 A practical algorithm

We introduce MO-CMA-ES-2, an HV-ISOOMOO algorithm that relies on the CMA-ES algorithm to solve single-objective subproblems. The acronym MO-CMA-ES-2 stands for Multi-Objective Optimization Covariance Matrix Adaptation Evolution Strategy 2. Additional stopping criteria were added to allow a better increase of the precision of the solution returned by the single-objective solver through a MO-CMA-ES-2 run. The CMA-ES are initialized in two distinct ways that share the function evaluations budget evenly: randomly (`random` initialization) and close to promising regions (`best_chv` initialization).

We examine empirically the performance of MO-CMA-ES-2. More precisely, we investigate the following questions:

> Does the MO-CMA-ES-2 reach optimal convergence speed ?

and

> How does MO-CMA-ES-2 compare to other bi-objective optimization algorithms ?

We examine empirically the convergence speed of MO-CMA-ES-2 on the **double-sphere** problem, a well-conditioned convex quadratic problem. The convergence speed of FI-PFA with respect to meta-iterations appears to be in $\Theta(1/n)$ and close to the one of ideal HV-ISOOMOO and n-optimal distributions. The convergence speed with respect to iterations appears to be slightly slower, but resembles $\Theta(1/k^{\alpha})$ with α as large as 0.99 for small search space dimension and with using `best_chv` initialization only.

We compare MO-CMA-ES-2 with three algorithms: COMO-CMA-ES, HMO-CMA-ES and UP-MO-CMA-ES on the COCO benchmarking platform. It succeeds in providing an anytime extension of COMO-CMA-ES, that performs well for any budget without needing to adapt any algorithm parameter. It achieves similar performance as the Hybrid Multi-Objective CMA-ES (HMO-CMA-ES), an hybrid of algorithms (most based on CMA-ES) that achieved the best performance on the COCO platform so far. It also performs better than the Unbounded Population Multi-Objective CMA-ES (UP-MO-CMA-ES) combined with a simultaneous run of multiple Multi-Objective CMA-ES (MO-CMA-ES) used to explore the search space at the beginning. The UP-MO-CMA-ES catches up after the end of the exploration phase.

1.4.4 Scope of the results

While we do not focus on the handling of constraints in the search-space, unconstrained problems are within the scope of the theoretical results. Indeed, assumptions are done only on the shape of the Pareto front and on the reference point. The assumptions can be verified by unconstrained problems as well as constrained problems. Coupling the HV-ISOOMOO framework with a constraint-handling single-objective optimizer would lead to a multi-objective algorithm able to handle constraints. However, the algorithm MO-CMA-ES-2 that we introduce relies on the CMA-ES algorithm and does not handle constraints.

The subject of this study is problems where variables are continuous, but the theoretical results of Chapter 3 may also apply to some *mixed-integer* problems, that is problems with both continuous and discrete variables. Problems with discrete variables only are outside the scope of this manuscript.

We focus on the case where the number of objectives m is equal to 2. We discuss in the conclusions of the chapters and of the manuscript the extension of the theoretical results to $m \geq 3$. The HV-ISOOMOO structure can be extended without modification to the $m \geq 3$ case. Extending the MO-CMA-ES-2 algorithm to the case $m \geq 3$ case would necessarily require a modification of the handling of extremes, and likely other modifications for MO-CMA-ES-2 to stay a competitive algorithm.

CHAPTER 2

Introduction (version française)

L'optimisation est le domaine des mathématiques appliquées qui traite du problème de la recherche du minimum (ou du maximum) d'une ou plusieurs fonctions objectif. Nous nous concentrons sur les problèmes qui peuvent être formulés comme suit :

$$\min_{X \in \Omega} F_1(X), \ldots, F_m(X) \ , \tag{2.1}$$

où Ω est typiquement un sous-ensemble d'un espace vectoriel de dimension finie et les fonctions objectif F_i sont des fonctions de Ω dans $\mathbb{R}$. Le nombre d'objectifs est désigné par la lettre m. Un tel problème est appelé un problème d'optimisation *mono-objectif* lorsque $m = 1$ et multi-objectifs pour $m \geq 2$.

Dans cette thèse, l'accent est mis sur l'optimisation bi-objectifs, mais nous mentionnerons aussi des problèmes et des algorithmes d'optimisation mono-objectif. Pour rendre la distinction plus claire, la fonction objectif d'un problème mono-objectif sera désignée par h.

2.1 Optimisation mono-objectif

L'idée de rechercher un minimum ou un maximum est assez ancienne et remonte à l'antiquité. À cette époque, le calcul infinitésimal n'existait pas encore, et ce problème était étudié d'un point de vue géométrique. Un exemple bien connu est le théorème élémentaire de la géométrie euclidienne qui énonce que le plus petit chemin entre deux points est une ligne droite. Un autre exemple classique se déroule quelques siècles plus tard lorsque Héron utilise un argument géométrique pour prouver que le chemin de la lumière lorsqu'elle est réfléchie sur une surface est le plus court chemin : [25, p261-265]. L'approche géométrique est limitée aux problèmes ayant une interprétation géométrique et manque de généralité.

Ce n'est qu'au XVII$^{\text{ème}}$ siècle que l'invention du calcul infinitésimal par Leibniz et Newton a donné naissance à des méthodes systématiques de résolution des problèmes d'optimisation. Le potentiel du calcul pour l'optimisation était évident dès le début. Il est significatif que le titre du premier article sur le calcul infinitésimal soit *Nova Methodus pro Maximis et Minimis*, que l'on peut traduire par "Nouvelle méthode pour les maxima et les minima" [60]. Dans cet article, Leibniz aborde la relation entre les dérivées et les optima. Bien qu'il n'ait pas été formulé ainsi à l'époque, cette œuvre pionnière contenait déjà dans l'esprit le résultat fondamental selon lequel les dérivées d'une fonction sont nulles à ses minima et maxima. Pour les fonctions différentiables simples, cette condition nécessaire et une analyse de base suffisent pour obtenir une expression analytique d'un optimum global.

Plus près de nous, le développement des ordinateurs au XX$^{\text{ème}}$ siècle a ouvert la voie à une nouvelle expansion de l'éventail des problèmes d'optimisation que nous sommes en mesure de résoudre au moins de manière approximative. La solution de nombreux problèmes insolubles sur le plan analytique peut être approchée en exécutant de manière automatisée un algorithme à l'aide d'un ordinateur. L'optimisation a bénéficié à la fois de la croissance exponentielle de la puissance de calcul et du développement d'une grande variété d'algorithmes. Les fonctions objectif sans dérivées ou sans formule analytique connue ne sont plus des obstacles insurmontables. Le traitement de tels problèmes est au cœur de l'optimisation sans dérivée et de l'optimisation blackbox : [4].

Les applications de l'optimisation ne se limitent pas aux quelques exemples du domaine de l'optique mentionnés plus haut. Elles sont nombreuses en physique. Les deux exemples discutés ci-dessus sont des cas particuliers du principe plus général de moindre action, qui a été introduit par Maupertuis au 18ème siècle. Ce principe permet de formuler de nombreuses lois de la physique, de la mécanique à la théorie quantique, comme la minimisation d'une quantité appelée *action* [69]. En outre, un certain nombre de phénomènes naturels concrets sont le produit de la sélection naturelle des habitudes ou de la génétique. De tels phénomènes peuvent être considérés comme des solutions approximatives de problèmes d'optimisation. C'est le cas, par exemple, de la forme des nids d'abeilles : [44]. Enfin, l'optimisation est largement utilisée dans l'industrie, notamment dans le domaine du design industriel. Au XVIII$^{\text{ème}}$ siècle, MacLaurin, le scientifique qui a élaboré la première théorie complète des minima et maxima en fonction du signe des dérivées, a utilisé l'optimisation pour concevoir des bâteaux et des moulins (à eau et à vent) [44]. De nos jours, les algorithmes d'optimisation sont utilisés dans un large éventail d'applications en ingénierie, de la conception de voitures [35] à la programmation des arrêts de réacteurs nucléaires [56]. Ce rapide aperçu montre que le problème de la recherche des minima ou des maxima d'une certaine quantité apparaît dans de nombreux domaines. L'optimisation est un outil important tant en science, pour comprendre et décrire les phénomènes naturels, que dans l'industrie.

2.2 Optimisation multi-objectifs

On considère généralement que l'optimisation multi-objectifs est née à la fin du XIX$^{\text{ème}}$ siècle avec les travaux pionniers de Edgeworth et Pareto : *Psychologie mathématique* (publié en 1881) [37] et *Manuel d'économie politique* (publié en 1906 et traduit en anglais en 1971) [66]. Edgeworth et Pareto sont deux économistes. Le domaine de l'économie a bénéficié des développements théoriques et algorithmiques antérieurs réalisés dans le contexte de l'optimisation mono-objectif.

Comme son homologue mono-objectif, l'optimisation multi-objectifs a de nombreuses applications dans des domaines variés : de l'économie, son domaine d'origine, [23] à l'écologie [55] en passant par les neurosciences [34]. L'optimisation multi-objectifs est également particulièrement pertinente en ingénierie. En effet, les problèmes d'ingénierie impliquent souvent plus d'une quantité intéressante dont la minimisation ou la maximisation est souhaitable. Lors de la conception d'un nouveau produit, un faible coût de production, une bonne qualité, une grande résistance aux chocs et une masse raisonnablement faible sont des objectifs fréquemment en conflit les uns avec les autres. Par exemple, [35] présente un problème multi-objectifs de conception de voiture impliquant la minimisation à la fois de la masse totale de la voiture et de l'intrusion dans l'habitacle en cas de collision. L'optimisation multi-objectifs est également en jeu dans l'apprentissage automatique (machine learning en anglais). En effet, l'apprentissage automatique cherche à fournir le modèle le plus précis possible tout en évitant le surajustement, un problème intrinsèquement multi-objectifs : [52].

La résolution de problèmes et la prise de décision sont plus faciles lorsqu'un seul objectif est en jeu. Il est donc tentant de transformer un problème multi-objectifs en un problème à objectif unique, en se concentrant soit sur l'objectif le plus important, soit sur une combinaison d'objectifs. Toutefois, cela se ferait au prix de l'effacement de la complexité intrinsèque de la décision. Une approche multi-objectifs, bien que plus complexe, permet un choix éclairé qui prend en compte simultanément plusieurs aspects importants de la décision.

Le concept central de l'optimisation multi-objectifs est la *domination*. On dit qu'une solution domine une autre solution si elle au moins aussi bonne pour tous les objectifs et meilleure pour au moins un objectif. Si, parmi deux solutions, aucune ne domine l'autre, on dit que ces solutions sont *incomparables*. La domination peut également s'appliquer à des ensembles de solutions. On dit qu'un ensemble de solutions domine un autre ensemble lorsque chaque solution du second est dominée par au moins une solution du premier. L'optimisation multi-objectifs s'intéresse à *l'ensemble de Pareto*, l'ensemble de toutes les solutions réalisables non dominées, et au *front de Pareto*, son image dans l'espace objectif. L'ensemble et le front de Pareto rassemblent respectivement les solutions et les compromis qui pourraient être choisis par une personne rationnelle qui ne s'intéresse qu'aux objectifs pris en considération dans la formulation du problème.

Aux débuts du domaine, la finalité de l'optimisation multi-objectifs était de trouver un seul vecteur appartenant à l'ensemble de Pareto, une *solution efficace*. Aujourd'hui, la finalité s'est déplacée vers la recherche d'une bonne approximation de l'ensemble de

Pareto ou du front de Pareto. Rechercher une approximation du front de Pareto revient à se contenter d'une solution efficace par compromis raisonnable. Au contraire, recherche une approximation de l'ensemble de Pareto revient à vouloir connaître toutes les solutions efficaces. Selon l'application, la connaissance de solutions efficaces redondantes présente un intérêt variable.

Le problème fondamental qui se pose lorsque le nombre d'objectifs est plus grand que 1 est la disparition d'un ordre naturel sur les solutions ou les ensembles de solutions. De nombreuses solutions et ensembles de solutions sont incomparables du point de vue de la domination. Cela soulève les questions pratiques de comment guider les algorithmes et mesurer la qualité des solutions ou ensembles de solutions qu'ils génèrent.

Une option consiste à s'appuyer sur un *indicateur de qualité* qui attribue un nombre réel à tout ensemble de solutions. Des dizaines d'indicateurs de qualité existent déjà [80, 2]. À l'exception de ceux basés sur la cardinalité, qui ne sont pas très informatifs, ils sont généralement construits pour favoriser les ensembles de solutions dont l'image est proche du front de Pareto, ou qui offrent une meilleure couverture du front de Pareto [2]. Si les indicateurs attribuent souvent la même valeur à plusieurs ensembles différents, il est rare qu'un indicateur non discret attribue la même valeur à deux ensembles distincts de solutions générés par un ou plusieurs algorithmes. La distance générationnelle inversée (IGD), la distance de Hausdorff moyennée et la famille des indicateurs ϵ, qui mesurent tous la proximité entre l'ensemble évalué et le front de Pareto [2] font partie des indicateurs les plus populaires. Dans ce manuscrit, nous utilisons l'hypervolume, un indicateur de qualité d'ensemble populaire qui favorise les ensembles de solutions dont l'image est proche du front de Pareto et bien distribuée sur le front de Pareto. Contrairement aux indicateurs mentionnés ci-dessus, l'hypervolume ne repose pas sur la connaissance du front de Pareto. Nous aborderons cet indicateur plus en détail dans la suite de l'introduction.

Nous avons développé un algorithme qui s'inscrit dans la tradition des *Algorithmes évolutionnaires multi-objectifs* (MOEAs). *Agorithmes évolutionnaires* (EAs) forment à l'origine une classe d'algorithmes mono-objectif qui constitue une bonne base pour le développement d'algorithmes d'optimisation multi-objectifs. Ce sont des algorithmes stochastiques inspirés du processus biologique de sélection naturelle qui font évoluer une population de solutions. L'évolution de la population est pilotée par une *fonction d'aptitude*, qui attribue une valeur à toute solution échantillonnée. Les MOEAs ont été développés dans les années 90. Leur structure est la même que celle des EAs mais la fonction qui mesure la valeur sélective des solutions et guide l'évolution de la population est conçue pour traiter des problèmes multi-objectifs. Avec un guidage approprié, la population peut évoluer jusqu'à former une bonne approximation de l'ensemble de Pareto. L'algorithme que nous développons n'est pas un EA multi-objectifs en soi. Cependant, il s'agit d'un algorithme d'optimisation multi-objectifs qui s'appuie sur un EA, CMA-ES, pour résoudre des sous-problèmes mono-objectif.

Il existe une grande variété de MOEAs. Emmerich et al. les séparent en trois catégories selon qu'il sont basé sur la domination de Pareto, une décomposition ou sur des indicateurs de qualité [39]. Ces catégories diffèrent par la manière dont la mesure

la valeur sélective des solutions est construite. Les algorithmes basés sur la domination de Pareto classent d'abord itérativement les solutions en utilisant la domination. Les solutions qui ont le même rang de domination sont ensuite classées selon un critère qui capture leur contribution à la diversité. C'est le principe de l'un des MOEAs les plus connus, NSGA-II [31]. Les algorithmes basés sur la décomposition décomposent le problème multi-objectifs en sous-problèmes mono-objectif à l'aide d'une *scalarisation*. La scalarisation transforme un problème multi-objectifs en une classe paramétrée de problèmes mono-objectif. Avec une scalarisation bien choisie, les solutions optimales des problèmes scalaires sont efficaces et chaque vecteur de l'ensemble de Pareto est une solution optimale d'un problème scalaire. Les solutions sont attribuées dynamiquement à ces sous-problèmes. Les algorithmes NSGA-III [33] et MOEA/D [68] sont deux algorithmes célèbres basés sur la décomposition. Enfin, la valeur sélective utilisée dans les algorithmes basés sur des indicateurs est construite à partir d'indicateurs de qualité d'ensemble. L'algorithme que nous développons est proche de cette dernière catégorie.

2.3 Vitesse de convergence

L'objectif de ce manuscrit est d'étudier la vitesse de convergence des algorithmes d'optimisation multi-objectifs. La vitesse de convergence d'un algorithme détermine les ressources de calcul qui doivent être dépensées afin d'obtenir un ensemble de solutions d'une qualité donnée. La connaissance de la vitesse de convergence guide les choix pratiques de l'algorithme utilisé et le budget alloué. Dans cette section, nous définissons le vocabulaire et les notations nécessaires pour décrire la vitesse de convergence et détaillons le type de vitesse de convergence qui nous intéresse dans ce manuscrit.

2.3.1 Notations et vocabulaire

Nous présentons les notations et le vocabulaire standard que nous allons utiliser pour décrire la vitesse de convergence d'algorithmes. Nous les définissons formellement dans le contexte de la convergence d'une suite de nombres réels vers une limite.

Les notations de Bachmann-Landau o, O, Θ et Ω permettent de comparer la vitesse de convergence de séquences convergeant soit vers $+\infty$ soit vers 0. Nous nous intéressons ici uniquement à ce dernier cas. Nous définissons rigoureusement ces notations ci-dessous.

Définition 1 (Notations de Bachmann-Landau) *Soient f et g deux fonctions de $\mathbb{N}^*$ dans $\mathbb{R}$.*

On dit que:

- $f(n) = o(g(n))$ *si pour tout $\epsilon > 0$, il existe $N \in \mathbb{N}^*$ tel que pour tout $n \geq N$, $f(n) \leq \epsilon \times g(n)$,*
- $f(n) = O(g(n))$ *s'il existe $N \in \mathbb{N}^*$ et $C > 0$ tels que pour tout $n \geq N$, $f(n) \leq C \times g(n)$,*

- $f(n) = \Omega(g(n))$ *s'il existe* $N \in \mathbb{N}^*$ *et* $C > 0$ *tels que pour tout* $n \geq N$, $f(n) \geq C \times g(n)$ *et*

- $f(n) = \Theta(g(n))$ *si* $f(n) = O(g(n))$ *et* $f(n) = \Omega(g(n))$.

Dans cette thèse, nous nous référons régulièrement aux vitesses de convergence $\Omega(1/n)$, $O(1/n)$ et $\Theta(1/n)$. Une séquence typique avec une vitesse de convergence de $\Theta(1/n)$ est simplement $(1/n)_{n\in\mathbb{N}^*}$. Pour cette séquence particulière, améliorer la précision d'un facteur dix nécessite de multiplier l'indice d'itération par dix. Cela devient de plus en plus coûteux à mesure que l'indice d'itération n augmente.

La convergence en $\Theta(1/n^\alpha)$ avec $0 < \alpha$ est une sous-catégorie de *convergence sous-linéaire*. On dit d'une suite (u_n) de nombres réels qu'elle *converge linéairement* vers une limite L lorsque le rapport entre $|u_{n+1} - L|$ et $|u_n - L|$ est inférieur à une constante $C < 1$ pour n suffisamment grand. Ainsi, au bout d'un certain temps, la précision s'améliore toujours d'un facteur 10 au moins après $1/\log_{10}(1/C)$ itérations. Par exemple, pour $C = 0.99$ et $C = 0.999$, il faut respectivement 229 et 2301 itérations pour garantir une amélioration d'un facteur 10. Plus la constante C est petite, plus la convergence est rapide. Un exemple de séquence qui converge linéairement vers 0 est $(1/10^n)_{n\in\mathbb{N}^*}$. Pour cette séquence particulière, la précision s'améliore d'un facteur dix à chaque itération. La convergence linéaire est atteinte par de nombreux algorithmes mono-objectif sur des fonctions faciles à résoudre. Il a récemment été prouvé que les stratégies évolutionnaires (une sous-catégorie d'algorithmes évolutionnaires) avec un mécanisme de contrôle sur la taille des pas atteignent une vitesse de convergence linéaire sur une large classe de fonctions, y compris les transformations monotones de fonctions fortement convexes avec gradient continu de Lipschitz [1]. Bien que cela n'ait pas encore été prouvé mathématiquement, la stratégie évolutionnaire populaire CMA-ES que nous exploitons est également connue empiriquement pour atteindre une convergence linéaire sur les fonctions convexes-quadratiques. Les algorithmes évolutionnaires, comme tous les algorithmes d'optimisation basés sur la comparaison, ne peuvent converger plus rapidement que linéairement [72]. Cependant, certains algorithmes mono-objectif non basés sur la comparaison sont capables d'atteindre une vitesse de convergence superlinéaire. Par exemple, on sait que la vitesse de convergence locale de la méthode de Newton est quadratique pour des fonctions objectif deux fois différentiables dont la hessienne est Lipschitz [65, Theorem 3.7]. On dit d'une suite (u_n) de nombres réels qu'elle *converge quadratiquement* vers une limite L lorsque le rapport entre $|u_{n+1} - L|$ et $|u_n - L|^2$ est inférieur à une constante C pour n suffisamment grand. Un exemple de séquence à convergence quadratique est $(1/10^{2^n})_{n\in\mathbb{N}^*}$.

Cette catégorisation des vitesses de convergence asymptotique nous informe sur la manière dont un algorithme converge après qu'une certaine quantité non spécifiée de ressources ait été dépensée. En pratique, cette quantité non spécifiée est importante, de même que le délai avant d'atteindre le régime asymptotique et les constantes impliquées dans les définitions. Par exemple, converger linéairement comme $10^{10} \times 1/2^n$ est moins efficace à très court terme (pour $n \leq 55$) que converger linéairement comme $10^{-5} \times 1/n$. A partir de $n = 56$, un petit nombre comparé à l'écart de 10^{15} entre les constantes en

jeu, la première séquence a rattrapé son retard.

2.3.2 Définition de la vitesse de convergence utilisée dans ce manuscrit

Dans cette section, nous détaillons la vitesse de convergence que nous avons choisi d'étudier. Une définition de la vitesse de convergence est composée d'une mesure de progrès et d'une mesure de coût. Il existe de nombreuses façons de mesurer le coût et encore plus de mesurer le progrès.

Mesure du coût : nombre d'évaluations de fonctions, d'itérations ou de méta-itérations

Le coût doit refléter les ressources informatiques dépensées par un algorithme. Ces ressources dépendent non seulement du problème et de l'algorithme, mais elles peuvent également varier considérablement en fonction du matériel, de l'implémentation de l'algorithme et du coût de calcul des fonctions objectif. Elles peuvent être difficiles à analyser théoriquement.

Ici, nous avons choisi de définir le coût comme le nombre d'évaluations de la fonction objective. Le nombre d'évaluations de fonctions ne dépend pas du matériel et se prête bien à une analyse théorique. L'évaluation d'une fonction objectif a toujours un coût, qui peut être plus ou moins important. Il est particulièrement important lorsque le calcul de la fonction objectif nécessite une simulation ou une expérience coûteuse. C'est le cas dans le problème réaliste exposé dans [35], où le calcul de la fonction objectif nécessite une simulation de crash de voiture dont le temps de calcul se situe entre 10 et 20h sur 8 CPUs. Dans le cas plus général, le nombre d'évaluations de fonctions objectif est un coût à considérer parmi d'autres. Par exemple, le gain de temps important qui peut être obtenu par la parallélisation n'impacte pas le nombre d'évaluations de fonctions.

Nous nous intéresserons également à la vitesse de convergence des algorithmes par rapport au nombre d'itérations ou de méta-itérations (dans le cas des algorithmes embarqués).

Mesure de progrès : hypervolume

Nous nous appuyons sur l'hypervolume $\mathrm{HV}_r(.)$, un indicateur de qualité populaire, pour mesurer la qualité des ensembles de solutions fournis par les algorithmes. L'*écart d'optimalité* associé à un ensemble S est $\mathrm{HV}_r(\mathrm{PF}_f) - \mathrm{HV}_r(S)$. Un progrès correspond à un gain de qualité, ou de manière équivalente à une diminution de l'écart d'optimalité.

L'hypervolume mesure la taille de la région de l'espace objectif dominée par un ensemble de solutions (et dominant un point de référence r), voir Figure 1.1. Cela favorise les ensembles proches du front de Pareto et bien distribués. Dans une certaine mesure, les résultats théoriques éclairent les préférences en matière d'hypervolume. Il a été prouvé que la densité asymptotique de l'image des distributions n-optimales, c'est-à-dire des ensembles de n solutions à l'hypervolume le plus élevé, est maximale dans les parties où la pente est -1 [8]. Cela montre que l'hypervolume favorise les ensembles

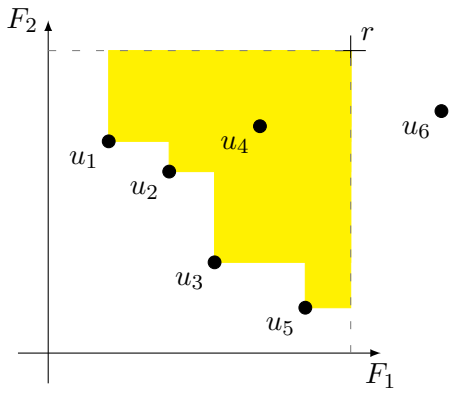


Figure 2.1: Illustration de l'hypervolume. L'hypervolume par rapport au point de référence r de l'ensemble $\{u_1, \ldots, u_6\}$ est l'aire de la région jaune.

dont la distribution est plus clairsemée dans les parties du front de Pareto où un petit sacrifice pour un objectif offre une récompense particulièrement importante pour l'autre. Une telle distribution peut être plus informative qu'une distribution uniforme pour les décideurs. En effet, il est probable qu'ils choisissent de faire ce sacrifice intéressants et de sélectionner un compromis entre les objectifs qui ne se trouve pas dans de telles parties du front de Pareto.

L'hypervolume est un indicateur populaire. Il est utilisé dans de nombreux MOEAs, parmi lesquels SMS-EMOA [15], HypE [11], IBEA [82], POSEA [77] et COMO-CMA-ES [74]. Cela le rend particulièrement adapté à l'étude des performances de ces algorithmes. Il est déjà utilisé pour le benchmark des algorithmes. L'indicateur utilisé dans les suites de tests d'algorithmes bi-objectifs de la plateforme de benchmarking COCO [22] est construit à partir de l'hypervolume. Il est également disponible dans d'autres outils d'évaluation comparative tels que Nevergrad [12].

Forces et faiblesses de l'hypervolume

Comme mentionné précédemment, les préférences de l'hypervolume vis-à-vis d'ensembles de solutions sont intéressantes pour un décideur. De plus, cet indicateur est facile à appréhender et peut être calculé sans connaître front de Pareto. La popularité de l'indicateur d'hypervolume est largement due à sa stricte conformité avec la domination de Pareto, ce qui signifie qu'un ensemble de solutions qui en domine un autre a toujours un hypervolume strictement supérieur. Tous les indicateurs strictement conformes à la domination de Pareto connus sont construits à partir de l'hypervolume. On considère depuis longtemps qu'il s'agit d'une propriété souhaitable, mais sa pertinence pratique est encore à l'étude : [40].

L'hypervolume présente deux inconvénients principaux, qui ne sont de véritables inconvénients que pour plus de deux objectifs. Premièrement, il nécessite de fixer un point de référence. En pratique, trouver un point de référence qui ne retire pas de

solutions intéressantes est relativement facile. Ses coordonnées peuvent être fixées aux valeurs les plus grandes qui intéressent le décideur ou à des bornes supérieures des valeurs des fonctions objectif pour des solutions non dominées. Cependant, le choix du point de référence n'est pas trivial et a un impact sur la préférence sur les ensembles de solutions de l'indicateur. Lorsqu'il y a plus de deux objectifs, un point de référence trop éloigné conduit parfois à une préférence problématique, où les extrêmes - qui sont en nombre infini - sont sur-favorisés [71]. Deuxièmement, le calcul de l'hypervolume est un problème **#P**-hard en m, le nombre d'objectifs [17]. Cela implique que l'hypervolume ne peut pas être calculé en un temps polynomial en m (tant que **P**$\neq$ **NP**). Pour deux et trois objectifs, la complexité temporelle du calcul de l'hypervolume d'un ensemble de n points est de $O(nlog(n))$ [45, 13]. Pour plus de 3 objectifs, le meilleur algorithme connu a une complexité temporelle de $O(n^{m/3}\text{polylog}(n))$ [24, 45]. Heureusement, des approximations de l'hypervolume peuvent être calculées en un temps plus raisonnable. Par exemple, un schéma d'approximation randomisé en temps entièrement polynomial a été présenté par Bringmann et Friedrich en 2010 [17]. Le développement de méthodes efficaces pour calculer de bonnes approximations de l'hypervolume pour plus de trois objectifs est un sujet actif : [71].

2.4 Résultats

Les résultats présentés dans le présent manuscrit apportent des réponses à la question suivante :

> À quelle vitesse un algorithme d'optimisation bi-objectifs peut-il converger vers le front de Pareto tout entier ?

Nous utilisons cette question comme fil conducteur.

2.4.1 Une limite supérieure théorique sur la vitesse de convergence

Tout d'abord, nous abordons cette question d'un point de vue purement théorique dans Chapter 3 :

> Considérons un problème d'optimisation bi-objectifs donné. Comment le plus petit écart d'optimalité associé à un ensemble de n solutions évolue-t-il avec n ?

Les ensembles de n solutions admissibles ayant le plus grand hypervolume sont appelés *distributions n-optimales*. Dans Chapter 3, nous démontrons pour une large classe de fronts de Pareto que la vitesse de convergence de l'écart d'optimalité des distributions n-optimales est $\Omega(1/n)$. Une conséquence directe est que la vitesse de convergence des algorithmes d'optimisation bi-objectifs est également $\Omega(1/l)$ sur les problèmes pour lesquels le front de Pareto satisfait les hypothèses pertinentes, avec l le nombre de solutions évaluées jusqu'à présent par l'algorithme. En particulier, cette affirmation est vraie lorsque les deux fonctions objectives sont convexes quadratiques [73], une classe de problèmes parmi les plus faciles à résoudre. Une vitesse de convergence de $\Omega(1/l)$ est

sous-linéaire et radicalement plus lente que les vitesses de convergence atteignable pour des fonctions objectifs quadratiques convexes par des optimiseurs mono-objectif.

Nous étudions également de manière empirique la vitesse de convergence de l'écart d'optimalité des distributions n-optimales. Pour six fronts de Pareto, nous calculons une distribution n-optimale pour n allant jusqu'à 1000. Pour les valeurs de n supérieures à 10, l'écart d'optimalité de la distribution n-optimale calculée s'avère être proche de $1/n$ multiplié par une constante.

2.4.2 Une structure d'algorithmes au potentiel d'atteindre une vitesse de convergence optimale

Dans la Chapter 4, nous présentons la structure algorithmique HV-ISOOMOO qui s'appuie sur une approche gourmande pour construire une approximation de l'ensemble du front de Pareto. Une *méta-iteration* d'HV-ISOOMOO correspond à une exécution complète d'un solveur d'optimisation mono-objectif, qui peut elle-même être décomposé en itérations. Nous définissons l'*approximation du front de Pareto formée par les solutions finales*(FI-PFA) comme le sous-ensemble non-dominé de l'ensemble des solutions finales fournies par toutes les exécutions du solveur mono-objectif effectuées jusqu'à présent. À chaque méta-itération, l'objectif est de trouver une solution dont l'ajout aura la plus grande contribution à l'hypervolume de cette approximation du front de Pareto.

Le couplage de la structure algorithmique HV-ISOOMOO avec un solveur mono-objectif idél qui renvoie toujours un optimum global exact est appelée *HV-ISOOMOO idéal*. À mi-chemin entre la théorie et la pratique, nous étudions la question suivante :

> Considérons un problème d'optimisation bi-objectifs donné. À quelle vitesse la séquence des approximation du front de Pareto formées par les solutions finales (FI-PFA) associée à l'algorithme ideal HV-ISOOMOO converge-t-elle vers l'ensemble du front de Pareto ?

Nous prouvons trois résultats principaux concernant la vitesse de convergence de l'HV-ISOOMOO idéal. Nous prouvons que lorsque le front de Pareto est convexe, la vitesse de convergence de l'écart d'optimalité de la FI-PFA de l'HV-ISOOMOO idéal est $O(1/n)$, avec n le nombre de méta-itérations. En combinant ce résultat avec la borne supérieure symétrique fournie dans Chapter 3, il s'ensuit que la vitesse de convergence est alors $\Theta(1/n)$. Une limite inférieure plus faible sur la vitesse de convergence de l'HV-ISOOMOO idéall est également trouvée pour les fronts de Pareto bilipschitz. Enfin, nous prouvons pour une certaine classe de fronts de Pareto que l'impact asymptotique du doublement du nombre de méta-itérations est de réduire de moitié l'écart d'optimalité.

En plus de cette analyse théorique, nous examinons empiriquement les écarts d'optimalité des FI-PFA de l'HV-ISOOMOO idéal pour sept fronts de Pareto, pour n allant jusqu'à 1000. L'écart d'optimalité de la FI-PFA après n méta-itérations s'avère être proche de celui des distributions n-optimales pour tous les fronts de Pareto étudiés. Dans nos expériences, il est toujours moins de 24% plus important.

2.4.3 Un algorithme concret

Nous présentons MO-CMA-ES-2, un algorithme qui suit la structure d'HV-ISOOMOO et s'appuie sur l'algorithme CMA-ES pour résoudre les sous-problèmes mono-objectif. Des critères d'arrêt ont été ajoutés pour augmenter plus finement la précision de la solution renvoyée par CMA-ES lors d'une exécution de MO-CMA-ES-2. Les CMA-ES sont initialisés de deux manières distinctes qui partagent équitablement le budget des évaluations de fonctions : de manière aléatoire (initialisation `random`) et à proximité des régions prometteuses (initialisation `best_chv`).

Nous avons examiné empiriquement les performances de MO-CMA-ES-2. Plus précisément, nous étudions les questions suivantes :

> La vitesse de convergence de MO-CMA-ES-2 est-elle optimale ?

et

> Comment MO-CMA-ES-2 se compare-t-il aux autres algorithmes d'optimisation bi-objectifs ?

Nous examinons empiriquement la vitesse de convergence de MO-CMA-ES-2 sur le problème **doublesphere**, un problème quadratique convexe bien conditionné. La vitesse de convergence de la FI-PFA de MO-CMA-ES-2 par rapport au nombre de méta-itérations n est $\Theta(1/n)$. La vitesse exacte est proche de celle de l'HV-ISOOMOO idéal et des distributions n-optimales. La vitesse de convergence par rapport au nombre d'itérations k semble être légèrement plus lente, mais ressemble à $\Theta(1/k^{\alpha})$ avec α qui va jusqu'à 0,99 pour une petite dimension de l'espace de recherche et en utilisant uniquement l'initialisation `best_chv`.

Nous comparons MO-CMA-ES-2 avec trois algorithmes : COMO-CMA-ES, HMO-CMA-ES et UP-MO-CMA-ES sur la plateforme COCO. L'algorithme MO-CMA-ES-2 est une bonne extension de COMO-CMA-ES, performante pour n'importe quel budget même sans adapter de paramètre au budget. Il obtient des performances similaires à HMO-CMA-ES, un hybride d'algorithmes (la plupart basés sur le CMA-ES) qui a obtenu les meilleures performances sur la plateforme COCO jusqu'à présent. Il est également plus performant que UP-MO-CMA-ES combiné à une exécution simultanée de plusieurs MO-CMA-ES utilisés pour explorer l'espace de recherche au début. L'algorithme UP-MO-CMA-ES rattrape le retard à la fin de la phase d'exploration.

2.4.4 Portée des résultats

Bien que nous ne nous intéressons pas au traitement des contraintes dans l'espace de recherche, les problèmes sans contraintes sont dans le champ des résultats théoriques. En effet, des hypothèses sont faites uniquement sur la forme du front de Pareto et sur le point de référence. Les hypothèses peuvent être vérifiées aussi bien par des problèmes non contraints que par des problèmes contraints. Le couplage de la structutre HV-ISOOMOO avec un optimiseur mono-objectif gérant les contraintes conduirait à un algorithme multi-objectifs capable de gérer les contraintes. Cependant, l'algorithme

MO-CMA-ES-2 que nous présentons repose sur l'algorithme CMA-ES et ne gère pas les contraintes.

L'objet de cette étude est les problèmes à variables continues, mais les résultats théoriques de Chapter 3 peuvent également s'appliquer à certains problèmes en nombres entiers mixtes, c'est-à-dire des problèmes avec à la fois des variables continues et des variables discrètes. Les problèmes qui ne comportent que des variables discrètes sortent du cadre de ce manuscrit.

Nous nous concentrons sur le cas où le nombre d'objectifs est égal à 2. Nous discutons dans les conclusions des chapitres et du manuscrit de l'extension des résultats théoriques à $m \geq 3$. La structure HV-ISOOMOO peut être étendue sans modification au cas $m \geq 3$. L'extension de l'algorithme MO-CMA-ES-2 au cas $m \geq 3$ demandera nécessairement une modification de la gestion des extrêmes, et vraisemblablement d'autres modifications pour qu'il demeure un algorithme compétitif.

CHAPTER 3

Hypervolume in Bi-objective Optimization Cannot Converge Faster Than $\Omega(1/n)$

Adapted with minor modifications from a paper presented at the GECCO conference in July 2021 [63].

Abstract

The hypervolume indicator is widely used by multi-objective optimization algorithms and for assessing their performance. We investigate a set of n vectors in the bi-objective space that maximizes the hypervolume indicator with respect to some reference point, referred to as *n-optimal distribution*. We prove explicit lower and upper bounds on the gap between the hypervolumes of the n-optimal distribution and the ∞-optimal distribution (the Pareto front) as a function of n, of the reference point, and of some Lipschitz constants. On a wide class of functions, this optimality gap can not be smaller than $\Omega(1/n)$, thereby establishing a bound on the optimal convergence speed of any algorithm. For functions with either bilipschitz or convex Pareto fronts, we also establish an upper bound and the gap is hence $\Theta(1/n)$. The presented bounds are not only asymptotic. In particular, functions with a linear Pareto front have the normalized exact gap of $1/(n+1)$ for any reference point dominating the nadir point.

We empirically investigate on a small set of Pareto fronts the exact optimality gap for values of n up to 1000 and find in all cases a dependency resembling $1/(n+\mathrm{CONST})$.

3.1 Introduction

Multi-objective optimization aims at minimizing a vector-valued function F. Nowadays, a main goal of multi-objective optimization is to find a good approximation of the Pareto set, the set of all non-dominated feasible vectors of the search space. When measuring the performance in the objective space, there are at least three different ways to define the convergence of a multi-objective optimization algorithm towards the Pareto front. First, the F-values of a subsequence of vectors explored by the algorithm may converge to a vector of the Pareto front. Proof of such convergence already exists for many multi-objective algorithms, such as MultiGLODS [26], Newton's method [42], the projected gradient method [43] and (1+1) evolutionary multi-objective algorithms [14], often with a guarantee on the convergence rate. Second, for evolutionary algorithms, the set of F-values of the population can converge to a good approximation of the Pareto front. In the case of evolutionary algorithms with a hypervolume-based selection, ideally, the image of the population would converge to a set of n vectors of the objective space maximizing the hypervolume, with n being the population size. We call such a set *a n-optimal distribution*. Finally, a dynamic subset of the archive can converge to the entire Pareto front, in some sense which is to be defined. For example, it has been known for a long time that under the hypothesis of an uncountably infinite population, the set of all non-dominated vectors explored by an evolutionary algorithm converges almost surely to the Pareto front [75].

Set-quality indicators are widely used for assessing the performance of multi-objective optimization algorithms. They create a total order where there only was a partial one. Many indicators have been invented and are thoroughly used, like the multiplicative and additive epsilon indicators [78] or the hypervolume indicator. The hypervolume and its variants such as the weighted hypervolume [79] are the only known strictly Pareto compliant indicators [81]. The multiplicative epsilon indicator is also called multiplicative approximation ratio when the Pareto front is used as reference set. In the bi-objective case, it has been proven that the multiplicative approximation ratio of the set of all vectors explored by the algorithm cannot converge to 1 more rapidly than $\Omega(1/l)$, with l being the number of function evaluations [18]. It is a direct consequence of Corollary 3.2 in [18], which gives a lower bound of the form $1 + \Theta(1/l)$ of the minimum multiplicative approximation ratio of a set of n vectors.

In this chapter, we derive lower and upper bounds of the form $\Theta(1/n)$ of the difference in hypervolume between a n-optimal distribution and the Pareto front. We call this difference *optimality gap*. For bilipschitz Pareto fronts, we have a tight lower bound on the optimality gap of the form $\mathrm{CONST}/(n+1)$. The constant depends on the bilipschitz constants and on the position of the reference point of the hypervolume indicator with respect to the nadir point. The lower bound we found is exact in the case of a linear Pareto front where the reference point dominates the nadir point. In this case, the constant is simply the hypervolume of the Pareto front. We generalize this result to Pareto fronts with a bilipschitz subsection which dominates the reference point. For this wide class of Pareto fronts (see Figure 3.6), the rate of convergence of multi-objective

algorithms in terms of hypervolume cannot be better than $\Omega(1/l)$, with l being the number of solutions evaluated so far. For bilipschitz or convex Pareto fronts, we prove an upper bound on the optimality gap of n-optimal distributions of the form CONST $\times$ $(n+1)/n^2$. The constant depends on the extreme values of the Pareto front, on the reference point, and additionally on the bilipschitz constants in the bilipschitz case. Since any convex Pareto front has a bilipschitz subsection, both convex and bilipschitz Pareto fronts abide by the above lower bound. As a consequence, for either bilipschitz or convex Pareto fronts, the optimality gap of n-optimal distributions evolves as $\Theta(1/n)$.

We empirically evaluated the optimality gap for n up to 1000 on six different Pareto fronts, among which three are convex and three are concave. We observe convergence of n times the optimality gap to a constant, even for non-bilipschitz and non-convex Pareto fronts.

The rest of the chapter is organized as follows. In Section 2, we define formally the Pareto front, the hypervolume and the n-optimal distribution. Additionally, we introduce the concept of gap region. In Section 3, we derive first lower and upper bounds on the optimality gap, respectively for bilipschitz and for bilipschitz or convex Pareto fronts. In Section 4, we derive sufficient conditions on the objective functions and the search space for the Pareto front to be convex and bilipschitz. Additionally, we generalize the lower bound for Pareto fronts with only a bilipschitz subsection. In Section 5, we examine the empirical optimality gap on six different Pareto fronts.

Notations We denote the search space by Ω. We denote elements of Ω by X, which should not be confused with x denoting the first coordinate of a vector of the Pareto front. A vector of the objective space, v, is called feasible if $v \in F(\Omega)$. In order to avoid confusion, we always use the term *area* to refer to the Lebesgue-measure in dimension 2 and never to refer to a part of the objective space. Further notations are defined in the next section.

3.2 Preliminaries

We focus on bi-objective optimization, which aims at minimizing two objective functions, F_1 and F_2 over the search space $\Omega \subset \mathbb{R}^d$. We denote $F : X \in \Omega \mapsto (F_1(X), F_2(X))$.

3.2.1 Domination and Pareto front

A vector $u \in \mathbb{R}^2$ of the objective space is said to weakly dominate a vector $v \in \mathbb{R}^2$ when $u_1 \leq v_1$ and $u_2 \leq v_2$. We denote it $u \preceq v$. A vector u is said to dominate a vector v when $u \preceq v$ and $u \neq v$. We denote it $u \prec v$. If a vector u does not dominate a vector v, we denote it $u \not\prec v$. The *Pareto front* is the set of all non-dominated feasible vectors: $\{u \in F(\Omega) : \forall v \in F(\Omega), v \not\prec u\}$. We will assume here that the Pareto front has an explicit representation via f, namely that it can be written as $\{(x, f(x)) : x \in [x_{\min}; x_{\max}]\}$ with $x_{\min} \neq x_{\max}$. By definition of the Pareto front, f must be strictly decreasing. We say that f is ($L_{\min}$,$L_{\max}$)-bilipschitz when $|f(x) - f(y)|$ is between $L_{\min} \times |x - y|$ and

$L_{\max} \times |x - y|$ for all $x, y \in \mathbb{R}$. This is one of the main assumptions of interest here. We will also consider Pareto fronts for which f is convex and the (much) wider class of Pareto fronts with a bilipschitz subsection.

Given a reference point r, we denote the extremes of the part of the Pareto front dominating r with $\tilde{u}_{\min,r} := (\tilde{x}_{\min,r}, f(\tilde{x}_{\min,r}))$ and $\tilde{u}_{\max,r} := (\tilde{x}_{\max,r}, f(\tilde{x}_{\max,r}))$ where $\tilde{x}_{\min,r} := x_{\min}$ when $r_2 \geq f(x_{\min})$ and $f^{-1}(r_2)$ otherwise, and $\tilde{x}_{\max,r} := \min(x_{\max}, r_1)$, see Figure 3.1. In this chapter, we will also assume that the reference point r is valid, in the sense that there exists a feasible vector of the objective space dominating r. The

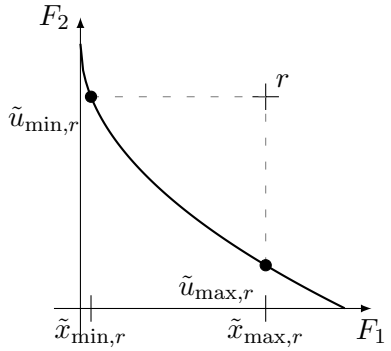


Figure 3.1: Depiction of $\tilde{u}_{\min,r}$ and $\tilde{u}_{\max,r}$ with $r = (0.7, 0.8)$ for the Pareto front associated with the function $f : x \mapsto 1 - \sqrt{x}$ for $x \in [0, 1]$.

nadir point is the vector which coordinate in each objective is the worst value achieved on the Pareto front for this objective. It equals $(x_{\max}, f(x_{\min}))$.

3.2.2 Hypervolume and n-optimal distribution

The hypervolume is a set-quality indicator which depends on a reference point r. The hypervolume of a set S is the Lebesgue measure of the region weakly dominated by S and dominating r : $\lambda(\{u \in F(\Omega) : \exists s \in S, s \preceq u \prec r\})$. We denote it $\mathrm{HV}_r(S)$. The hypervolume improvement and hypervolume contribution of a vector u to a set S quantifies how much adding u to or removing u from the set S, respectively, affects its hypervolume: $\mathrm{HVI}_r(u, S) := \mathrm{HV}_r(S \cup \{u\}) - \mathrm{HV}_r(S)$ and $\mathrm{HVC}_r(u, S) := \mathrm{HVI}_r(u, S \backslash \{u\})$.

Here, we will study sets of n feasible vectors of the objective space maximizing the hypervolume. We call them *n-optimal distributions* and denote them S_r^n. We denote $v_{n,1}, \ldots, v_{n,n}$ the vectors of S_r^n ordered by increasing F_1 values and $x_{n,1}, \ldots, x_{n,n}$ their first coordinates. We also denote $x_{n,0} := \tilde{x}_{\min,r}$ and $x_{n,n+1} := \tilde{x}_{\max,r}$.

3.2.3 Gap regions

In this chapter, we will examine the dependency of the *optimality gap* $\mathrm{HV}_r(\mathrm{PF}) - \mathrm{HV}_r(S_r^n)$ in n. The optimality gap is the area of the region of the objective space dominated by the Pareto front PF but not by the n-optimal distribution S_r^n. We call this region of the objective space *total gap region* and denote it $\mathcal{G}_r^n$. We can now write

the optimality gap as $\lambda(\mathcal{G}_r^n)$. Since the n-optimal distribution S_r^n is a subset of the Pareto front, the total gap region can be decomposed into $n+1$ disjoints regions, that we call *gap regions*, see Figure 3.2. The i-th gap region of S_r^n is the region of the objective space dominated by the Pareto front and dominating the reference point r_i^n with $r_1^n := (x_{n,1}, r_2)$, $r_{n+1}^n := (r_1, f(x_{n,n}))$ and $r_i^n := (x_{n,i}, f(x_{n,i-1}))$ for all $i \in [\![2, n]\!]$. We denote it $\mathcal{G}_{r,i}^n$.

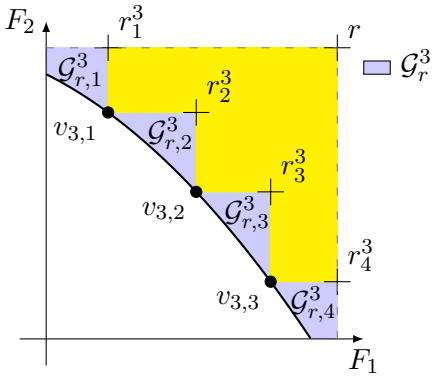


Figure 3.2: The four gap regions of the 3-optimal distribution S_r^3 for the concave and bilipschitz function $f : x \mapsto 1 - 0.5x - 0.5x^2$ for $x \in [0, 1]$.

3.3 Theoretical bounds on the optimality gap $\mathbf{HV}_r(\mathbf{PF}) - \mathbf{HV}_r(S_r^n)$

In this section, we will prove both upper and lower theoretical bounds on the optimality gap $\mathrm{HV}_r(\mathrm{PF}) - \mathrm{HV}_r(S_r^n)$ when f is bilipschitz, and a theoretical upper bound when f is convex. The lower bound is generalized to a wider class of functions in Section 3.4.2. All bounds are equivalent to a constant times $1/n$.

3.3.1 Lower bound on the optimality gap

We will prove that if f is $(L_{\min}, L_{\max})$-bilipschitz, then the optimality gap $\lambda(\mathcal{G}_r^n) = \mathrm{HV}_r(\mathrm{PF}) - \mathrm{HV}_r(S_r^n)$ is greater than $\frac{1}{n+1}$ times a constant depending only on the hypervolume of the Pareto front, on the reference point and on the bilipschitz constants $L_{\min}$ and $L_{\max}$.

Theorem 1 *If f restricted to $[\tilde{x}_{\min,r}, \tilde{x}_{\max,r}]$ is $(L_{\min}, L_{\max})$-bilipschitz, then the normalized optimality gap is bounded from below as*

$$\frac{\lambda(\mathcal{G}_r^n)}{HV_r(PF)} \geq \frac{1}{n+1} \times \frac{L_{\min}}{L_{\max}} \times \frac{1}{1 + 2 \times (q_1 + q_2 + q_1 \times q_2)} \tag{3.1}$$

with $q_1 := \frac{r_1 - \tilde{x}_{\max,r}}{\tilde{x}_{\max,r} - \tilde{x}_{\min,r}}$ *and* $q_2 := \frac{r_2 - f(\tilde{x}_{\min,r})}{f(\tilde{x}_{\min,r}) - f(\tilde{x}_{\max,r})}$.

In particular, when r dominates the nadir point $(x_{\max}, f(x_{\min}))$, both q_1 and q_2 equal 0, and thus the lower bound is simply $\frac{1}{n+1} \times \frac{L_{\min}}{L_{\max}}$.

Proof *We note $c_{n,i} := x_{n,i} - x_{n,i-1}$ for $i \in [\![1, n+1]\!]$, $\Delta_1 := r_1 - \tilde{x}_{\max,r}$ and $\Delta_2 := r_2 - f(\tilde{x}_{\min,r})$, see Figure 3.3.*

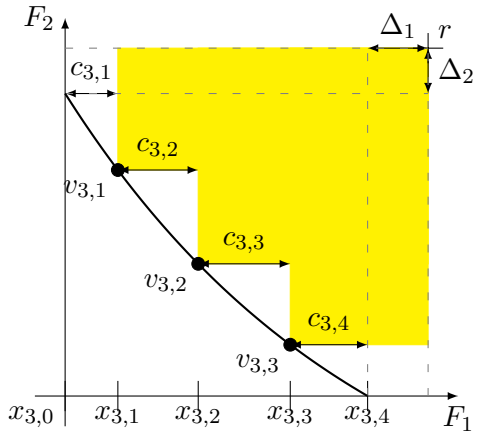


Figure 3.3: Illustration of the notations of the proof of Theorem 1 for the convex and bilipschitz function $f : x \mapsto \frac{e}{e-1} \times e^{-x} + 1 - \frac{e}{e-1}$ for $x \in [0, 1]$ and $r = (1.2, 1.15)$.

The optimality gap $HV_r(PF) - HV_r(S_r^n)$ is greater than $\sum_{i=1}^{n+1} \int_{x=x_{n,i-1}}^{x_{n,i}} (f(x_{n,i-1}) - f(x))dx$. Since the function f is decreasing and its restriction to $[\tilde{x}_{\min,r}, \tilde{x}_{\max,r}]$ is ($L_{\min}$,$L_{\max}$)-bilipschitz, $f(x_{n,i-1}) - f(x)$ is greater than $L_{\min}$ times $x - x_{n,i-1}$. Therefore, the optimality gap $\lambda(\mathcal{G}_r^n)$ is greater than $\frac{1}{2} \times L_{\min} \times \sum_{i=1}^{n+1} c_{n,i}^2$.

The hypervolume of the Pareto front is equal to

$$\Delta_1 \times (f(\tilde{x}_{\min,r}) - f(\tilde{x}_{\max,r})) + \Delta_2 \times (\tilde{x}_{\max,r} - \tilde{x}_{\min,r}) + \Delta_1 \times \Delta_2 + \int_{\tilde{x}_{\min,r}}^{\tilde{x}_{\max,r}} (f(\tilde{x}_{\min,r}) - f(x))dx.$$

The integral and $\Delta_1 \times (f(\tilde{x}_{\min,r}) - f(\tilde{x}_{\max,r}))$ are respectively smaller than $\frac{1}{2} \times L_{\max} \times (\tilde{x}_{\max,r} - \tilde{x}_{\min,r})^2$ and $\Delta_1 \times L_{\max} \times (\tilde{x}_{\max,r} - \tilde{x}_{\min,r})$. By rewriting $\tilde{x}_{\max,r} - \tilde{x}_{\min,r}$ as the sum over $i \in [\![1, n+1]\!]$ of $c_{n,i}$, we obtain that the optimality gap divided by the hypervolume of the Pareto front is greater than

$$\frac{\frac{1}{2} L_{\min} \times \sum_{i=1}^{n+1} c_{n,i}^2}{\frac{1}{2} L_{\max} \times (\sum_{i=1}^{n+1} c_{n,i})^2 + (\Delta_1 L_{\max} + \Delta_2 \times (\sum_{i=1}^{n+1} c_{n,i})) + \Delta_1 \Delta_2}.$$

The terms $\sum_{i=1}^{n+1} c_{n,i}^2$ and $\left(\sum_{i=0}^{n} c_{n,i}\right)^2$ being respectively $\|c_n\|_2^2$ and $\|c_n\|_1^2$ with $c_n := (c_{n,i})_{i \in [\![1,n+1]\!]} \in \mathbb{R}^{n+1}$, it is well-known that their ratio is superior to $\frac{1}{n+1}$. Therefore, the normalized optimality gap is greater than

$$\frac{1}{n+1} \times \frac{L_{\min}}{L_{\max} + 2 \times \frac{\Delta_1 \times L_{\max}}{\sum_{i=1}^{p+1} c_{n,i}} + 2 \times \frac{\Delta_2}{\sum_{i=1}^{n+1} c_{n,i}} + 2 \times \frac{\Delta_1 \times \Delta_2}{(\sum_{i=1}^{n+1} c_{n,i})^2}}.$$

We can rewrite back $\sum_{i=1}^{n+1} c_{n,i}$ *as* $\tilde{x}_{\max,r} - \tilde{x}_{\min,r}$*. Additionally, since* f *is* ($L_{\min}$,$L_{\max}$)*-bilipschitz,* $\frac{\Delta_2 \times L_{\max}}{f(\tilde{x}_{\max,r}) - f(\tilde{x}_{\min,r})}$ *is greater than* $\frac{\Delta_2}{\tilde{x}_{\max,r} - \tilde{x}_{\min,r}}$*.*

The larger $L_{\max}/L_{\min}$ is, the less information we have on the shape of ($L_{\min}$,$L_{\max}$)-bilipschitz Pareto fronts, and the looser is the bound. The quantities q_1 and q_2 reflect the distance between r and the nadir point normalized by the scale of the Pareto front, respectively in the first and in the second objective. When the nadir point dominates the reference point r, the bound gets looser as r moves away from the nadir point. In case of a linear front, we get the following tight result.

Corollary 1 *When* r *dominates the nadir point and* $L_{\min} = L_{\max}$*, the normalized optimality gap* $\frac{\lambda(\mathcal{G}_r^n)}{HV_r(PF)}$ *equals* $\frac{1}{n+1}$*, which corresponds to the lower bound* (3.1) *in Theorem 1.*

Proof *- In this case, the Pareto front is linear, and thus* S_r^n *is evenly distributed [8, Theorem 6]. As a consequence, the normalized optimality gap of the* n*-optimal distribution is exactly* $1/(n+1)$*, which is the lower bound given by Theorem 1.*

3.3.2 Upper bound on the optimality gap

We will prove that if f is either ($L_{\min}$,$L_{\max}$)-bilipschitz or convex, then the optimality gap is smaller than $\frac{n+1}{n^2}$ times a constant. The proof idea is largely inspired by the proof of Theorem 4.4 in [18]. This theorem gives an upper bound on the multiplicative approximation ratio of the n-optimal distribution of the form $1 + \frac{C}{n-4}$ with C being a constant depending on the Pareto front.

The following lemma is a natural extension of [18, Lemma 4.3], where we consider not only the $n-2$ inner vectors of the n-optimal distribution S_r^n but also its two extreme vectors. It states that the smallest hypervolume contribution of any element in S_r^n is below $1/n^2$ times a constant depending only on the extreme values of the Pareto front and on the reference point r.

Lemma 1 *For any* f *such that there exists a* n*-optimal distribution* S_r^n*, we have*

$$\min_{v \in S_r^n} HVC_r(v, S_r^n) \leq \frac{(r_1 - \tilde{x}_{\min,r})(r_2 - f(\tilde{x}_{\max,r}))}{n^2}. \tag{3.2}$$

Proof *We denote* $a_{n,i} := x_{n,i} - x_{n,i-1}$ *and* $b_{n,i} := f(x_{n,i}) - f(x_{n,i-1})$ *for* $i \in [\![1, n+1]\!]$ *except for* $a_{n,n+1}$ *and* $b_{n,0}$ *which are respectively* $r_1 - x_{n,n}$ *and* $r_2 - f(x_{n,1})$*, see Figure 3.4. For all* $i \in [\![1, n]\!]$*, the hypervolume contribution of* $v_{n,i}$ *to* S_r^n *is* $a_{n,i+1} \times b_{n,i}$*. In particular, this implies that for any* i *the quantity* $a_{n,i+1}$ *is greater than* $\frac{\min_{v \in S_r^n} HVC_r(v, S_r^n)}{b_{n,i}}$*. Summing over* i*, we obtain that* $(\sum_{i=1}^n \frac{1}{b_{n,i}}) \times \min_{v \in S_r^n} HVC_r(v, S_r^n)$ *is no greater than* $\sum_{i=1}^n a_{n,i+1}$*, and thus no greater than* $r_1 - \tilde{x}_{\min,r}$*. Additionally, the harmonic mean of the vector* $(b_{n,i})_{i \in [\![1,n]\!]}$ *is no greater than its arithmetic mean:*

$$\frac{n}{\sum_{i=1}^n \frac{1}{b_{n,i}}} \leq \frac{\sum_{i=1}^n b_{n,i}}{n}.$$

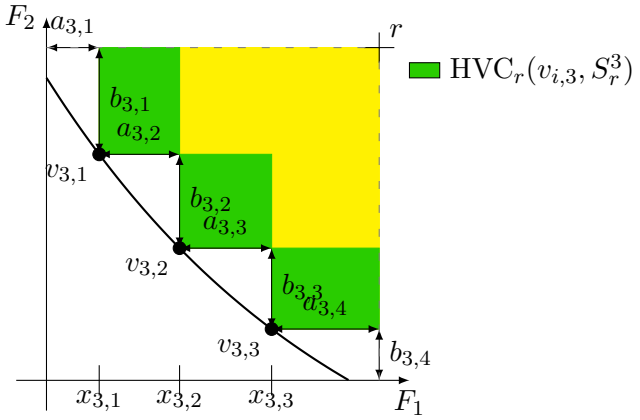


Figure 3.4: Illustration of the notations of the proof of Lemma 1 for $f : x \mapsto \frac{e}{e-1} \times e^{-x} + 1 - \frac{e}{e-1}$ for $x \in [0, 1]$ with $r = (1.1, 1.1)$ and $n = 3$.

As a consequence, $\min_{v \in S_r^n} HVC_r(v, S_r^n)$ *is no greater than* $(r_1 - \tilde{x}_{\min,r}) \times \frac{\sum_{i=1}^{n+1} b_{n,i}}{n^2}$*, that is* $\frac{(r_1 - \tilde{x}_{\min,r}) \times (r_2 - f(\tilde{x}_{\max,r}))}{n^2}$.

We will also use the lower bounds on the maximum hypervolume of a single vector of the Pareto front for f either convex or ($L_{\min}$,$L_{\max}$)-bilipschitz stated in Chapter 4. We recall these results below, with a slight reformulation, for sake of completeness.

Proposition 1 *If f restricted to $[\tilde{x}_{\min,r}, \tilde{x}_{\max,r}]$ is convex, we have*

$$\max_{u \in PF} HV_r(u) \geq \frac{1}{2} \times HV_r(PF). \tag{3.3}$$

If f restricted to $[\tilde{x}_{\min,r}, \tilde{x}_{\max,r}]$ is $(L_{\min}, L_{\max})$-bilipschitz, we have

$$\max_{u \in PF} HV_r(u) \geq \frac{1}{2} \times \frac{L_{\min}}{L_{\max}} \times HV_r(PF). \tag{3.4}$$

For all n, for all $i \in [\![1, n+1]\!]$, the hypervolume associated with the reference point r_i^n of a vector u is equal to $\mathrm{HVI}_r(u, S_r^n)$ when $u \in \mathcal{G}_{r,i}^n$ and 0 otherwise. Thus, $\max_{u \in \mathrm{PF}} \mathrm{HV}_{r_i^n}(u)$ equals $\max_{u \in \mathcal{G}_{r,i}^n} \mathrm{HVI}_r(u, S_r^n)$. Additionally, the hypervolume associated with r_i^n of the Pareto front is simply the area of the gap region $\mathcal{G}_{r,i}^n$. Using the optimality of the n-optimal distribution, we can deduce from Lemma 1 and Proposition 1 an upper bound on the area of any gap region $\mathcal{G}_{r,i}^n$, and thus an upper bound on the optimality gap at iteration n.

By applying Proposition 1 in the convex case, we obtain the following theorem.

Theorem 2 *If f restricted to $[\tilde{x}_{\min,r}, \tilde{x}_{\max,r}]$ is convex, then the optimality gap is bounded from above as*

$$\lambda(\mathcal{G}_r^n) \leq 2 \times (r_1 - \tilde{x}_{\min,r}) \times (r_2 - f(\tilde{x}_{\max,r})) \times \frac{n+1}{n^2} \tag{3.5}$$

Proof *The n-optimal distribution S_r^n is a set of n feasible vectors of the objective space maximizing the hypervolume. As a consequence, for any $u \in F(\Omega)$, for any $v \in S_r^n$, the hypervolume of $S_r^n \setminus \{v\} \cup \{u\}$ is lower than the hypervolume of S_r^n. In other words, the hypervolume improvement with respect to the set $S_r^n \setminus \{v\}$ of any feasible vector u is lower than the one of v itself. Additionally, for any feasible u, the hypervolume improvement of u to S_r^n is lower than the hypervolume improvement of u to $S_r^n \setminus \{v\}$. Indeed, they are equal to the area of the region dominated by u but not by respectively S_r^n and $S_r^n \setminus \{v\}$. As a consequence, for any feasible vector u, $HVI_r(u, S_r^n)$ is lower than $HVI_r(v, S_r^n \setminus \{v\})$, that is $HVC_r(v, S_r^n)$.*

Let $\mathcal{G}_{r,i}^n$ be the i-th gap region of S_r^n. By Proposition 1, the area of $\mathcal{G}_{r,i}^n$ is lower than $2 \times \max_{u \in \mathcal{G}_{r,i}^n} HVI_r(u, S_r^n)$. We just proved that $\max_{u \in \mathcal{G}_{r,i}^n} HVI_r(u, S_r^n)$ is lower than $\min_{v \in S_r^n} HVC_r(v, S_r^n)$, and thus it is lower than $\frac{(r_1 - \tilde{x}_{\min,r})(r_2 - f(\tilde{x}_{\max,r}))}{n^2}$ by Lemma 1. Therefore, the area of any gap region $\mathcal{G}_{r,i}^n$ of S_r^n is lower than two times $\frac{(r_1 - \tilde{x}_{\min,r})(r_2 - f(\tilde{x}_{\max,r}))}{n^2}$. Since the optimality gap $HV_r(PF) - HV_r(S_r^n)$ is the sum of the areas of the $n+1$ gap regions of S_r^n, we can conclude.

By applying Proposition 1 in the bilipschitz case instead of the convex case, we obtain the following looser upper bound on the optimality gap.

Theorem 3 *If f restricted to $[\tilde{x}_{\min,r}, \tilde{x}_{\max,r}]$ is $(L_{\min}, L_{\max})$-bilipschitz, then the optimality gap is bounded from above by*

$$\lambda(\mathcal{G}_r^n) \leq 2 \times \frac{L_{\max}}{L_{\min}} \times (r_1 - \tilde{x}_{\min,r}) \times (r_2 - f(\tilde{x}_{\max,r})) \times \frac{n+1}{n^2} \tag{3.6}$$

3.4 Sufficient assumptions and generalization of the lower bound

In this section, we give sufficient conditions for deriving bounds on the optimality gap from Theorems 1 and 3. We also generalize Theorem 1 to Pareto fronts that only need to have some bilipschitz subsection.

3.4.1 Sufficient assumptions on the objective functions

First, we will examine the bilipschitz assumption, under which we have both a lower and an upper bound on the optimality gap. It is simple to prove that, as soon as both objective functions F_1 and F_2 are bilipschitz, the function f characterizing the Pareto front is bilipschitz too. We will say that F_i is ($L_{\min}$,$L_{\max}$)-bilipschitz when for all $X, Y \in \Omega$, the quantity $|F_i(Y) - F_i(X)|$ is between $L_{\min} \times \|Y - X\|$ and $L_{\max} \times \|Y - X\|$.

Proposition 2 *If F_1 and F_2 are respectively ($L_{\min,1}$, $L_{\max,1}$)-bilipschitz and ($L_{\min,2}$, $L_{\max,2}$)-bilipschitz.*

Then, f is ($\frac{L_{\min,2}}{L_{\max,1}}$, $\frac{L_{\max,2}}{L_{\min,1}}$)-bilipschitz.

Proof *Let x, y be in $[x_{\min}, x_{\max}]$. Since the vectors $(x, f(x))$ and $(y, f(y))$ belong to the Pareto front, they are feasible, and thus there exist $X, Y \in \Omega$ such that $(x, f(x)) = (F_1(X), F_2(X))$ and $(y, f(y)) = (F_1(Y), F_2(Y))$. Thus, $|f(y) - f(x)|$ equals $|F_2(Y) - F_2(X)|$, which is superior to $L_{\min,2}$ times $\|Y - X\|$, which is superior to $\frac{L_{\min,2}}{L_{\max,1}}$ times $|F_1(Y) - F_1(X)|$, that is $|x - y|$. Conversely, $|x - y|$ is superior to $\frac{L_{\min,1}}{L_{\max,2}}$ times $|f(y) - f(x)|$.*

However, as soon as an objective function has a critical point in the interior of the search space Ω, it is not bilipschitz. In particular, this is the case when the objective function is differentiable and has a local minimum. In that case, there is no guarantee that f will be bilipschitz. Setting the reference point such that the optimum in question is excluded can account for this problem. From a practical perspective, this is not a strong restriction.

Likewise, the convexity assumption on f is met as soon as both objective functions F_1 and F_2 are convex. This is a known result (see [38, p68]), but since we did not find a proof in the literature, we include the proof below.

Proposition 3 *If the search space Ω and the objective functions F_1 and F_2 are convex, then f is convex.*

Proof *Let $u := (u_1, u_2)$ and $v := (v_1, v_2)$ be two vectors of the epigraph of f. Since the vectors $(u_1, f(u_1))$ and $(v_1, f(v_1))$ belong to the Pareto front, they are feasible, and thus there exist $X, Y \in \Omega$ such that $(u_1, f(u_1)) = (F_1(X), F_2(X))$ and $(v_1, f(v_1)) = (F_1(Y), F_2(Y))$. Let note $Z := \frac{X+Y}{2}$. By convexity of Ω, Z also belongs to Ω. By convexity of F_1, $F_1(Z)$ is smaller than $\frac{F_1(X)+F_1(Y)}{2}$, that is $\frac{u_1+v_1}{2}$. Therefore, f being decreasing, $f\left(\frac{u_1+v_1}{2}\right)$ is smaller than $f(F_1(Z))$, that is $F_2(Z)$. By convexity of F_2, $F_2(Z)$ is smaller than $\frac{F_2(X)+F_2(Y)}{2}$, that is $\frac{f(u_1)+f(v_1)}{2}$ and thus than $\frac{u_2+v_2}{2}$ since u and v belong to the epigraph of f. Therefore, $\frac{u+v}{2}$ also belongs to $epif$. We can conclude that the epigraph of f, and thus the function f itself, are convex.*

Convexity of each objective is a sufficient but not a necessary condition. For example, the Pareto front of the test problem ZDT1 [80] is convex, while the second objective function is not.

3.4.2 Generalization of the lower bound to functions with a bilipschitz subsection

We prove that the optimality gap associated with any reference point r' dominating the reference point r provides a lower bound on the optimality gap associated with r.

Lemma 2 *Given a reference point r' that dominates the reference point r. The optimality gap $HV_r(PF) - HV_r(S_r^n)$ associated with r is bounded from below by the optimality gap $HV_{r'}(PF) - HV_{r'}(S_{r'}^p)$ associated with r'.*

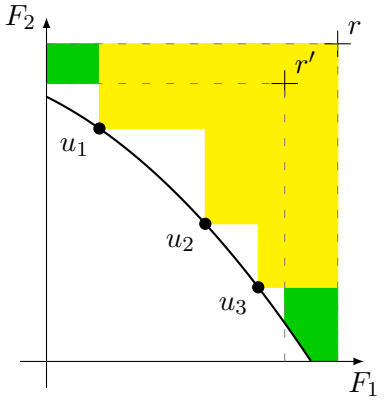


Figure 3.5: Illustration of the proof of Lemma 2.

Proof *For any set S, the difference $HV_r(PF) - HV_r(S)$ is greater than $HV_{r'}(PF) - HV_{r'}(S)$, see Figure 3.5. Indeed, the difference between the hypervolumes of S and of the Pareto front is the area of the intersection of the region dominated by the Pareto front but not by the set S and the region dominating the reference point. Since r' dominates r, the region of the objective space dominating r' is included in the region dominating r. Therefore, $HV_r(PF) - HV_r(S_r^n)$ is greater than $HV_{r'}(PF) - HV_{r'}(S_r^n)$, which is itself greater than $HV_{r'}(PF) - HV_{r'}(S_{r'}^n)$ by definition of S_r^n.*

Hence, to get a lower bound on the optimality gap for any reference point r, it suffices to find a reference point r' such that Theorem 1 applies and r' dominates r. As a consequence, as soon as any part of the Pareto front dominating r is bilipschitz, see Figure 3.6, Theorem 1 provides a, generally non-tight, lower bound on the optimality gap for any reference point that covers at least some part of a bilipschitz subsection.

Theorem 4 *Assume there exists a reference point r' dominating both r and the nadir point such that f restricted to $[\tilde{x}_{\min,r}, \tilde{x}_{\max,r}]$ is $(L_{\min},L_{\max})$-bilipschitz. Then, the optimality gap $HV_r(PF) - HV_r(S_r^n)$ is greater than $HV_{r'}(PF) \times \frac{L_{\min}}{L_{\max}} \times \frac{1}{n+1}$.*

In particular, this is the case when the function f is either convex or concave. Indeed, since f is strictly decreasing, it has finite and nonzero left and right derivatives everywhere outside the extremes $x_{\min}$ and $x_{\max}$. This generalization of the lower bound extends the scope of the study to non-continuous Pareto fronts such as piecewise continuous Pareto fronts. Lemma 2 and Theorem 1 also provide a way to find a lower bound for Pareto fronts which do not even have an explicit formula, contrary to the assumptions detailed in the preliminaries. It suffices that a part of the Pareto front has such a characterization.

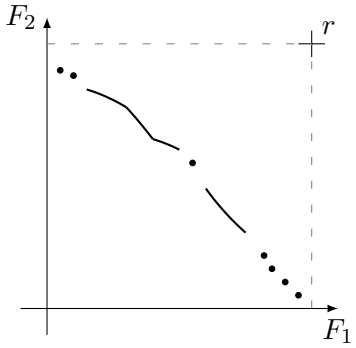


Figure 3.6: An example of Pareto front with a bilipschitz section dominating the reference point r. Both continuous sections are bilipschitz.

3.5 Experimental results

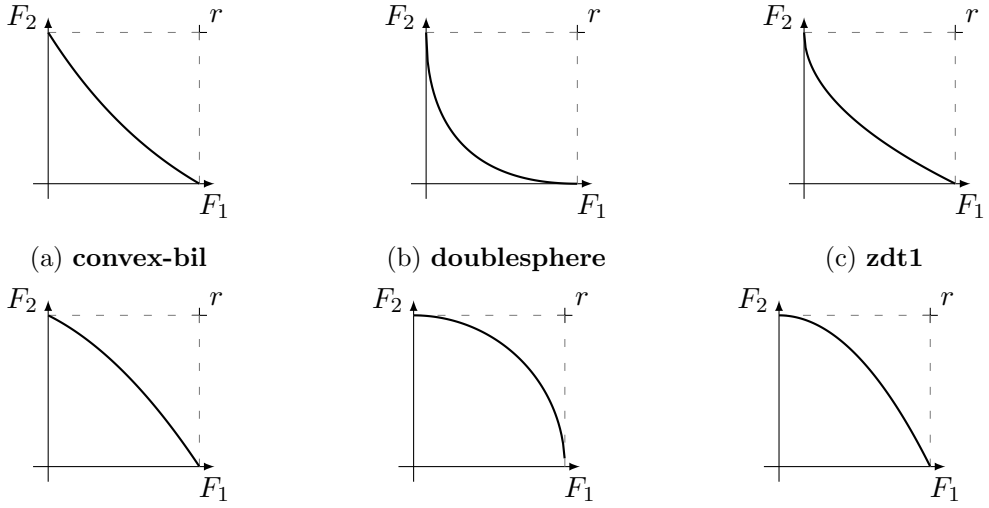
In this section, we will compare the dependency in n of the optimality gap $\mathrm{HV}_r(\mathrm{PF}) - \mathrm{HV}_r(S_r^n)$ and the theoretical bounds on six different Pareto fronts.

3.5.1 Benchmark Pareto fronts

We will look at six different Pareto fronts, among which three are convex and three are concave, see Figure 3.7 (a)-(f).

The **doublesphere** Pareto front (b) corresponds to the objective functions $F_1 : X \mapsto \|X - X_1^*\|_2^2$ and $F_2 : X \mapsto \|X - X_2^*\|_2^2$ with $\|X_1^* - X_2^*\|_2 = 1$, see [73]. **zdt1** (c) and **zdt2** (f) belong to the ZDT test suite [80] while **dtlz2** (e) belongs to the DTLZ test suite [32]. None of these Pareto fronts are bilipschitz. For this reason, we construct ourselves **convex-bil** (a) and **concave-bil** (d) to be bilipschitz convex and concave Pareto fronts, respectively. We can easily build constrained multi-objective optimization problems whose Pareto fronts are respectively **convex-bil** and **concave-bil**. For example, these Pareto fronts correspond to the problem of minimizing $F_1 : X \mapsto X_1$ and $F_2 : X \mapsto f(X_1) + \sum_{i=2}^{d} X_i$ for $X \in [0,1]^n$ with $f : x \mapsto \frac{e}{e-1} \times e^{-x} + 1 - \frac{e}{e-1}$ and $f : x \mapsto 1 - 0.5x - 0.5x^2$, respectively. The letter d represents the dimension of the search space.

We chose these Pareto fronts because they have a known analytic formula. It allows us to estimate the optimality gap for a n-optimal distribution for high n in reasonable time with high confidence.



(a) **convex-bil** (b) **doublesphere** (c) **zdt1**

(d) **concave-bil** (e) **dtlz2** (f) **zdt2**

Figure 3.7: The benchmark Pareto fronts, corresponding to the functions (a) : $f : x \mapsto \frac{e}{e-1} \times e^{-x} + 1 - \frac{e}{e-1}$ (b) : $f : x \mapsto 1 + x - 2 \times \sqrt{x}$ (c) : $f : x \mapsto 1 - \sqrt{x}$ (d) : $f : x \mapsto 1 - 0.5x - 0.5x^2$ (e) : $f : x \mapsto \sqrt{1 - x^2}$ (f) : $f : x \mapsto 1 - x^2$. The reference point is $r = (1, 1)$.

3.5.2 Computation of a n-optimal distribution and the optimality gap

Since S_r^n is a subset of the Pareto front, if we know the explicit representation of the Pareto front $\{(x, f(x)) : x \in [x_{\min}, x_{\max}]\}$, a n-optimal distribution S_r^n can be obtained from the first coordinates of its vectors, that is, $x_{n,i}$ for $i \in [\![1, n]\!]$. They are a solution of

$$\max_{x_1,\dots,x_n \in [x_{\min}, x_{\max}]} \mathrm{HV}_r(\{(x_1, f(x_1)), \dots, (x_n, f(x_n))\}) \ .$$

We do not solve this problem directly. We exploit the following parametrization to solve it faster: $\delta_i := x_i - x_{i-1}$ for all $i \in [\![2, n-1]\!]$, $\delta_1 := x_1 - x_{\min}$ and $\delta_n := x_{\max} - x_n$, see Figure 3.8. For **dtlz2**, we use a slightly different parametrization to cancel the bad conditioning: $\delta'_n := \mathrm{sgn}(\delta_n) \times \sqrt{|\delta_n|}$.

We use the algorithm CMA-ES with bounds between 0 and 1. These bounds do not guarantee that the x_i corresponding to the δ_i are in $[0, 1]$. We ensure that a x_i outside $[0, 1]$ does not contribute to the hypervolume by setting $f(x) = r_2$ outside $[0, 1]$. Additionally, to prevent having a flat fitness for δ_i such that the corresponding x_i are outside $[0, 1]$, we add the penalization $\sum_{i=1}^n (x-1)^2 \mathbf{1}_{x_i > 1} + x^2 \mathbf{1}_{x_i < 0}$. The source code is available at https://github.com/eugeniemarescaux/gecco2021.

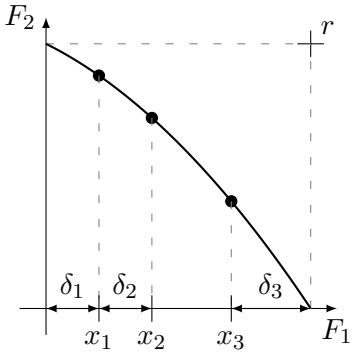
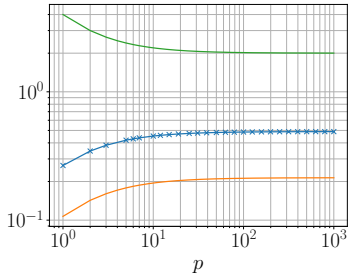


Figure 3.8: Illustration of the reparametrization of a set of n points of the Pareto front by δ.

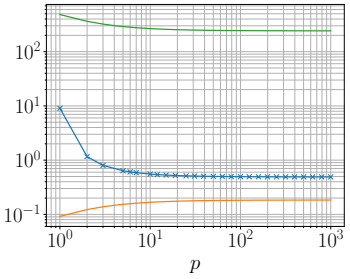
3.5.3 Results

Figure 3.9 shows n times the optimality gap of a n-optimal distribution for n between 1 and 1000. In the same figure, we also plotted for comparison n times the theoretical lower and upper bounds described in Section 3.3. For non-bilipschitz Pareto fronts, we exploit the generalization of the theoretical lower bound done in Section 3.4.2. For Pareto fronts which are neither bilipschitz nor convex, since there is no known theoretical upper bound, only two curves are represented. For a reference point r equal to the nadir point $(1, 1)$, i.e. in the first and third column of Figure 3.9, we see a close to constant shift between the curves of the empirical optimality gap and of the theoretical lower bound. Hence, the optimality gap evolves roughly as $1/(n + 1)$ even for the smallest values of n. For a reference point r equal to $(11, 11)$, i.e. in the second and fourth column of Figure 3.9, the shift between the empirical curve and the theoretical bound curve takes very different values for n small. However, the empirical curve seems to converge to a horizontal line, which is a marker that n times the optimality gap of S_r^n converges to a constant. For every Pareto front, the curve of the empirical optimality gap is almost horizontal at least for n greater than 10.

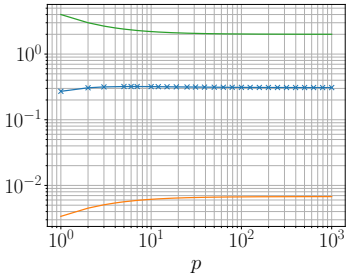
In the proof of Theorem 1, more precisely in the first phrase after the introduction of the notations, we neglected the area of the part of the total gap region which does not dominate the nadir point. It could explain why the optimality gap decreases with n in the same way as the theoretical lower bound for r equal to the nadir point $(1, 1)$ but not for r equal to $(11, 11)$.



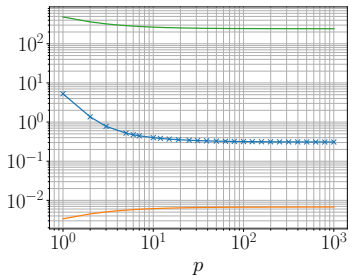
(a) **convex-bil** with $r = (1, 1)$



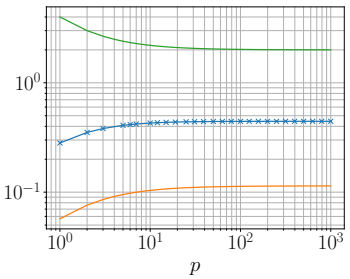
(b) **convex-bil** with $r = (11, 11)$



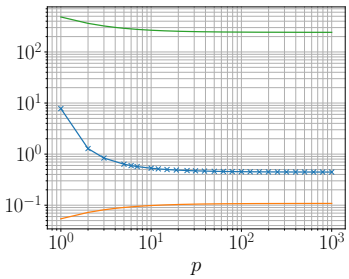
(c) **doublesphere** with $r = (1, 1)$



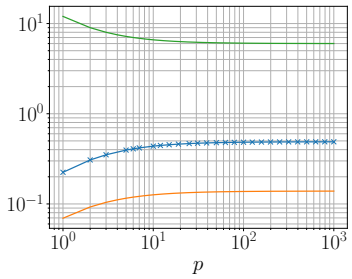
(d) **doublesphere** with $r = (11, 11)$



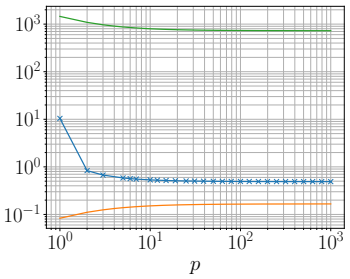
(e) **zdt1** with $r = (1, 1)$



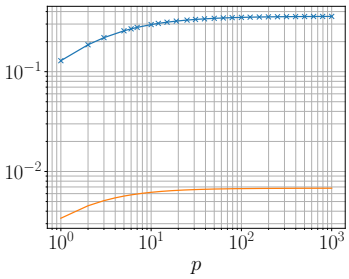
(f) **zdt1** with $r = (11, 11)$



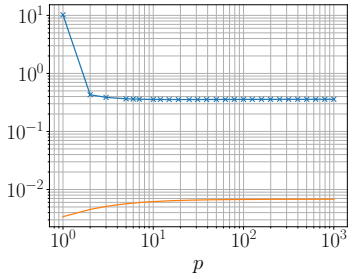
(g) **concave-bil** with $r = (1, 1)$



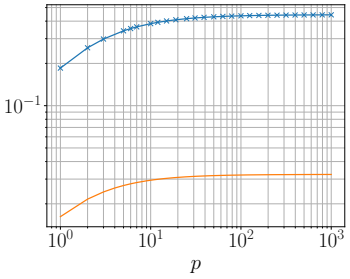
(h) **concave-bil** with $r = (11, 11)$



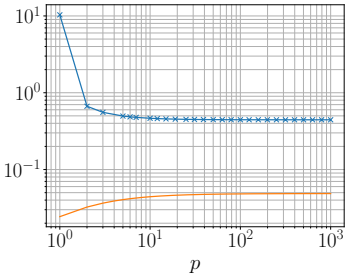
(i) **dtlz2** with $r = (1, 1)$



(j) **dtlz2** with $r = (11, 11)$



(k) **zdt1** with $r = (1, 1)$



(l) **zdt2** with $r = (11, 11)$

Figure 3.9: Comparison between n times the empirical optimality gap $\mathrm{HV}_r(\mathrm{PF}) - \mathrm{HV}_r(S_r^n)$ and its theoretical bounds on the Pareto fronts described in Section 3.5.1. The empirical optimality gap, $\mathrm{HV}_r(\mathrm{PF}) - \mathrm{HV}_r(S_r^n)$ is represented in blue and the theoretical lower and upper bounds are represented in orange and green, respectively.

3.6 Conclusion

We have proven that for a wide class of bi-objective Pareto fronts, the hypervolume of all the solutions visited by a bi-objective optimization algorithm converges to the hypervolume of the Pareto front in $\Omega(1/l)$, with l being the number of function evaluations. This is true for any algorithm, but also for objective functions as easy as convex quadratic functions, and for a search space of any dimension. The maximum rate of convergence to the entire Pareto front is slow compared to the convergence rate observed when converging to a single point in the Pareto set or, likewise, in single-objective optimization.

Several evolutionary algorithms achieve linear convergence on convex-quadratic functions [5]: their distance to the optimum stays close to $1/\alpha^n$ for some $\alpha > 1$ (typical convergence speeds depend on the search space dimension d and α rarely exceeds $1.2^{1/l}$). For random-search algorithms, the precision ϵ evolves on most functions as $\Theta(1/l^{1/l})$ with l being here the expected number of function evaluations to reach the ball of radius ϵ, see [5, Theorem 10.8].

A convergence rate which would seem slow for a single-objective algorithm does not come from an inefficient algorithm or from hard to optimize objective functions. The slowness is inherent to the set-quality indicators used. The simplicity of the proof, the fact that the convergence rate is $\Omega(1/l)$ for both the hypervolume indicator and the multiplicative approximation ratio, and the very general (weak) assumptions suggest that this is a fundamental limitation on the convergence rate in multi-objective optimization. This is however not a very surprising result when considering that, contrary to single-objective optimization, the goal is not to approximate a single vector, but an entire set.

We empirically observed on six functions that n times the optimality gap of n-optimal distributions converges rapidly to a constant, even for the two functions for which we have no upper bound. We suspect that the optimality gap is equivalent to a constant times $1/n$ on most if not all bi-objective optimization problems with a partially continuous Pareto front.

In general, theoretical lower bounds are quite useful in algorithm development. Apart from designing algorithms that actually reach the lower bound, they can in particular avoid futile but long lasting attempts to improve algorithms that already reach the bound. This requires a non-asymptotic bound, as presented in this chapter. Yet, the question arises how the lower bound on the optimality gap of the n-optimal distribution transfers to a practical algorithm. A practical algorithm faces at least two additional problems: it can only approximate any n-optimal distribution (but never reaches it), and the intersection between n-optimal distributions for different values of n is often small. Therefore, the presented bounds need to be carefully combined with bounds on the convergence speed towards points in a l-optimal distribution. We have currently no conjecture as to whether a convergence rate of $\Theta(1/l)$ is achievable by a real bi-objective algorithm.

For finite discrete Pareto fronts, we cannot talk of a convergence rate in $\theta(1/n)$ since n is bounded by the size of the Pareto front, that we will denote N. In that case, it

is trivial that for $n \neq N$ and for C_1 and C_2 respectively small and large enough, the optimality gap of n-optimal distributions is between $C_1 \times 1/n$ and $C_2 \times 1/n$. However, we can still expect that for some discrete Pareto fronts the optimality gap of n-optimal distributions will resemble $CONST \times 1/(n + CONST)$ for medium values of n. Indeed, we expect that for regular discretization of continous Pareto fronts, the impact on the optimality gap of the lack of precision of the approximation of a n-optimal distribution is negligable for $n \ll N$. Thus, the optimality gaps for the discrete approximation and for the original continuous Pareto front should be alike for these values of n. For example, consider a discrete Pareto front PF_d which is the regular discretization of a continuous linear Pareto front PF_c. If N is even, then the optimality gap for PF_d of any n-optimal distribution with n even is exactly $(1/(n+1) - 1/(N+1)) \times \mathrm{HV}_r(\mathrm{PF}_c)$. For $n \ll N$, it is close to $1/(n+1) \times \mathrm{HV}_r(\mathrm{PF}_c)$, the optimality gap in the continuous case.

CHAPTER 4

Bi-objective Hypervolume Based HV-ISOOMOO Algorithms Converge with At Least Sublinear Speed to the Entire Pareto Front

Adapted with minor modifications from a paper submitted to SIAM Optimization in August 2021 and currently at the second stage of the review process [62].

Abstract

In multi-objective optimization, one is interested in finding a good approximation of the Pareto set and the Pareto front, i.e the sets of best compromises in the decision and objective spaces, respectively. In this context, we introduce a new algorithm framework based on the hypervolume and called HyperVolume based Incremental Single-Objective Optimization for Multi-Objective Optimization (HV-ISOOMOO) for approximating the Pareto front with an increasing number of points. The hypervolume indicator is a set-quality indicator widely used for algorithms design and performance assessment. The class of HV-ISOOMOO algorithms builds a Pareto front approximation by a greedy maximization of this indicator. The greedy maximization consists here in iteratively finding a new solution with the largest hypervolume contribution to the set of previously found solutions. At each meta-iteration of HV-ISOOMOO algorithms, a single-objective subproblem is solved. We study the convergence to the entire Pareto front of HV-ISOOMOO under the assumption that these subproblems are solved perfectly. The convergence is defined as the convergence of the hypervolume of the sets of all meta-iterations incumbents towards the hypervolume of the Pareto front. We prove tight lower bounds on the speed of convergence for convex and bilipschitz Pareto fronts in $O(1/n^c)$ with $c = 1$ and $c \leq 1$, respectively. The index n denotes the number of meta-iterations of HV-ISOOMOO. For convex Pareto fronts, the convergence speed is $\Theta(1/n)$, namely the fastest convergence achievable by a bi-objective optimization algorithm. These are the first results on the speed of convergence of multi-objective optimization algorithms towards the entire Pareto front. We also analyze theoretically the asymptotic convergence behavior.

4.1 Introduction

Real-world problems often involve the simultaneous optimization of several conflicting objectives. The solution of such problems is the set of non-dominated *decision vectors* (vectors of the search space), the *Pareto set*. It is defined as the set of solutions that cannot be improved along one objective without degrading along another one. Its image in the objective space is the Pareto front. A decision maker is then often involved to choose, based on its preferences, a single best compromise. The shape of the Pareto front informs him on the trade-off between objectives. Many algorithms such as evolutionary algorithms approximate the Pareto front with a number of points fixed in the beginning. But some algorithms, in particular stemming from direct search methods [3, 16, 26, 27] approximate the Pareto set or the Pareto front with more and more points during the run. Ideally, the quality of the Pareto front approximation (e.g. measured with its size) increases gradually with time.

The speed of convergence towards a critical decision vector or a vector of the Pareto front has been examined for many algorithms such as $(1+1)$ evolutionary multi-objective algorithms [14] or Newton's method [43]. Convergence speeds are typically similar to the ones obtained for single-objective optimization. Both aforementioned publications describe convergence towards a single point and the first analysis is even reduced to the study of the convergence of a single-objective optimization algorithm [14]. The convergence of algorithms towards the whole Pareto set or front is of a different kind because iterates are sets and not points. Such convergence has already been theoretically investigated for some algorithms [26] and more abstract frameworks [75], but analysis of the speed of convergence are generally missing. Empirical studies typically focus on determining which algorithm is faster but more rarely provide information on the speed of convergence such as order of convergence or complexity. Yet, while often overlooked, the study of the speed of convergence, both theoretically and empirically, is important. In this context, it has been proven that the convergence of bi-objective optimization algorithms towards the whole Pareto front is always sublinear in the number of function evaluations, at least when measuring convergence with the hypervolume indicator Chapter 3 or the multiplicative ϵ-indicator [18], and thus much lower than typical speeds of convergence to a single point. The hypervolume and the multiplicative ϵ-indicator are set-quality indicators widely used in multi-objective optimization, both to guide algorithms and for performance assessment. The hypervolume is at the core of all known stricly Pareto-compliant indicators [81].

In this chapter, we introduce a new algorithm framework: HyperVolume based Incremental Single-Objective Optimization for Multi-Objective Optimization (HV-ISOOMOO). Algorithms following this framework try to greedily maximize the hypervolume by adding points approximating the largest hypervolume contribution achievable by a feasible vector. Such points are obtained by running a single-objective solver. Relying on single-objective optimization is a traditional approach in multi-objective optimization, the most significant example being scalarization [38]. A greedy approach to finding a set of n points with a large hypervolume is already used in the selection part of some

multi-objective optimization evolutionary algorithms such as SMS-EMOA [15]. The hypervolume of such discrete greedy approximation is proven to be at least $(e-1)/e$ times the one of a n-optimal distribution [64]. To the best of our knowledge, we provide the first continuous equivalent of this result.

The HV-ISOOMOO framework shares some high-level similarities with a recent hyperboxing algorithm [29]. At each meta-iteration, this hyperboxing algorithm finds a new point by solving the Pascoletti-Serafini scalarization problem defined by the upper corner and the diagonal of a box. The choice of the box relates to minimizing the ϵ-additive indicator. In contrast, HV-ISOOMOO algorithms are built to minimize the hypervolume.

We analyze an ideal version of HV-ISOOMOO algorithms where the single-objective solver returns a global optimum of the single-objective subproblems. This analysis is relevant for practical HV-ISOOMOO algorithms (whose construction is left for future publications) when the single-objective solver returns good approximations of global optima. Good approximations are typically unavailable when the single-objective solver does not converge to a solution, stops too early or converges to a local optimum. We investigate the speed of convergence of the ideal version of HV-ISOOMOO towards the whole Pareto front when measuring the convergence with the hypervolume. For convex and bilipschitz Pareto fronts, we prove that the convergence speed is in $O(1/n^c)$ with $c = 1$ and $c \leq 1$, respectively, with n being the number of single-objective optimization runs performed. For convex Pareto fronts, the convergence is exactly in $\Theta(1/n)$ as no bi-objective algorithm can converge faster to the Pareto front Chapter 3. Additionally, we prove that for simultaneously bilipschitz and smooth enough Pareto fronts doubling the number of points in the approximation divides the optimality gap by a factor which converges asymptotically to two. In the proof process, we obtain interesting intermediary results such as bounds on the normalized maximum hypervolume and a geometric interpretation of optimality conditions.

The chapter is organized as follows. In Section 4.2, we recall some classic concepts and define our assumptions. In Section 4.3, we introduce the HV-ISOOMOO framework together with connected mathematical abstractions and present numerical results on its convergence rate. In Section 4.4, we prove lower bounds on the so-called *normalized maximum hypervolume*, later used to investigate convergence. In Section 4.5, we derive lower bounds on the speed of convergence of HV-ISOOMOO algorithms under the assumption that every single-objective subproblem is solved perfectly and a theoretical result giving an insight on its asymptotic convergence behavior. For readability of the proof of this insight, we put some intermediary results and their proofs in the Appendices A and B.

Notations and conventions For $a, b \in \mathbb{N}$ with $a < b$, we note $[\![a; b]\!]$ the set $\{a, a+1, \ldots, b-1, b\}$. For a vector $u \in \mathbb{R}^2$, we note u_1 and u_2 respectively its first and its second coordinate. If the vector notation already contains an index, we separate the two indices with a comma. For simplicity sake, we often replace the set $\{u\}$ by u in the notations. We say that a function $f : \mathbb{R} \to \mathbb{R}$ is decreasing (resp. strictly decreasing) when for all $x < y$, we have $f(x) \geq f(y)$ (resp. $f(x) > f(y)$). We only consider two-dimensional

objective spaces and refer to the Lebesgue measure of a set as its area.

4.2 Background and assumptions

In this section, we recall some classic concepts of multi-objective optimization and define the assumptions used in this chapter.

4.2.1 Bi-objective optimization problems, the Pareto front and the hypervolume indicator

We consider a bi-objective minimization problem:

$$\min_{X \in \Omega \subset \mathbb{R}^d} F(X) \tag{4.1}$$

with $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}^2 : X \mapsto (F_1(X), F_2(X))$. We define two dominance relations for vectors in the objective space. We say that u weakly dominates v denoted by $u \preceq v$ if $u_1 \leq v_1$ and $u_2 \leq v_2$ and that u dominates v denoted by $u \prec v$ if $u \preceq v$ and $u \neq v$. A vector of the objective space $\mathbb{R}^2$ is said *feasible* when it belongs to $F(\Omega)$. Solving the optimization problem consists in finding a good approximation of the *Pareto front*, the set of non-dominated feasible vectors, $\{F(X) : X \in \Omega,\ \forall Y \in \Omega, F(Y) \not\prec F(X)\}$. We restrict ourselves to Pareto fronts with an explicit representation:

$$\mathrm{PF}_f = \{(x, f(x)) : x \in [x_{\min}, x_{\max}]\} \tag{4.2}$$

with $f : \mathbb{R} \mapsto \mathbb{R}$ decreasing. In the general case, the function f can be discontinuous but we assume from now on that f is continuous. Continuity will also be implied by the assumptions on f detailed in Section 4.2.3 which are used in the main results. When both objective functions have and reach finite global minimum values in the search space, that we denote respectively v_1 and v_2, we have $x_{\min} := v_1$ and $x_{\max} := \min_{X \in \Omega : F_2(X) = v_2} F_1(X)$.

We denote by $u_{\min} := (x_{\min}, f(x_{\min}))$ and $u_{\max} := (x_{\max}, f(x_{\max}))$ the extreme vectors of the Pareto front. We often focus on the part of the Pareto front dominating a reference point r. We only consider reference points dominated by at least one vector of the Pareto front, that we call *valid*. We denote by $\tilde{u}_{\min,r} := (\tilde{x}_{\min,r}, f(\tilde{x}_{\min,r}))$ and $\tilde{u}_{\max,r} := (\tilde{x}_{\max,r}, f(\tilde{x}_{\max,r}))$ the extremes vectors of the part of the Pareto front dominating a reference point r, with $\tilde{x}_{\min,r} := \min_{x \in [x_{\min}, x_{\max}] : f(x) \leq r_2} x$ and $\tilde{x}_{\max,r} := \min(x_{\max}, r_1)$. The vector $(x_{\max}, f(x_{\min}))$ is called the *nadir point*. All these notations are illustrated in Figure 4.1.

The *hypervolume* with respect to a reference point r of a set S of objective vectors is the Lebesgue measure of the region of the objective space dominated by S and strictly dominating the reference point r. We denote it $\mathrm{HV}_r(S)$. When no vector of the Pareto front dominates the reference point r, $\mathrm{HV}_r(S) = 0$ for any set S of feasible points of the objective space. This situation is prevented by choosing a valid reference point.

When there are more than two points in the Pareto front, any $(F_1(X_2), F_2(X_1))$ with $X_1 \in \arg\min_{X\in\Omega} F_1$ and $X_2 \in \arg\min_{X\in\Omega} F_2$ is a valid reference point.

The region of the objective space weakly dominated by S and dominating r (see the righthand plot of Figure 4.1) is denoted by $\mathcal{D}_S^r$ and formally defined as:

$$\mathcal{D}_S^r = \{w \in \mathbb{R}^2 : \exists u \in S : u \preceq w \prec r\} \ . \tag{4.3}$$

The hypervolume of a set S relative to the reference point r equals $\lambda(\mathcal{D}_S^r)$ with $\lambda(.)$ being the Lebesgue measure. The set $\mathcal{D}_S^r$ is the union of the $\mathcal{D}_u^r$ for $u \in S$, $\mathcal{D}_u^r$ being the rectangle $[u_1, r_1] \times [u_2, r_2]$ when u dominates r and $\emptyset$ otherwise, see the righthand plot of Figure 4.1. Note that the $\mathcal{D}_u^r$ are not disjoint.

We use the hypervolume to characterize the convergence of a sequence of sets (S_n) of objective vectors to the entire Pareto front. For a fixed valid reference point r, a sequence of sets (S_n) is said to converge to the Pareto front when the hypervolume difference $\mathrm{HV}_r(\mathrm{PF}_f) - \mathrm{HV}_r(S_n)$ converges asymptotically to 0. We define the *optimality gap* of a set S with respect to a valid reference point r as $\mathrm{HV}_r(\mathrm{PF}_f) - \mathrm{HV}_r(S)$. Another quantity of interest is how much adding a vector to a set affects its hypervolume. The *hypervolume improvement* with respect to r of the vector u to the set S is $\mathrm{HVI}_r(u, S) = \mathrm{HV}_r(S \cup \{u\}) - \mathrm{HV}_r(S)$. We also use the term hypervolume improvement to refer to the hypervolume increase of an increasing[1] sequence of sets $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ at iteration n, $\mathrm{HV}_r(\mathcal{S}_{n+1}) - \mathrm{HV}_r(\mathcal{S}_n)$.

4.2.2 Decomposition of the optimality gap using gap regions

The optimality gap is the Lebesgue measure of the *total gap region* introduced below.

Definition 2 *The* total gap region *of S with respect to a fixed valid reference point r, $\mathcal{G}_S^r$, is defined as the region of the objective space which dominates r and is weakly dominated by PF_f but not by S, namely $\mathcal{D}_{PF_f}^r \setminus \mathcal{D}_S^r$.*

We introduced $\mathcal{D}_S^r$ in (4.3). Its Lebesgue measure is $\mathrm{HV}_r(S)$.

The concept of *total gap region* is strongly connected to the concept of *search region* [57]. They can both be seen as the part of the objective space which may contain feasible nondominated points, but only the total gap region relies on the knowledge of the Pareto front. The total gap region is composed of the vectors of the search region which are weakly dominated by the Pareto front, a condition satisfied by all feasible objective-vectors.

When S is a subset of the Pareto front dominating the reference point r, the total gap region $\mathcal{G}_S^r$ has a particular shape which can be visualized in the rightmost plot of Figure 4.1. It can be decomposed into the disjoint union of $|S| + 1$ sets of the form $\mathcal{D}_{S'}^{r'}$ that are formally defined below.

Definition 3 (Gap regions and local nadir points) *Let $S = \{v_1, ..., v_n\}$ be a set of n distinct vectors of the Pareto front dominating a valid reference point r. Let σ be the permutation ordering the v_i by increasing F_1-values: $v_{\sigma(1),1} < v_{\sigma(2),1} < \ldots < v_{\sigma(n),1}$.*

[1] A sequence of set $\{A_n, n \geq 0\}$ is increasing if the inclusions $A_0 \subset A_1, \ldots \subset A_n \subset \ldots$ hold.

- *For all* $i \in [\![1, n+1]\!]$*, the* i*-th gap region of the set* S*,* $\mathcal{G}^r_{S,i}$*, is the set* $\mathcal{D}^{r_i^S}_{PF_f}$ *with the associated* local nadir point r_i^S *being* $r_1^S = (v_{\sigma(1),1}, r_2)$*,* $r_{n+1}^S = (r_1, v_{\sigma(n),2})$ *and* $r_i^S = (v_{\sigma(i),1}, v_{\sigma(i-1),2})$ *for all* $i \in [\![2, n]\!]$*.*
- *We refer to* $\mathcal{G}^r_{S,1}$ *and* $\mathcal{G}^r_{S,n+1}$ *as the left and the right extreme gap region of* S*, respectively.*

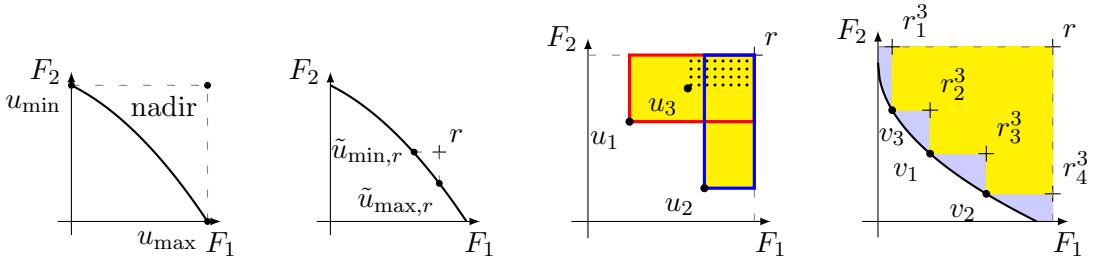


Figure 4.1: Illustration of notations. The extreme vectors $u_{\min}$ and $u_{\max}$ and the nadir point (leftmost) ; the extreme vectors relative to the reference point r $\tilde{u}_{\min,r}$ and $\tilde{u}_{\max,r}$ (left) ; three vectors u_1, u_2 and u_3 and the regions weakly dominated by them and dominating r, $\mathcal{D}^r_{u_1}$ ( ) , $\mathcal{D}^r_{u_2}$ ( ) and $\mathcal{D}^r_{u_3}$ ( ) (right) ; the set $S = \{v_1, v_2, v_3\}$, its four gap regions ( ) and the associated local nadir points (rightmost).

Local nadir points are also called *local upper bounds* [67, 57]. *Gap regions* are to *search zones* what total gap regions are to search regions. A gap region is composed of the vectors of a search zone which are weakly dominated by the Pareto front.

The left (resp. right) extreme gap region is empty when the left (resp. right) extreme vector of the Pareto front belongs to S. Non-extreme gap regions are never empty.

The total gap region of such a set S is the disjoint union of its gap regions: $\mathcal{G}^r_S = \dot{\cup}_{i=1}^{|S|+1} \mathcal{G}^r_{S,i}$. This decomposition of the total gap region, and thus of the optimality gap, is the cornerstone of our convergence analysis. Since the area of a gap region $\mathcal{G}^r_{S,i}$ is $\mathrm{HV}_{r_i^S}(\mathrm{PF}_f)$, we can write the optimality gap of a set S as the sum of $|S|+1$ hypervolumes of the Pareto front with respect to the local nadir points.

Lemma 3 *Let* S *be a set of* n *distinct vectors of the Pareto front dominating a valid reference point* r*. The optimality gap of* S *with respect to a valid reference point* r *can be decomposed as*

$$HV_r(PF_f) - HV_r(S) = \sum_{i=1}^{n+1} HV_{r_i^S}(PF_f) \ . \tag{4.4}$$

Proof *The optimality gap of* S *is the Lebesgue measure of the total gap region* $\mathcal{G}^r_S$*, which is the disjoint union of the gap regions* $\mathcal{G}^r_{S,i} = \mathcal{D}^{r_i^S}_{PF_f}$*. Therefore, the optimality gap equals* $\sum_{i=1}^{n+1} \lambda(\mathcal{D}^{r_i^S}_{PF_f}) = \sum_{i=1}^{n+1} HV_{r_i^S}(PF_f)$*.*

Additionally, as stated in Lemma 4, we can express the hypervolume improvement to a set S relative to a reference point r of any vector as its hypervolume relative to another reference point. It is a trivial assertion for vectors which do not dominate S. For other vectors, the reference point depends on the gap region to which the vector belongs.

Lemma 4 *Let S be a set of n distinct vectors of the Pareto front dominating a valid reference point r. For any u belonging to the i-th gap region of S, $\mathcal{G}^r_{S,i}$, we have*

$$HVI_r(u, S) = HV_{r_i^S}(u) \ . \tag{4.5}$$

Proof *The hypervolume improvement of any $u \in \mathcal{G}^r_{S,i}$ is the Lebesgue-measure of the intersection between $\mathcal{G}^r_{S,i}$ and $\mathcal{D}^r_u$. Therefore, it is equal to $\lambda(\mathcal{D}^{r_i^S}_u)$, that is $HV_{r_i^S}(u)$.*

4.2.3 Assumptions on the Pareto front and the objective functions

We present and discuss here the assumptions on the function f describing the Pareto front under which we derive convergence results. We typically assume that the function f is bilipschitz, convex or simultaneously bilipschitz and with a Hölder continuous derivative. Under any of these three assumptions, f is continuous. For the sake of conciseness, we transfer the properties of f to the Pareto front. For example, we call *convex Pareto front* a Pareto front described by a convex function. The ZDT test suite [80] contains two bi-objective optimization problems with a convex Pareto front: ZDT1 (see Figure 4.2) and ZDT4. We recall that a function f is Hölder continuous with exponent α, namely $\mathcal{C}^{1,\alpha}$, when there exists $H \geq 0$ such that $|f(x) - f(y)| \leq H \times |x - y|^\alpha$ for all x, y [41]. We note $[f]_\alpha$ the *minimum Hölder coefficient* with respect to the exponent α of a $\mathcal{C}^{1,\alpha}$ function f, that is $[f]_\alpha := \sup_{x \neq y} \frac{|f(x)-f(y)|}{|x-y|^\alpha}$. We recall that a function f is bilipschitz if there exists $L_{\min}$ and $L_{\max}$ with $L_{\max} \geq L_{\min} > 0$ such that for all $x, y \in [x_{\min}, x_{\max}]$, we have $L_{\min} \times |x - y| \leq |f(x) - f(y)| \leq L_{\max} \times |x - y|$. When needed, we detail the bilipschitz constants and refer to f as ($L_{\min}$,$L_{\max}$)-bilipschitz. For example, the problem of minimizing $F_1(X) = X^2$ and $F_2(X) = (X - 1)^2$ for $X \in [0.1, 0.9]$ has a bilipschitz Pareto front. Its representation is the function $f : x \mapsto 1 - \sqrt{x}$ for $x \in [F_1(0.1), F_1(0.9)] = [0.01, 0.81]$, which is ($L_{\min}$,$L_{\max}$)-bilipschitz with $L_{\min} := |f'(0.81)| = 0.555\ldots$ and $L_{\max} := |f'(0.01)| = 5$. In this example, the Pareto front would no longer be bilipschitz if the search space was extended to contain the two optima, for example to $[0, 1]$ or $\mathbb{R}$. This is representative of an important limitation of the bilipschitz assumption. To illustrate the range of this limitation, we introduce a general class of problems with non-artificial assumptions on F and Ω for which the Pareto front is not bilipschitz. Let F_1 and F_2 be two convex differentiable objective functions from $\mathbb{R}^d$ to $\mathbb{R}$ that have two unique distinct global optima X_1^* and X_2^*. Optimality conditions yield that $\nabla F_1(X_1^*) = 0$ and $\nabla F_2(X_2^*) = 0$. As $X_1^* \neq X_2^*$ and both F_1 and F_2 are convex, we have $\nabla F_1(X_2^*) \neq 0$ and $\nabla F_2(X_1^*) \neq 0$. As a consequence, as soon as the interior of Ω contains either X_1^* or X_2^*, the Pareto front of the problem $\min_{X \in \Omega}(F_1(X), F_2(X))$ is not bilipschitz. This reasoning could be applied to non-convex differentiable objective functions under the additional assumptions that X_1^* is not a critical point of F_2 and vice-versa.

We also talk of *affine Pareto fronts* when $f(x) = ax + b$ with $a < 0$ and $b \in \mathbb{R}$. As they form a line in the bi-objective case, they are usually referred to as *linear Pareto fronts*. They help to understand the results we prove on the asymptotic convergence behavior. The bi-objective optimization problem DTLZ1 of the DTLZ test suite [32] has an affine Pareto front. Affine Pareto fronts are a special case of ($L_{\min}$,$L_{\max}$)-bilipschitz Pareto fronts where $L_{\min} = L_{\max}$.

We remind below sufficient conditions on the search space and on the objective functions which guarantee that f is convex and bilipschitz.

Proposition 4 *Given a bi-objective minimization problem as in* (4.1) *whose Pareto front is described by a function f. If F_1 and F_2 are respectively $(L_{\min,1}\,,\ L_{\max,1})$-bilipschitz and $(L_{\min,2}\,,\ L_{\max,2})$-bilipschitz, then f is $\left(\frac{L_{\min,2}}{L_{\max,1}}\,,\ \frac{L_{\max,2}}{L_{\min,1}}\right)$-bilipschitz. If the search space Ω and the objective functions F_1 and F_2 are convex, then f is convex.*

The proofs of this proposition can be found for instance in Chapter 3. The conditions on F_1, F_2 and Ω are sufficient but non-necessary conditions. Indeed, adding small discontinuity in the objective functions far from the Pareto set makes them non-convex and non-bilipschitz without modifying the Pareto front.

Representing F_1-values on the abscissa and F_2-values on the ordinate instead of the converse is an arbitrary choice. When f is a bijection, if we chose to represent the F_2 values on the x-axis instead of on the y-axis, we would have another representation of the Pareto front : $\{(y, f^{-1}(y)) : y \in [f(x_{\max}); f(x_{\min})]\}$. If so, the inverse function f^{-1} would play the role of f. It is interesting to notice that the choice of the objective function represented on the horizontal axis does not impact whether the function characterizing the Pareto front is bilipschitz or convex. Indeed, f being bilipschitz is equivalent to both f and f^{-1} being lipschitz. Additionally, given that the function f is decreasing, f being convex is equivalent to its inverse f^{-1} being convex.

4.3 The HV-ISOOMOO framework

We introduce the HV-ISOOMOO framework. We formalize in Definition 5 its mathematical abstraction under the assumption that every single-objective subproblem is solved perfectly, that we call the *greedy set sequences*. Finally, we present numerical results on the rate of convergence of these greedy set sequences towards the Pareto front.

4.3.1 Presentation of the framework

The HV-ISOOMOO framework builds incrementally an increasing sequence $(\mathcal{I}_n)_{n\in\mathbb{N}^*}$ of sets of vectors of the objective space. The pseudocode of HV-ISOOMOO is given in Algorithm 1, where the current value of $\mathcal{I}_n$ is denoted by $\mathcal{I}$. At each so-called meta-iteration, a single-objective maximization solver SOOPTIMIZER (line 2 in Algorithm 1) is run on the criterion $X \in \Omega \subset \mathbb{R}^d \mapsto J(\mathcal{I}, X)$ and the resulting solution is added to $\mathcal{I}$ (line 3 in Algorithm 1). We use the term meta-iteration to separate between the (meta-)iterations of HV-ISOOMOO and the iterations of SOOPTIMIZER. Since the set

$\mathcal{I}$ is composed of the final objective incumbents of previous runs of SOOPTIMIZER and (ideally) provides an approximation of the Pareto front, we call it *final incumbents Pareto front approximation*.

The single-objective optimization procedure may vary between meta-iterations. More precisely, the run of SOOPTIMIZER depends on data about precedent runs stored in D (line 4 in Algorithm 1). This allows to alternate between various single-objective optimization solvers with different features, but also to adapt their initialization. This could be done by storing in D an iteration index and the final search space incumbents of SOOPTIMIZER runs.

Algorithm 1 HV-ISOOMOO Framework

1: **while** not **stopping criterion do**
2: $Y, d \leftarrow$ SOOPTIMIZER $(X \mapsto J(\mathcal{I}, X),\ D)$
3: $\mathcal{I} \leftarrow \mathcal{I} \cup \{F(Y)\}$ # update of the approximation of the Pareto front
4: $D \leftarrow D \cup \{d\}$ # update of the data collected
5: **end while**

The criterion J is chosen such that its maximization is *compliant* with the maximization of the hypervolume improvement with respect to a reference point r as defined below.

Assumption 1 *(Compliance to HVI_r maximization) The maximization of a criterion J as in HV-ISOOMOO is* compliant *with the maximization of HVI_r if for any set $\mathcal{I}$ of objective vectors, we have*

$$\operatorname{argmax}_{X \in \mathbb{R}^d} J(\mathcal{I}, X) = \operatorname{argmax}_{X \in \mathbb{R}^d} HVI_r(F(X), \mathcal{I}) \ . \tag{4.6}$$

In other words, at each meta-iteration n, an HV-ISOOMOO algorithm seeks a feasible vector maximizing the hypervolume improvement to the final incumbents Pareto front approximation $\mathcal{I}_n$. Ideally, when n goes to infinity, the non-dominated subset of $(\mathcal{I}_n)_{n \in \mathbb{N}^*}$ converges to the (entire) Pareto front, a set which maximizes the hypervolume. In a nutshell, HV-ISOOMOO algorithms try to approximate the Pareto front with a greedy approach.

Definition 4 *We define the convergence of an HV-ISOOMOO algorithm as the convergence of $HV_r(\mathcal{I}_n)$ towards $HV_r(PF_f)$.*

The performance of a specific HV-ISOOMOO algorithm depends crucially on the choice of the criterion J. In this respect, $\mathrm{HVI}_r(\mathcal{I}, F(.))$ itself is not a good candidate for $J(\mathcal{I}, .)$. Indeed, it is constant equal to zero in the region dominated by $\mathcal{I}$, which makes it difficult to optimize. A criterion whose maximization is compliant with the maximization of the hypervolume improvement and designed to be easier to optimize has already been introduced in [74] under the name *uncrowded hypervolume improvement* (UHVI). For $F(X)$ not dominated by $\mathcal{I}$, UHVI_r and HVI_r are equal. Otherwise, in the region where the hypervolume improvement is null, UHVI_r is negative and equals minus

the distance to the empirical non-dominated front of the set $\mathcal{I}$ relative to r. It is easy to see that UHVI_r satisfies (4.6).

The choice of SOOPTIMIZER also plays a key role in the performance of an HV-ISOOMOO algorithm. In this chapter, we analyze the HV-ISOOMOO framework under the assumption of perfect single-objective optimization formalized below.

Assumption 2 (Perfect single-objective optimization) *At every meta-iteration n, for any final incumbents Pareto front approximation $\mathcal{I}_n$, the run of SOOPTIMIZER (line 2 in Algorithm 1) returns $Y \in \text{argmax}_{X \in \Omega} J(\mathcal{I}_n, F(X))$.*

The assumption of perfect single-objective optimization is reminiscent of the assumption of perfect line search which is sometimes used in the analysis of gradient based methods [30]. Under this assumption, all choices of criterions verifying Assumption 1 are equivalent.

Convergence results under perfect conditions are useful to investigate the convergence (speed) we can expect with a given framework and in turn guide the construction of practical algorithms. A good theoretical convergence speed suggests that an approach is worth exploring while an unexplained gap between practical and ideal convergence speeds suggests that there is room for improvement. In our case, convergence rates under perfect conditions also give us an idea of what the convergence rates with respect to meta-iterations of practical HV-ISOOMOO algorithms will look like when SOOPTIMIZER returns good approximations of global optima.

4.3.2 Greedy sets and greedy set sequences

We introduce below mathematical abstractions of the HV-ISOOMOO framework under Assumption 2 of perfect single-objective optimization, *greedy sequences* and *greedy set sequences*.

Definition 5 (Greedy sequence, set sequence and vector) *Given a valid reference point r, we define as greedy sequence relative to r, a sequence $(v_n)_{n \in \mathbb{N}^*}$ satisfying*

$$v_1 \in \arg\max_{v \in F(\Omega)} HV_r(v) \text{ and} \tag{4.7}$$

$$v_{n+1} \in \arg\max_{v \in F(\Omega)} HV_r(\{v_1, \cdots, v_n, v\}) \text{ for all } n \geq 1 \ . \tag{4.8}$$

We refer to the vectors v_n of a greedy sequence $(v_n)_{n \in \mathbb{N}^*}$ as greedy vectors. The greedy set sequence $(\mathcal{S}_n)_{n \in \mathbb{N}^*}$ associated to the greedy sequence $(v_n)_{n \in \mathbb{N}^*}$ is composed of the greedy sets $\mathcal{S}_n := \{v_k, k \leq n\}$.

When considering a greedy set $\mathcal{S}_n$, we denote the i-th gap region and the associated local nadir point defined in Definition 3 by $\mathcal{G}_i^n$ and r_i^n, respectively.

There is a bijection between greedy sequences and greedy set sequences. The n-th element of the greedy sequence $(v_n)_{n \in \mathbb{N}^*}$ associated to a greedy set sequence $(\mathcal{S}_n)_{n \in \mathbb{N}^*}$ is simply the unique element of $\mathcal{S}_n \setminus \mathcal{S}_{n-1}$ if $n > 1$ and of $\mathcal{S}_1$ if $n = 1$.

The recurrence relation of the greedy sequence (4.8) is equivalent to v_{n+1} belonging to $\arg\max_{v\in F(\Omega)} \mathrm{HVI}_r(v, \mathcal{S}_n)$ for all $n \geq 1$. It is immediate to see that under Assumption 2, the final incumbents generated by HV-ISOOMOO constitute *a* greedy sequence while the final incumbents Pareto front approximations form the associated greedy set sequence $(\mathcal{I}_n)_{n\in\mathbb{N}^*}$. The indices n of both sequences correspond to HV-ISOOMOO meta-iterations. In this chapter, we derive convergence results for greedy set sequences, which transfer to HV-ISOOMOO under Assumption 2.

Since the hypervolume indicator associated to a valid reference point is strictly Pareto-compliant (see [58]), this sequence is composed of vectors of the Pareto front.

Proposition 5 *If the Pareto front is described by a lower semi-continuous function f, then any vector of a greedy sequence relative to a valid reference point r belongs to the Pareto front. Consequently, for such Pareto front and reference point and under Assumption 2, all final incumbents Pareto front approximations $\mathcal{I}_n$ of an HV-ISOOMOO algorithm relative to r are subsets of the Pareto front.*

Proof *Since for any valid reference point r, $HV_r(.)$ is strictly Pareto-compliant [58], its maximum and the maximum of every function of the form $v \mapsto HV_r(S \cup \{v\})$ are non-dominated and belong to the Pareto front. Thus, in particular, a vector v verifying either* (4.7) *or* (4.8) *belongs to the Pareto front.*

We can express the hypervolume improvement of a greedy set sequence at iteration $n+1$, $\mathrm{HV}_r(\mathcal{S}_{n+1}) - \mathrm{HV}_r(\mathcal{S}_n)$, as the maximum of $n+1$ hypervolume maximization problems.

Lemma 5 *Let $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ be a greedy set sequence relative to a valid reference point r. The hypervolume improvement of $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ at iteration $n+1$ equals*

$$HV_r(\mathcal{S}_{n+1}) - HV_r(\mathcal{S}_n) = \max_{i\in[\![1,n+1]\!]} \max_{u\in PF_f} HV_{r_i^n}(u) \ . \tag{4.9}$$

Proof *The hypervolume improvement $HV_r(\mathcal{S}_{n+1}) - HV_r(\mathcal{S}_n)$ is the hypervolume improvement of v_{n+1} to $\mathcal{S}_n$, namely the highest hypervolume improvement of a vector $u \in PF_f$ to $\mathcal{S}_n$ by* (4.8) *and Proposition 5. We can reformulate the hypervolume improvement of any vector u to $\mathcal{S}_n$ as $\max_{i\in[\![1,n+1]\!]} HV_{r_i^n}(u)$ by Lemma 4 since gap regions are disjoints and $HV_{r_i^n}(.)$ is null outside the i-th gap region of $\mathcal{S}_n$.*

Similarly, the problem of maximizing the hypervolume improvement to a greedy set $\mathcal{S}_n$ can be rewritten as the maximum of a finite number of hypervolume maximization problems.

Lemma 6 *At any iteration n, the recurrence relation satisfied by v_{n+1}, i.e.* (4.8)*, can be reformulated as*

$$v_{n+1} \in \arg \max_{i\in[\![1,n+1]\!]} \max_{u\in PF_f} HV_{r_i^n}(u) \ . \tag{4.10}$$

Proof *It is a direct consequence of the fact that the hypervolume improvement of any vector u to $\mathcal{S}_n$ is $\max_{i\in[\![1,n+1]\!]} HV_{r_i^n}(u)$, as stated in the proof of Lemma 5, and that $v_{n+1} \in PF_f$ by Proposition 5.*

As a consequence, we can infer from [7, Theorem 1] that as soon as the Pareto front is lower semi-continuous, there exists a greedy sequence, and thus a greedy set sequence.

Proposition 6 *If the Pareto front is described by a lower semi-continuous function f, then there exists a greedy sequence $(v_n)_{n\in\mathbb{N}^*}$ relative to any valid reference point.*

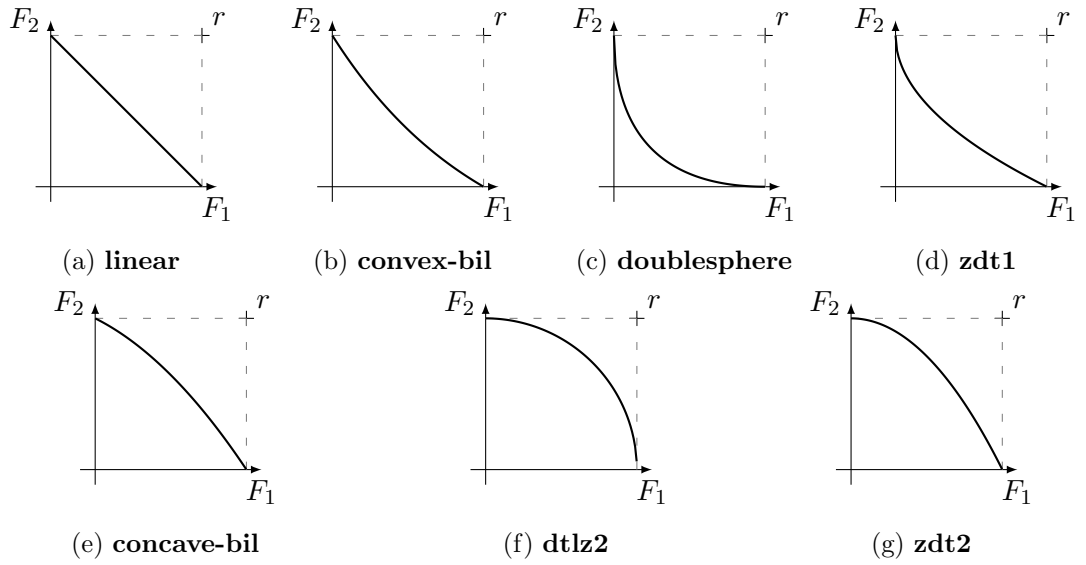
Proof *If f is lower semi-continuous, then for any valid reference point r, the maximum of $HV_r(.)$ exists, see [7, Theorem 1]. Therefore, there exists a vector verifying* (4.7) *and the problem of maximizing the maximum of a finite number of hypervolume functions defined in* (4.10) *admits a solution. Since* (4.8) *and* (4.10) *are equivalent by Lemma 6, we can build a sequence $(v_n)_{n\in\mathbb{N}^*}$ verifying* (4.7) *and* (4.8)*, namely a greedy sequence.*

Yet, in general, there exists more than one greedy sequence, and thus greedy set sequence. For example, there are infinitely many greedy sequences associated to any affine Pareto front with a reference point dominating the nadir point. This statement relies on the fact that the unique maximizer of the hypervolume relative to a reference point r dominating the nadir point is the middle of the section of the Pareto front dominating r, see [8, Theorem 5]. As a consequence, the middle of the section of the Pareto front dominating r is the only candidate for v_1 but v_2 can be either at $1/4$ or at $3/4$ of this section. Similarly, v_3 has to be in the position where v_2 is not but v_4 can be at $1/8$, $3/8$, $5/8$ or $7/8$ of the section of the Pareto front dominating r. For any n, we can find an iteration m such that v_m can be placed at 2^n different points, whatever the $m-1$ first terms of the greedy sequence are.

4.3.3 Numerical results

In this section, we investigate empirically the convergence rate of HV-ISOOMOO algorithms with respect to meta-iterations on seven benchmark Pareto fronts (see Figure 4.2). We try to be as close as possible to the Assumption 2 of perfect single-objective optimization. We iteratively find a vector of the Pareto front that we estimate close to a global optimum of the single-objective optimization subproblem of maximizing the hypervolume improvement to the vectors found so far. More precisely, we estimate that this vector differs from a global optimum by less than 10^{-12} (measured in terms of objective function values). We assimilate these vectors to greedy vectors in the following.

We consider the six concrete Pareto fronts with an explicit representation: $\{(x, f(x)) : x \in [0,1]\}$ considered in Chapter 3, plus an affine Pareto front (see Figure 4.2). The Pareto fronts **zdt1**, **zdt2** and **dtlz2** belong to the ZDT and DTLZ test suites [80, 32]. Four of the Pareto fronts examined are convex (**linear**, **convex-bil**, **doublesphere** and **zdt1**), while three are bilipschitz (**linear**, **convex-bil** and **concave-bil**). Two are neither convex nor bilipschitz (**dtlz2** and **zdt2**), and thus do not belong to the class of



(a) **linear** (b) **convex-bil** (c) **doublesphere** (d) **zdt1**

(e) **concave-bil** (f) **dtlz2** (g) **zdt2**

Figure 4.2: The benchmark Pareto fronts, corresponding to the functions (a): $f : x \mapsto 1-x$, (b): $f : x \mapsto \frac{e}{e-1} \times e^{-x} + 1 - \frac{e}{e-1}$, (c): $f : x \mapsto 1 + x - 2 \times \sqrt{x}$, (d): $f : x \mapsto 1 - \sqrt{x}$, (e): $f : x \mapsto 1 - 0.5x - 0.5x^2$, (f): $f : x \mapsto \sqrt{1-x^2}$ and (g): $f : x \mapsto 1 - x^2$ for $x \in [0, 1]$. The reference point is $r = (1, 1)$.

Pareto fronts investigated theoretically in this chapter. We take the nadir point $(1, 1)$ as reference point.

Greedy vectors defined in Definition 5 are true solutions of single-objective problems involving the objective functions F_1 and F_2 in the search space $\Omega \subset \mathbb{R}^d$. In order to compute an approximation of the greedy vectors, we exploit the explicit representation f of the Pareto front. Any greedy vector v_n belongs to the Pareto front (see Proposition 5), and thus is of the form $(v_{n,1}, f(v_{n,1}))$. We compute (an approximation of) $v_{n,1}$ by solving numerically the one-dimensional optimization problems defined in (4.12) and (4.13).

For all n, let note σ_n the permutation of $[\![1, n]\!]$ which orders the vectors of $\mathcal{S}_n$ by increasing F_1-values and the so-called *ordered greedy set F_1-values*:

$$w^n_{i,r} := v_{\sigma_n(i),1} \text{ for } i \in [\![1, n]\!],\ w^n_{0,r} := \tilde{x}_{\min,r} \text{ and } w^n_{n+1,r} := \tilde{x}_{\max,r} \ . \tag{4.11}$$

The following single-objective optimization problems

$$v_{1,1} \in \arg \max_{x \in [0,1]} \mathrm{HV}_r((x, f(x))) \tag{4.12}$$

$$v_{n+1,1} \in \arg \max_{i \in [\![1,n+1]\!]} \max_{x \in [w^n_{1,r}, w^n_{n+1,r}]} \mathrm{HV}_{w^n_{n+1,r}, f(w^n_{1,r})}((x, f(x))) \tag{4.13}$$

are solved using the SLSQP version implemented in the python module `scipy.optimize` with the stopping criterions being set to `ftol:1e-13` and `maxiter:1000`. For each problem, we run the solver SLSQP three times starting it uniformly at random in the

search interval. We ensured that the objective functions optimized did not differ by more than 10^{-12} between the runs. The source code is available at https://github.com/eugeniemarescaux/hypervolume-greedy-sequences.

In Figure 4.3, we display the optimality gap of $\mathcal{S}_n$ with respect to the meta-iteration n for all benchmark Pareto fronts. We rely on the analytical expression of the Pareto front PF_f to compute its hypervolume and the optimality gap. We observe very similar convergence rates for all benchmark Pareto fronts. They are all close to the $(1/(2n))_{n\in\mathbb{N}}$ line in the log-log scale. It is compliant with theory for the **linear** Pareto front. Indeed, let define the sequence of indices $(n_k)_{k\in\mathbb{N}}$ such that $n_0 = 1$ and $n_{k+1} = 2n_k + 1$. For an affine Pareto front and the nadir chosen as reference point, there is a unique greedy set $\mathcal{S}_{n_k}$, which consists of objective-vectors regularly distributed on the Pareto front and is equal to the n_k-th optimal distribution. A direct consequence is that for any such n_k, the optimality gaps of $\mathcal{S}_{n_k}$ and of the n_k-th optimal distribution are equal. For the **linear** Pareto front and $r = (1, 1)$, the optimality gap of the n-th optimal distribution is $1/(2n + 2)$ [8, Theorem 5].

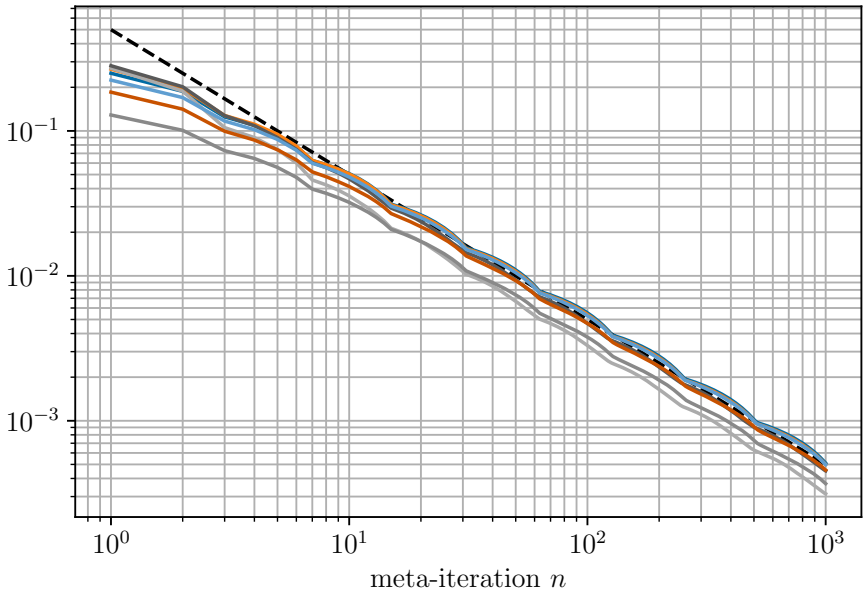


Figure 4.3: The optimality gap of $\mathcal{S}_n$ and $1/2n$ (the dashed line) for n up to 1000 for all the benchmark Pareto fronts: **linear** —, **convex-bil** —, **doublesphere** —, **zdt1** —, **concave-bil** —, **zdt2** — and **dtlz2** —.

Our underlying assumption of perfect single-objective optimization is theoretical and cannot be verified by a real solver. For HV-ISOOMOO algorithms coupled with a real single-objective optimization solver SOOPTIMIZER, we can still display convergence graphs with respect to meta-iterations as in Figure 4.3. We can also display convergence graphs with respect to cumulated SOOPTIMIZER iterations.

In Figures 4.4 and 4.5, we compare the optimality gap of $\mathcal{S}_n$ with the smallest opti-

mality gaps achievable by a set of n points. These are the optimality gaps of *n-optimal distributions* [8], the sets of n objective vectors with the highest hypervolume achievable by a set of this cardinal. We reuse the optimality gaps of n-optimal distributions which were computed for Chapter 3, for $n = 1$, 2, 3, 5, 6, 7, 10, 12, 15, 19, 25, 31, 39, 50, 63, 79, 100, 125, 158, 199, 251, 316, 398, 501, 630, 794 and 1000. The details of the computation method can be found in [63, Section 5.2].

In Figure 4.4, we display the optimality gap of $\mathcal{S}_n$ with respect to the meta-iteration index n for the **linear** Pareto front only, along with the optimality gap of n-optimal distributions. The curve of the optimality gap of the greedy set sequence $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ follows the one of the n-optimal distribution, moving away and getting closer periodically. This is what we would expect theoretically, as detailed above.

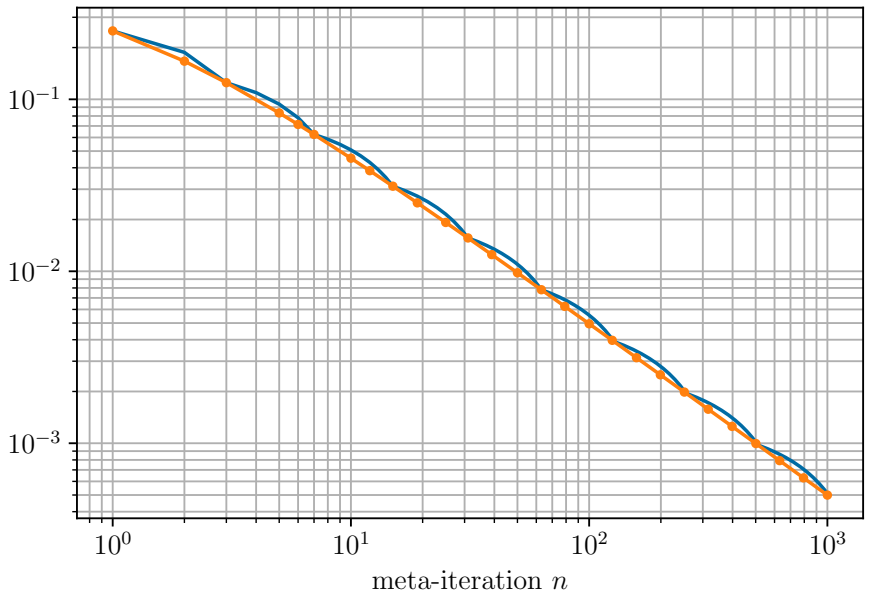


Figure 4.4: The optimality gap of $\mathcal{S}_n$ — and of a n-optimal distribution —•— for n up to 1000 for the **linear** Pareto front.

In Figure 4.5, we display the relation between the optimality gap of $\mathcal{S}_n$ and of n-optimal distributions for all benchmark Pareto fronts. We see similar fluctuations as for the **linear** Pareto front, with the same periodicity. At the bottom of the curve, the optimality gap of $\mathcal{S}_n$ is only a few percent larger than the one of a n-optimal distribution. In the worst case, that is for **doublesphere** Pareto front and for $n = 2$, the optimality gap of $\mathcal{S}_n$ is only 23% larger than the one of a n-optimal distribution. For $n \geq 10$, what is lost in proportion by taking $\mathcal{S}_n$ instead of a n-optimal distribution is never significantly higher than for the **linear** Pareto front for the displayed value of n. We conjecture that it is true for all $10 \leq n \leq 1000$. The **linear** curve stops reaching regularly the value 1, in contrast to what is known theoretically. It is explained by the discretization in n.

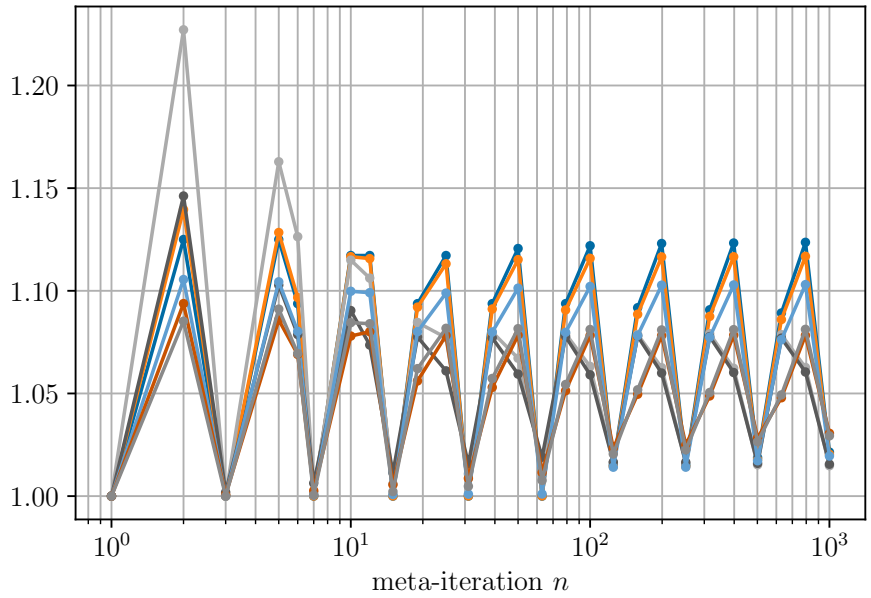


Figure 4.5: The optimality gap of $\mathcal{S}_n$ normalized by the optimality gap of a n-optimal distribution for n up to 1000 for all benchmark Pareto fronts: **linear** —, **convex-bil** —, **doublesphere** —, **zdt1** —, **concave-bil** —, **zdt2** — and **dtlz2** —.

4.4 Lower bounds on the normalized maximum hypervolume

In this section, we provide bounds on the maximum hypervolume achievable by a single feasible vector normalized by the maximum hypervolume of a feasible set: $\frac{\max_{u \in \mathrm{PF}_f} \mathrm{HV}_r(u)}{\mathrm{HV}_r(\mathrm{PF}_f)}$. We refer to this ratio as the *normalized maximum hypervolume* with respect to r. Bounds on the normalized maximum hypervolume are exploited in Section 4.5 to provide bounds on the speed of convergence of greedy set sequences towards the Pareto front.

4.4.1 Lower bound on the normalized maximum hypervolume for convex Pareto fronts

The hypervolume relative to a reference point r of a vector $u = (x, f(x))$ of the Pareto front is $\mathrm{HV}_r(u) = (r_1 - x) \times (r_2 - f(x))$. From this simple formula, we derive in the next proposition necessary conditions for a vector of the Pareto front to be an hypervolume maximizer when f has at least left and right derivatives in x^*.

Proposition 7 *Let $x^* \in]x_{\min}, x_{\max}[$ such that $u^* := (x^*, f(x^*))$ maximizes the hypervolume with respect to a valid reference point r. If the function f describing the Pareto*

front admits left and right derivatives in x^*, respectively $f'_-(x^*)$ and $f'_+(x^*)$, then

$$-f'_+(x^*) \leq \frac{r_2 - f(x^*)}{r_1 - x^*} \leq -f'_-(x^*) \ . \tag{4.14}$$

Proof *We define the function $HV_{x,r}(.)$ as $x \mapsto HV_r((x, f(x)))$. If x^* maximizes $HV_{x,r}(.)$, then the left and the right derivatives of $HV_{x,r}(.)$ are positive and negative, respectively. By replacing the left and right derivatives of $HV_{x,r}(.)$ by their explicit formulas and reorganizing the terms we obtain* (4.14).

Equation (4.14) states that the slope of the diagonal of the rectangle $\mathcal{D}^r_{u^*}$ is between the absolute values of the slopes of the right and the left derivatives of f at x^* (see the middle plot of Figure 4.6). To the best of our knowledge, this geometric interpretation is new. It becomes simpler when f is differentiable. Then, the absolute value of the slope of the tangent of the front at a non-extreme vector u^* is equal to the slope of the diagonal of the rectangle $\mathcal{D}^r_{u^*}$ (see the lefthand plot of Figure 4.6).

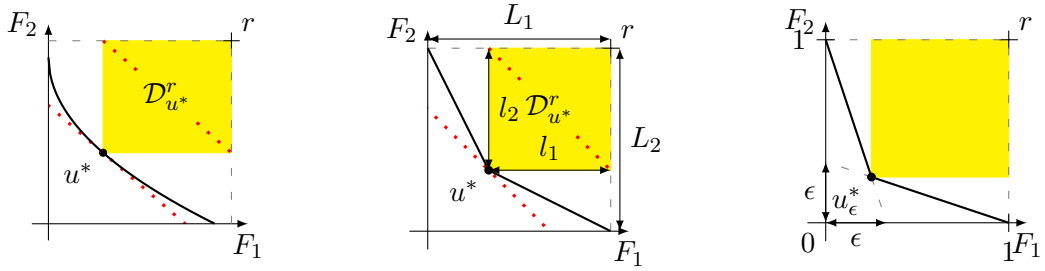


Figure 4.6: Left and middle: Two convex Pareto fronts and their respective hypervolume maximizers u^*, one differentiable (left) and one non-differentiable (middle). The slopes of the two dotted lines, namely PF_g and the diagonal of $\mathcal{D}^r_{u^*}$, are equal. Right: The Pareto front PF_ϵ and the hypervolume maximizer u^*_ϵ for $\epsilon = 1/3$ and $r = (1, 1)$.

Corollary 2 *Let $x^* \in]x_{\min}, x_{\max}[$ be such that $u^* := (x^*, f(x^*))$ maximizes the hypervolume with respect to a valid reference point r. If the Pareto front is described by a differentiable function f in x^*, then $f'(x^*)$ satisfies*

$$-f'(x^*) = \frac{r_2 - f(x^*)}{r_1 - x^*} \ . \tag{4.15}$$

Proof *It is a direct consequence of Proposition 7*

A convex function may not be differentiable, but it always has left and right derivatives. It is also above its left and right tangent lines respectively on the left and on the right of x^*. Therefore, Proposition 7 implies that the affine function $g : x \mapsto f(x^*) - \frac{r_2 - f(x^*)}{r_1 - x^*} \times (x - x^*)$ is a minorant of f. This is the key idea of the proof of the following lower bound on the normalized maximum hypervolume.

Proposition 8 *If the Pareto front is described by a convex function f, then the following lower bound on the normalized maximum hypervolume with respect to any valid reference point r holds:*

$$\frac{\max_{u\in PF_f} HV_r(u)}{HV_r(PF_f)} \geq \frac{1}{2} \tag{4.16}$$

where the inequality is an equality if and only if the Pareto front is affine and r dominates the nadir point.

Proof *As explained in the above paragraph, the convexity of f implies that the affine function $g : x \mapsto f(x^*) - \frac{r_2-f(x^*)}{r_1-x^*} \times (x - x^*)$ is a minorant of f. Therefore, $PF_g := \{g(x) : x \in [x_{\min}, x_{\max}]\}$ dominates PF_f, and thus has a higher hypervolume. We denote $L_1 := r_1 - \tilde{x}_{\min,r}$ and $L_2 := r_2 - f(\tilde{x}_{\max,r})$ the lengths of the rectangle $\mathcal{R} := [\tilde{x}_{\min,r}, r_1] \times [f(\tilde{x}_{\max,r}), r_2]$. We denote $l_1 := r_1 - x^*$ and $l_2 := r_2 - f(x^*)$ the lengths of the rectangle $\mathcal{D}^r_{u^*}$. The region of $\mathcal{R}$ which dominates PF_g is a right-angled triangle. Additionally, by definition, the slope of its hypotenuse is l_2/l_1, and thus the lengths of the other sides are $L_1 - l_1 + (L_2 - l_2) \times \frac{l_1}{l_2}$ and $L_2 - l_2 + (L_1 - l_1) \times \frac{l_2}{l_1}$ (see the middle plot of Figure 4.6). Therefore, we have*

$$\begin{aligned} HV_r(PF_g) &= \lambda(\mathcal{R}) - \lambda(\{u \in \mathbb{R}^2 : u \in \mathcal{R}, u \preceq PF_g\}) \\ &= L_1L_2 - \frac{1}{2} \times (L_1 - l_1 + (L_2 - l_2) \times \frac{l_1}{l_2}) \times (L_2 - l_2 + (L_1 - l_1) \times \frac{l_2}{l_1}) \\ &= l_1l_2 \times \Big[-2 + 2 \times \frac{L_2}{l_2} - \frac{1}{2} \times \Big(\frac{L_2}{l_2}\Big)^2 + 2 \times \frac{L_1}{l_1} - \frac{1}{2} \times \Big(\frac{L_1}{l_1}\Big)^2 \Big] \ . \end{aligned}$$

For all x, we have $(x-2)^2 \geq 0$ and thus $2x - \frac{1}{2}x^2 \leq 2$. Therefore, we can conclude that $HV_r(PF_g)$, and thus $HV_r(PF_f)$ is smaller than $2 \times l_1l_2$, that is $2 \times HV_r(u^*)$. If either $L_1/l_1 \neq 2$ or $L_2/l_2 \neq 2$, the inequality is strict. Thus, when the inequality is an equality, the center of $\mathcal{R}$ belongs to the Pareto front. Since f is convex, it happens only when f is affine and the reference point r dominates the nadir point. Conversely, if both conditions are met, we know that the optimum is in the middle of the Pareto front and that we have the equality (see [8, Theorem 5]).

We just proved that one half is a tight lower bound on the normalized maximum hypervolume for convex Pareto fronts. However, except for the trivial upper bound 1, there is no upper bound valid for every convex Pareto front, even when r dominates the nadir point. Let consider the Pareto fronts $\mathrm{PF}_\epsilon := \{\max(1 - \frac{x}{\epsilon}, \epsilon - \epsilon \times x) : x \in [0, 1]\}$ (represented in the righthand plot of Figure 4.6 for $\epsilon = \frac{1}{3}$). The normalized maximum hypervolume of PF_ϵ for the reference point $r = (1, 1)$ converges to 1 when ϵ goes to 0.[2]

4.4.2 Lower and upper bounds on the normalized maximum hypervolume for bilipschitz Pareto fronts

In this section, we examine lower and upper bounds on the normalized maximum hypervolume in the case of bilipschitz Pareto fronts.

[2]The normalized hypervolume equals $\frac{(1-\epsilon+\epsilon^2)^2}{1-\epsilon\times(1-\epsilon)^2+(\epsilon-\epsilon^2)^2}$ which converges to 0 when ϵ goes to 0.

We consider two affine fronts with the same left extreme vector as PF_f and slopes $-L_{\min}$ and $-L_{\max}$, see the lefthand plot of Figure 4.7. We call them $\mathrm{PF}_{\min}$ and $\mathrm{PF}_{\max}$, respectively. Formally:

$$\mathrm{PF}_{\max} := \{(x, f_{\max}(x)) : x \in [x_{\min}, x_{\max}]\} \text{ and} \tag{4.17}$$
$$\mathrm{PF}_{\min} := \{(x, f_{\min}(x) : x \in [x_{\min}, x_{\max}]\} \tag{4.18}$$

with $f_{\min}(x) = f(x_{\min}) - (x - x_{\min}) \times L_{\min}$ and $f_{\max}(x) = f(x_{\min}) - (x - x_{\min}) \times L_{\max}$. For a ($L_{\min}$,$L_{\max}$)-bilipschitz function f, $f_{\max}(x) \leq f(x) \leq f_{\min}(x)$ for all $x \in [x_{\min}, x_{\max}]$, and thus the Pareto front is dominated by $\mathrm{PF}_{\max}$ and dominates $\mathrm{PF}_{\min}$. These two

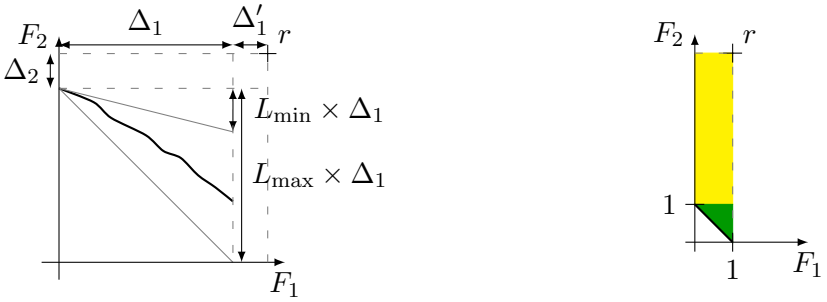


Figure 4.7: Left : The Pareto front PF_f surrounded by $\mathrm{PF}_{\max}$ (below) and $\mathrm{PF}_{\min}$ (above). Right : An illustration that $\mathrm{HV}_r(u_{\min}) - \mathrm{HV}_r(\mathrm{PF}_f)$ (▬) becomes negligible compared to $\mathrm{HV}_r(u_{\min})$ (▬) for $r_1 = 1$ and $r_2 \to \infty$.

affine fronts provide bounds on both the hypervolume of the Pareto front and the largest hypervolume of a vector of the Pareto front. They are key to prove the following lower bound on the normalized maximum hypervolume of a ($L_{\min}$,$L_{\max}$)-bilipschitz Pareto front.

Proposition 9 *If the Pareto front is described by a ($L_{\min}$,$L_{\max}$)-bilipschitz function f, then for any valid reference point r, we have*

$$\frac{\max_{u \in PF_f} HV_r(u)}{HV_r(PF_f)} \geq \frac{1}{2} \times \frac{L_{\min}}{L_{\max}} \ . \tag{4.19}$$

Proof *The fronts $PF_{\max}$ and $PF_{\min}$ are defined respectively in* (4.17) *and* (4.18)*. We note $\Delta_1 := \tilde{x}_{\max,r} - \tilde{x}_{\min,r}$, $\Delta'_1 := r_1 - \tilde{x}_{\max,r}$, $\Delta_2 := r_2 - f(\tilde{x}_{\min,r})$ and $V := \Delta_2 \times (r_1 - \tilde{x}_{\min,r})$, see the lefthand plot of Figure 4.7. Since the front $PF_{\max}$ dominates the Pareto front, the hypervolume of PF_f is smaller than the hypervolume of $PF_{\max}$, $V + L_{\max} \times \Delta_1 \times \Delta'_1 + \frac{1}{2} \times L_{\max} \times \Delta_1^2$. Additionally, since each vector of $PF_{\min}$ is dominated by a vector of PF_f, the maximum hypervolume of a vector of PF_f is larger than the maximum hypervolume of a vector of $PF_{\min}$. The front $PF_{\min}$ being an affine and therefore convex front, we know by Proposition 8 that the maximum hypervolume of a vector of $PF_{\min}$ is larger than half of $HV_r(PF_{\min})$, which is equal to*

$\frac{1}{2} \times (V + L_{\min} \times \Delta_1 \times \Delta'_1 + \frac{1}{2} \times L_{\min} \times \Delta_1^2)$. *To summarize, the maximum hypervolume of a vector of* PF_f *is larger than* $\frac{1}{2} \times (V + L_{\min} \times \Delta_1 \times \Delta'_1 + \frac{1}{2} \times L_{\min} \times \Delta_1^2)$. *Combining the upper bound on the hypervolume of* PF_f *and the lower bound on the maximum hypervolume of a vector of* PF_f*, the normalized maximum hypervolume is larger than* $\frac{\frac{1}{2}\times(V+L_{\min}\times\Delta_1\times\Delta'_1+\frac{1}{2}\times L_{\min}\times\Delta_1^2)}{V+L_{\max}\times\Delta_1\times\Delta'_1+\frac{1}{2}\times L_{\max}\times\Delta_1^2}$. *This quantity is itself larger than* $\frac{1}{2} \times \frac{L_{\min}\times\Delta_1\times\Delta'_1+\frac{1}{2}\times L_{\min}\times\Delta_1^2}{L_{\max}\times\Delta_1\times\Delta'_1+\frac{1}{2}\times L_{\max}\times\Delta_1^2}$. *As* $V \geq 0$ *and* $0 < \frac{L_{\min}\times\Delta_1\times\Delta'_1+\frac{1}{2}\times L_{\min}\times\Delta_1^2}{L_{\max}\times\Delta_1\times\Delta'_1+\frac{1}{2}\times L_{\max}\times\Delta_1^2} < 1$, *we conclude that the normalized maximum hypervolume is larger than* $\frac{1}{2} \times \frac{L_{\min}}{L_{\max}}$.

We cannot guarantee any upper bound strictly smaller than 1 on the normalized maximum hypervolume without adding an assumption on the reference point. Indeed, for a given bounded Pareto front, it is easy to show that the normalized maximum hypervolume goes to 1 for $r_1 = x_{\max}$ and $r_2 \to \infty$ (see the righthand plot of Figure 4.7). However, if f is ($L_{\min}$,$L_{\max}$)-bilipschitz and r dominates the nadir point, we can prove that the normalized maximum hypervolume is larger than $\frac{1}{2} \times \frac{L_{\max}}{L_{\min}}$. The proof relies on the fact that if the reference point r dominates the nadir point, the vector of an affine front with the largest hypervolume relative to r is its middle (see [8, Theorem 5]), whose hypervolume is half of the hypervolume of the entire Pareto front.

Proposition 10 *If the Pareto front is described by a* ($L_{\min}$,$L_{\max}$)*-bilipschitz function* f *and the reference point* r *is valid and dominates the nadir point, the following upper-bound on the normalized maximum hypervolume with respect to* r *holds*

$$\frac{\max_{u \in PF_f} HV_r(u)}{HV_r(PF_f)} \leq \frac{1}{2} \times \frac{L_{\max}}{L_{\min}} \quad . \tag{4.20}$$

Proof *We use the same notations as in the proof of Proposition 9. Since* r *dominates the nadir point, both* Δ'_1, Δ_2 *and* V *equal 0, and thus the hypervolumes of* $PF_{\max}$ *and* $PF_{\min}$ *equal* $\frac{1}{2} \times L_{\max} \times \Delta_1^2$ *and* $\frac{1}{2} \times L_{\min} \times \Delta_1^2$*, respectively. The domination of* $PF_{\min}$ *by* PF_f *implies that the hypervolume of the Pareto front is larger than* $\frac{1}{2} \times L_{\min} \times \Delta_1^2$. *Since* $PF_{\max}$ *is an affine front whose extremes dominate* r*, its middle is the unique hypervolume maximizer (see [6, Theorem 5]) with an hypervolume equal to* $\frac{1}{4} \times L_{\max} \times \Delta_1$.*The domination of* PF_f *by* $PF_{\max}$ *implies that the maximum hypervolume of a vector of* PF_f *is smaller than* $\frac{1}{4} \times L_{\max} \times \Delta_1^2$. *Gathering the lower bound on* $HV_r(PF_f)$ *and the upper bound on the maximum hypervolume of a vector of* PF_f*, we retrieve* (4.20).

This upper bound is only relevant for $L_{\max}/L_{\min} < 2$ and is the tightest for $L_{\max} = L_{\min}$, where it achieves the value 1/2. In this chapter, we use this upper bound for $L_{\max}/L_{\min}$ close to 1 to analyze the asymptotic convergence behavior of HV-ISOOMOO.

4.5 Convergence of HV-ISOOMOO coupled with perfect single-objective optimization

We prove in this section various convergence results for HV-ISOOMOO algorithms coupled with perfect single-objective optimization. We first prove that when the Pareto front

is either convex or bilipschitz, these algorithms converge to the entire Pareto front. We transform the bounds on the normalized maximum hypervolume proven in Section 4.4 into lower bounds on the speed of convergence. Then, we analyze the asymptotic convergence behavior when the Pareto front is bilipschitz with a Hölder continuous derivative.

To analyze the decrease of the optimality gap with n, we track in which gap regions the vectors of the greedy sequence are inserted over multiple iterations. Naturally, a gap region of $\mathcal{S}_n$ persists in being a gap region as long as no greedy vector is added in this specific gap region. The greedy vector v_{n+1} is said to *fill* the gap region of $\mathcal{S}_n$ to which it belongs. At iteration $n+1$, this gap region disappears, replaced by two gap regions that we call its *children*. More generally, we say that a gap region is a *descendant* of another gap region when it is a proper subset of this gap region.

4.5.1 Convergence of HV-ISOOMOO with guaranteed speed of convergence

We prove some upper bounds on the relation between the optimality gap at iteration $2n+1$ and at iteration n. These bounds translate into lower bounds on the speed of convergence of HV-ISOOMOO under Assumption 2 of perfect single-objective optimization. The proof relies on inequalities of the form

$$\max_{u \in \mathrm{PF}_f} \mathrm{HV}_{r'}(u) \geq C \times \mathrm{HV}_{r'}(\mathrm{PF}_f) \tag{4.21}$$

stated in Propositions 8 and 9 and equations regarding optimality gaps, areas of gap regions and hypervolume improvement presented in Section 4.2.2. A consequence of (4.21) being true for any reference point r' is that the optimality gap at iteration $2n+1$ is at most $(1-C)$ times the optimality gap at iteration n.

We sketch the proof idea in the simple case where each of the v_k ($k \in [\![n+1, 2n+1]\!]$) is inserted in a distinct gap region of $\mathcal{S}_n$, see the lefthand plot of Figure 4.8. Inserting v_k in a gap region leads to an hypervolume improvement larger than C times the area of this gap region by (4.21). Thus, the hypervolume improvement from iteration n to $2n+1$ is larger than C times the area of the union of all gap regions of $\mathcal{S}_n$, namely the optimality gap at iteration n. A detailed proof is presented after the theorem statement.

Proposition 11 *Consider a bi-objective optimization problem with a Pareto front described by a function f. Any greedy set sequence $(\mathcal{S}_n)_{n \in \mathbb{N}^*}$ relative to a valid reference point r satisfies for all n*

$$\frac{HV_r(PF_f) - HV_r(\mathcal{S}_{2n+1})}{HV_r(PF_f) - HV_r(\mathcal{S}_n)} \leq 1 - \frac{1}{2} \times \frac{L_{\min}}{L_{\max}} \text{ if } f \text{ is } (L_{\min}, L_{\max})\text{-bilipschitz and} \tag{4.22}$$

$$\frac{HV_r(PF_f) - HV_r(\mathcal{S}_{2n+1})}{HV_r(PF_f) - HV_r(\mathcal{S}_n)} \leq \frac{1}{2} \text{ if } f \text{ is convex.} \tag{4.23}$$

Proof *Fix $n \geq 1$. We note σ a permutation of $[\![1, n+1]\!]$ such that $n + \sigma(i)$ is the index of the first greedy vector v_k inserted in $\mathcal{G}^r_{\mathcal{S}_n, i}$ when possible. With this choice of σ, the*

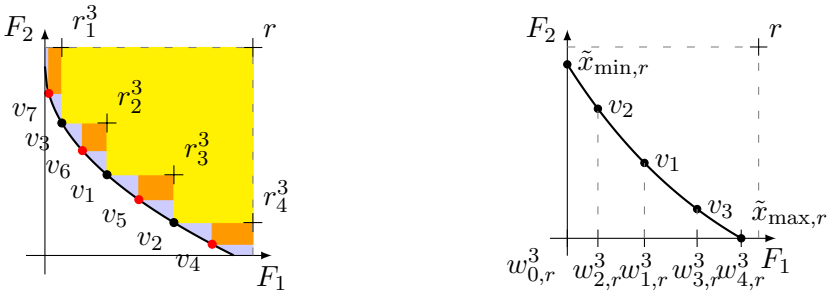


Figure 4.8: Left: A Pareto front where each of the gap regions of $\mathcal{S}_3$ is filled by one of the greedy vectors v_k for $k \in [\![4,7]\!]$. The front is described by $f(x) = 1 - \sqrt{x}$ for $x \in [0,1]$. We represent the region $\mathcal{D}^r_{\mathcal{S}_3}$ (▭) , the gap regions of $\mathcal{S}_3$ (▭) and the regions corresponding to $\text{HVI}_r(v_k, \mathcal{S}_{k-1})$ for $k \in [\![4,7]\!]$ (▭). Right: The ordered greedy set F_1-values $w^n_{i,r}$ corresponding to the greedy set $\mathcal{S}_3$. The Pareto front is described by $f(x) = \frac{e}{e-1} \times e^{-x} + 1 - \frac{e}{e-1}$ for $x \in [0,1]$.

i-th gap region of $\mathcal{S}_n$ is a gap region of $\mathcal{S}_{n+\sigma(i)-1}$. As a consequence, the hypervolume improvement to $\mathcal{S}_{n+\sigma(i)-1}$ of any vector u belonging to the i-th gap region of $\mathcal{S}_n$, $\mathcal{G}^r_{\mathcal{S}_n,i}$, is equal to $HV_{r^n_i}(u)$ by Lemma 4. The hypervolume improvement of the greedy vector $v_{n+\sigma(i)}$ to $\mathcal{S}_{n+\sigma(i)-1}$ being maximal, it is in particular larger than the one of any vector of $\mathcal{G}^r_{\mathcal{S}_n,i}$ and thus than $\frac{1}{2} \times \frac{L_{\min}}{L_{\max}} \times HV_{r^n_i}(PF_f)$ by Proposition 9. In other words, the hypervolume improvement at any iteration $n+\sigma(i)$ is larger than $\frac{1}{2} \times \frac{L_{\min}}{L_{\max}} \times HV_{r^n_i}(PF_f)$. By adding these inequations for all $i \in [\![1, n+1]\!]$, we deduce that the hypervolume improvement from iteration n to $2n+1$ is larger than $\frac{1}{2} \times \frac{L_{\min}}{L_{\max}} \times \sum_{i=1}^{n+1} HV_{r^n_i}(PF_f)$. Since the sum of the $HV_{r^n_i}(PF_f)$ is the optimality gap at iteration n, we have (4.22)*. If f is convex instead of bilipschitz, we use Proposition 8 instead of Proposition 9 and obtain* (4.23)*.*

Since the optimality gaps form a decreasing sequence, such lower bounds on the relation between the optimality gaps at iteration $2n+1$ and at iteration n imply that the optimality gap associated to a greedy set sequence converges asymptotically to 0. Equivalently, sets constructed by HV-ISOOMOO algorithms coupled with perfect single-objective optimization converge to the entire Pareto front as stated formally below.

Theorem 5 *Consider a bi-objective optimization problem with a Pareto front described by a bilipschitz or convex function f.*

The hypervolume of a greedy set sequence relative to a valid reference point r converges to the hypervolume of the entire Pareto front, i.e. $HV_r(\mathcal{S}_n) \xrightarrow[n\to\infty]{} HV_r(PF_f)$.

Equivalently, for such Pareto fronts and under Assumption 2 of perfect single-objective optimization, HV-ISOOMOO algorithms relative to a valid reference point r converge in the sense of Definition 4.

From the lower bounds on the relation between the optimality gaps at iteration $2n+1$ and at iteration n, we deduce the following upper bounds on the normalized optimality gap at any iteration.

Corollary 3 *Consider a bi-objective optimization problem with a Pareto front described by a ($L_{\min}$,$L_{\max}$)-bilipschitz function. A greedy set sequence $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ relative to a valid reference point r satisfies for all n*

$$\frac{HV_r(PF_f) - HV_r(\mathcal{S}_n)}{HV_r(PF_f)} \leq \left(1 - \frac{1}{2} \times \frac{L_{\min}}{L_{\max}}\right)^{\lfloor \log_2(n+1) \rfloor} \leq (2n+2)^{\log_2(1-\frac{1}{2}\times\frac{L_{\min}}{L_{\max}})} \ . \quad (4.24)$$

If the function f is convex, then any greedy set sequence relative to a valid reference point r satisfies for all n

$$\frac{HV_r(PF_f) - HV_r(\mathcal{S}_n)}{HV_r(PF_f)} \leq \left(\frac{1}{2}\right)^{\lfloor \log_2(n+1) \rfloor} \leq \frac{1}{2n+2} \ . \quad (4.25)$$

Hence, for such reference points and under Assumption 2 of perfect single-objective optimization, HV-ISOOMOO algorithms relative to r satisfy (4.24) if f is ($L_{\min}$,$L_{\max}$)-bilipschitz and (4.25) if f is convex where $\mathcal{S}_n$ is replaced by $\mathcal{I}_n$, the final incumbents Pareto front approximation at iteration n.

Proof *The k-th term of the sequence defined by $u_0 = 1$ and $u_{n+1} = 2 \times u_n + 1$ for all $n \geq 1$ is $2^k - 1$. Thus, (4.22) and (4.23) imply that when f is ($L_{\min}$,$L_{\max}$)-bilipschitz or convex, the normalized optimality gap at iteration $2^k - 1$ is smaller than $(1-C)^k$ with C equal to $\frac{1}{2} \times \frac{L_{\min}}{L_{\max}}$ and $\frac{1}{2}$, respectively. Since the hypervolume of the greedy set increases with n, and thus the optimality gap decreases with n, we deduce the first inequalities in (4.24) and (4.25) via the change of variable $k = \lfloor \log_2(n+1) \rfloor$.*

Additionally, for every n, $\lfloor \log_2(n+1) \rfloor$ is smaller than $\log_2(n+1) + 1$, that is $\log_2(2n+2)$. For every C, $\log_2(2n+2)$ equals $\log_C(2n+2) \times \log_2(C)$, and thus $C^{\log_2(2n+2)}$ equals $(2n+2)^{\log_2(C)}$. Therefore, we can infer that $(2n+2)^{\log_2(C)}$ is an upper bound of the normalized optimality gap with $C = 1 - \frac{1}{2} \times \frac{L_{\min}}{L_{\max}}$ and $C = \frac{1}{2}$ when f is ($L_{\min}$,$L_{\max}$)-bilipschitz and convex, respectively.

We focused on the relation between the optimality gap at iteration n and at iteration $2n+1$. We can similarly examine the relation between the optimality gap at iteration n and at any later iteration.

Lemma 7 *If f is ($L_{\min}$,$L_{\max}$)-bilipschitz (resp. convex), then for all n, for all $k \leq n+1$, $\frac{HV_r(PF_f) - HV_r(\mathcal{S}_{n+k})}{HV_r(PF_f) - HV_r(\mathcal{S}_n)}$ is smaller than $1-(1/2)(L_{\min}/L_{\max})\times k/(n+1)$ (resp. $1-(1/2)k/(n+1)$).*

Proof *Consider the k gap regions of $\mathcal{S}_n$ with the largest areas. If f is ($L_{\min}$,$L_{\max}$)-bilipschitz (resp. convex), the hypervolume improvement from iteration n to $n+k$ is at least $1/2 \times L_{\min}/L_{\max}$ (resp. $1/2$) times the area of the union of these gap regions, which is at least $\frac{k}{n+1}$ times the optimality gap at iteration n.*

Yet, the previous lemma leads to looser lower bounds on the convergence rate. To illustrate this, we detail the case $k = 1$ in the following remark.

Remark 1 *For all $m \in \mathbb{N}^*$, a direct consequence of Lemma 7 applied with $k = 1$ and $n \in [\![m, 2m]\!]$ is that*

$$\frac{HV_r(PF_f) - HV_r(\mathcal{S}_{2m+1})}{HV_r(PF_f) - HV_r(\mathcal{S}_m)} \leq \prod_{i=m+1}^{2m+1} (1 - \frac{C}{i})$$

with $C = 1/2$ and $C = (1/2)L_{\min}/L_{\max}$ for convex and bilipschitz Pareto fronts, respectively. These upper bounds are larger than the ones provided in Proposition 11: $1 - C$ with the same constants C.

Proof *Each term of the product $\prod_{i=m+1}^{2m+1}(1-\frac{C}{i})$ is larger than $1-C/(m+1)$. Thus, this product is larger than $(1-C/(m+1))^{m+1}$. The sequence $((1-C/i)^i)_{i\in\mathbb{N}^*}$ is an increasing sequence. This is a classic result that can be proven by using the inequality between the geometric and the arithmetic means of the vector $(1, 1-C/i, \ldots, 1-C/i) \in \mathbb{R}^{i+1}$. As a consequence, $(1 - C/(m + 1))^{m+1}$ is larger than $1 - C$, the first term of the sequence.*

4.5.2 Asymptotical behavior of the convergence of $\mathbf{HV}_r(\mathcal{S}_n)$ to $\mathbf{HV}_r(\mathbf{PF}_f)$

In this section, we analyze the asymptotic convergence behavior for a Pareto front described by a bilipschitz function with a Hölder continuous derivative. We prove that, in this case, doubling the number of vectors in the greedy set divides the optimality gap by a factor which converges asymptotically to two as stated in Theorem 6. This asymptotic limit corresponds to the case of affine Pareto fronts with a reference point dominating the nadir point. For such Pareto fronts and reference points, the optimality gap is always halved when the number of vectors in the greedy set goes from n to $2n+1$, see Figure 4.9.

First, we study the properties of the part of the Pareto front corresponding to a specific gap region of $\mathcal{S}_n$. We introduced the *ordered greedy set F_1-values* in (4.11). Naturally, we have $w^n_{0,r} \leq w^n_{1,r} \leq \ldots \leq w^n_{n+1,r}$, and the intervals $[w^n_{i-1,r}, w^n_{i,r}[$ for $i \in [\![1, n+1]\!]$ form a partition of $[\tilde{x}_{\min,r}, \tilde{x}_{\max,r}[$, see the righthand plot of Figure 4.8. The interval $[w^n_{i-1,r}, w^n_{i,r}]$ corresponds to the part of the Pareto front belonging to the i-th gap region of $\mathcal{S}_n$. When the Pareto front is bilipschitz, the lengths of these intervals converge asymptotically to 0 as stated in the next lemma. This result is a direct consequence of the convergence of $\mathrm{HV}_r(\mathcal{S}_n)$ to $\mathrm{HV}_r(\mathrm{PF}_f)$ stated in Theorem 5.

Lemma 8 *If the Pareto front is described by a bilipschitz function f and the greedy set sequence is associated to a valid reference point r, then the ordered greedy set F_1-values satisfy $\max_{i\in[\![1,n+1]\!]}(w^n_{i,r} - w^n_{i-1,r}) \xrightarrow[n\to\infty]{} 0$ with the $w^n_{i,r}$ defined in (4.11).*

Proof *Let $L_{\min}$ and $L_{\max}$ be constants such that f is ($L_{\min}$,$L_{\max}$)-bilipschitz. The area of the i-th gap region of $\mathcal{S}_n$ is $\int_{w^n_{i-1,r}}^{w^n_{i,r}} (f(x) - f(w^n_{i,r}))dx$. This is larger than $\int_{w^n_{i-1,r}}^{w^n_{i,r}} L_{\min} \times (w^n_{i,r} - x)dx$, which equals $\frac{1}{2} \times L_{\min} \times (w^n_{i,r} - w^n_{i-1,r})^2$. Since the area of any gap region of $\mathcal{S}_n$ is smaller than the optimality gap at iteration n, this implies that the difference $w^n_{i,r} - w^n_{i-1,r}$ is smaller than $\sqrt{(2/L_{\min})(HV_r(PF_f) - HV_r(\mathcal{S}_n))}$ for all n, for all $i \in$*

$[\![1, n+1]\!]$. *Therefore, the convergence of* $HV_r(\mathcal{S}_n)$ *to* $HV_r(PF_f)$ *stated in Theorem 5 implies that the maximum over* i *of* $w^n_{i,r} - w^n_{i-1,r}$ *converges to 0.*

We prove in the next lemma that if the Pareto front is described by a bilipschitz function f with a Hölder continuous derivative, then the part of the Pareto front belonging to a specific gap region of $\mathcal{S}_n$ is bilipschitz for some constants whose ratio converges asymptotically to 1. Affine functions being the only functions to be ($L_{\min}$,$L_{\max}$)-bilipschitz with $L_{\min}/L_{\max} = 1$, it supports the interpretation that the convergence of a greedy set sequence for such Pareto fronts and for affine Pareto fronts share some asymptotic similarities.

When f is bilipschitz, its restriction to the part of the Pareto front dominating r^n_i, that is $[w^n_{i-1,r}, w^n_{i,r}]$, is ($L^{i,n}_{\min}$,$L^{i,n}_{\max}$)-bilipschitz with

$$\begin{aligned} L^{i,n}_{\min} &:= \inf\left\{\left|\frac{f(x)-f(y)}{x-y}\right|, x, y \in [w^n_{i-1,r}, w^n_{i,r}], x \neq y\right\} \text{ and} \\ L^{i,n}_{\max} &:= \sup\left\{\left|\frac{f(x)-f(y)}{x-y}\right|, x, y \in [w^n_{i-1,r}, w^n_{i,r}], x \neq y\right\}. \end{aligned} \tag{4.26}$$

At iteration n, the ratio between $L^{i,n}_{\max}$ and $L^{i,n}_{\min}$, the bilipschitz constants on the i-th gap region of $\mathcal{S}_n$, is by definition smaller than

$$q_n := \max\left\{\frac{L^{i,n}_{\max}}{L^{i,n}_{\min}}, i \in [\![1, n+1]\!] : [w^n_{i-1,r}, w^n_{i,r}] \neq \emptyset\right\} \quad . \tag{4.27}$$

The proof of the convergence of q_n to 1 relies on the fact that a differentiable function can be approximated locally by an affine function. The quality of this approximation is guaranteed by the Hölder continuity of the derivative.

Lemma 9 *We consider a greedy set sequence* $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ *relative to a valid reference point* r. *If the Pareto front is described by a bilipschitz function with a Hölder continuous derivative, then* q_n *defined in* (4.27) *converges asymptotically to 1.*

Proof *We take* α *such that* f' *is Hölder continuous with exponent* α, *i.e* f *is* $\mathcal{C}^{1,\alpha}$, *and* $L_{\min}, L_{\max} > 0$ *such that the function* f *describing the Pareto front is* ($L_{\min}$,$L_{\max}$)*-bilipschitz. We recall that* f *is decreasing, and thus for all* $x < y$, *we have* $f(x)-f(y) \geq 0$. *Since* f *is* $\mathcal{C}^{1,\alpha}$ *and therefore* $\mathcal{C}^1$, *the Taylor formula with Lagrange remainder states that for all* $x < y$, *there exists* $\xi \in [x, y]$ *such that* $f(y) = f(x) + (y-x) \times f'(\xi)$. *Since* f *is* $\mathcal{C}^{1,\alpha}$, *this implies that for all* $x < y$, $|f(y) - f(x) - (y-x) \times f'(x)| \leq (y-x)^{1+\alpha} \times [f']_{\mathcal{C}^\alpha}$. *Thus,* $\frac{f(y)-f(x)}{x-y}$ *is smaller than* $-f'(x) + [f']_{\mathcal{C}^\alpha} \times (y-x)^\alpha$. *We now restrict ourselves to* x *and* y *belonging to the non-empty interval* $[w^n_{i-1,r}, w^n_{i,r}]$. *Our goal is to find an upper bound depending on* i *but not on either* x *or* y. *Since* f *is* $\mathcal{C}^{1,\alpha}$, *the difference between* $-f'(x)$ *and* $-f'(w^n_{i-1,r})$ *is smaller than* $[f']_{\mathcal{C}^\alpha} \times (x - w^n_{i-1,1,r})^\alpha$, *and thus* $[f']_{\mathcal{C}^\alpha} \times (w^n_{i,r} - w^n_{i-1,r})^\alpha$. *Additionally, the difference between* x *and* y *is smaller than* $w^n_{i,r} - w^n_{i-1,r}$. *We conclude that for* $x, y \in [w^n_{i-1,r}, w^n_{i,r}]$, $\frac{f(y)-f(x)}{x-y}$ *is smaller than* $-f'(w^n_{i-1,1,r}) + 2[f']_{\mathcal{C}^\alpha} \times (w^n_{i,r} - w^n_{i-1,r})^\alpha$, *and thus so is* $L^{i,n}_{\max}$ *defined in* (4.26).

Following the same approach, we can also infer that $L^{i,n}_{\max}$ *defined in* (4.26) *is greater than the symmetric quantity* $-f'(w^n_{i-1,1,r}) - 2[f']_{\mathcal{C}^\alpha} \times (w^n_{i,r} - w^n_{i-1,r})^\alpha$. *The quantity* $-f'(w^n_{i-1,1,r})$ *is greater than* $L_{\min}$ *and* $(w^n_{i,r} - w^n_{i-1,r})^\alpha$ *is smaller than* $\max_{i\in[\![1,n+1]\!]}(w^n_{i,r} - w^n_{i-1,r})^\alpha$. *As a consequence,* q_n *is smaller than* $\frac{L_{\min}+2[f']_{\mathcal{C}^\alpha}\times\max_{i\in[\![1,n+1]\!]}(w^n_{i,r}-w^n_{i-1,r})^\alpha}{L_{\min}-2[f']_{\mathcal{C}^\alpha}\times\max_{i\in[\![1,n+1]\!]}(w^n_{i,r}-w^n_{i-1,r})^\alpha}$. *By Lemma 8,* $\max_{i\in[\![1,n+1]\!]}(w^n_{i,r} - w^n_{i-1,r})$ *converges to 0 and thus, this upper bound on* q_n *converges to 1. Since* q_n *is always larger than 1, it converges to 1.*

A consequence of the previous lemma is that the bounds on the hypervolume improvement of v_{n+1} to $\mathcal{S}_n$ normalized by the area of the gap region filled by v_{n+1} that we can infer from Propositions 9 and 10 converge asymptotically to 1/2, see (4.28). Similarly, the bounds on the normalized area of the child of a gap region that we can infer from Lemma 12 converge to 1/4, see (4.29). These asymptotic values correspond to the case of an affine Pareto front with a reference point dominating the nadir point, see Figure 4.9.

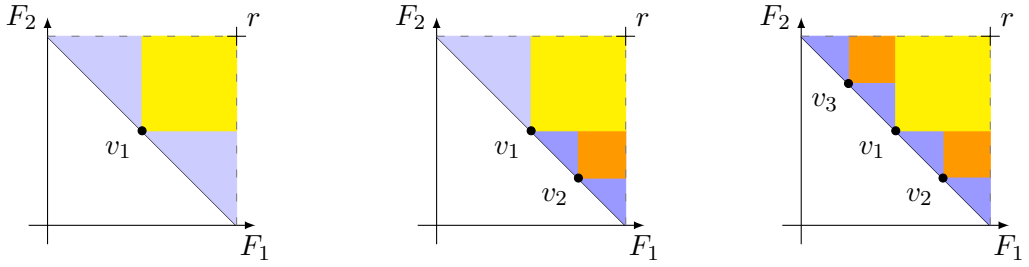


Figure 4.9: The three greedy sets $\mathcal{S}_1$ (left), $\mathcal{S}_2$ (middle) and $\mathcal{S}_3$ (right) and their gap regions for an affine Pareto front with a reference point r dominating the nadir point. The area of any of the gap regions of $\mathcal{S}_1$ are half of $\mathrm{HV}_r(\mathrm{PF}_f)$ (left). The area of any of the new gap regions of $\mathcal{S}_2$ is a quarter of the area of their parents (middle). The optimality gap of $\mathcal{S}_3$ (right) is half of the optimality gap of $\mathcal{S}_1$.

Lemma 10 *We consider a greedy set sequence* $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ *relative to a valid reference point* r. *If the Pareto front is described by a bilipschitz function* f *with a Hölder continuous derivative, then for all* $\epsilon > 0$, *for* n *large enough, for every non-empty gap region* $\mathcal{G}^r_{\mathcal{S}_n,i}$ *and every child* $\mathcal{G}^r_{\mathcal{S}_m,j}$ *of* $\mathcal{G}^r_{\mathcal{S}_n,i}$, *we have*

$$\frac{1}{2}\times(1-\epsilon) \leq \frac{\max_{u\in\mathcal{G}^r_{\mathcal{S}_n,i}} HVI_r(u,\mathcal{S}_n)}{\lambda(\mathcal{G}^r_{\mathcal{S}_n,i})} \leq \frac{1}{2}\times(1+\epsilon) \text{ and} \tag{4.28}$$

$$\frac{1}{4\times(1+\epsilon)} \leq \frac{\lambda(\mathcal{G}^r_{\mathcal{S}_m,j})}{\lambda(\mathcal{G}^r_{\mathcal{S}_n,i})} \leq \frac{1}{4\times(1-\epsilon)} \ . \tag{4.29}$$

Proof *The interval* $[w^n_{i-1,r}, w^n_{i,r}]$ *is the set of the first coordinates of the vectors of the Pareto front which dominate* r^n_i . *The restriction to* $[w^n_{i-1,r}, w^n_{i,r}]$ *of* f *is* ($L_{\min}$,$L_{\max}$)*-bilipschitz for some* $L_{\min}$ *and* $L_{\max}$ *such that* $\frac{L_{\max}}{L_{\min}} = q_n$ *with* q_n *defined in* (4.27).

Additionally, as stated in Proposition 15, for n large enough, all the r_i^n corresponding to non-empty gap regions dominate the nadir point. As a consequence, the conditions to apply Lemma 12 and Proposition 10 are met by non-extremes gap regions.

By Propositions 9 and 10, $\max_{u \in PF_f} HV_{r_i^n}(u)/HV_{r_i^n}(PF_f)$ is between $(1/2)1/q_n$ and $(1/2)q_n$. Additionally, by Lemma 12, $\lambda(\mathcal{G}^r_{\mathcal{S}_m,j})/HV_{r_i^n}(PF_f)$ is between $(1-(1/2)q_n)/(1+q_n^2)$ and $(1-(1/2)1/q_n)/(1+1/q_n^2)$. The maximum over the vectors u belonging to the Pareto front of $HV_{r_i^n}(u)$ is equal to the maximum over u belonging to the i-th gap region of $\mathcal{S}_n$ of $HVI_r(u, \mathcal{S}_n)$. Indeed, $HV_{r_i^n}(.)$ is null for vectors outside the i-th gap region of $\mathcal{S}_n$ while it is nonnegative, equal to $HVI_r(., \mathcal{S}_n)$, otherwise. Additionally, $HV_{r_i^n}(PF_f)$ equals $\lambda(\mathcal{G}^r_{\mathcal{S}_n,i})$. The convergence of q_n to 1 stated in Lemma 9 imply that the bounds proven so far converge to a half and a quarter, respectively. Thus, we have (4.28) *and* (4.29) *for n large enough.*

The following lemma states that for n large enough, the area of two non-empty gap regions relative to the same greedy set cannot be too different. More precisely, the area of any gap region of $\mathcal{S}_n$ cannot be more than $4 \times (1 + o(\epsilon))$ times greater than the area of another gap region of $\mathcal{S}_n$. The proof relies on considering the parents of the gap regions.

Lemma 11 *We consider a greedy set sequence $(\mathcal{S}_n)_{n \in \mathbb{N}^*}$ relative to a valid reference point r. If the Pareto front is described by a bilipschitz function with a Hölder continuous derivative, then for all $\epsilon > 0$, for n large enough and for any non-empty gap regions of $\mathcal{S}_n$, $\mathcal{G}^r_{\mathcal{S}_n,i}$ and $\mathcal{G}^r_{\mathcal{S}_n,j}$ with $i, j \in [\![1, n+1]\!]$, we have*

$$\frac{\lambda(\mathcal{G}^r_{\mathcal{S}_n,i})}{\lambda(\mathcal{G}^r_{\mathcal{S}_n,j})} \leq 4 \times \frac{(1+\epsilon)^2}{1-\epsilon} \quad . \tag{4.30}$$

Proof *Fix $\epsilon > 0$. By Lemma 10, there exists $N_1 \in \mathbb{N}^*$ such that for all n greater than N_1,* (4.28) *and* (4.29) *are verified for any non-empty gap region of $\mathcal{S}_n$ and its children. Since $\max_{i \in [\![1,n+1]\!]}(w^n_{i,r} - w^n_{i-1,r})$ converges to 0 by Lemma 8, every non-empty gap region is filled at some point. Take N_2 such that all the non-empty gap regions of $\mathcal{S}_{N_1}$ are filled at iteration N_2. For all n greater than N_2,* (4.28) *and* (4.29) *are true for any non-empty gap region of $\mathcal{S}_n$ and its children, but also for its parents.*

Take $n \geq N_2$. We note $\mathcal{G}_1 := \mathcal{G}^r_{\mathcal{S}_n,i}$ and $\mathcal{G}_2 := \mathcal{G}^r_{\mathcal{S}_n,j}$ two distinct non-empty gap regions of $\mathcal{S}_n$, and $\mathcal{P}_1$ and $\mathcal{P}_2$ their respective parents. When two sets correspond to gap regions relative to the same greedy set $\mathcal{S}_m$, we say that they cohabit at iteration m. Since only one vector is added to $\mathcal{S}_n$ at a time, the cohabitation of $\mathcal{G}_1$ and $\mathcal{G}_2$ implies that either $\mathcal{G}_1$ and $\mathcal{P}_2$ or $\mathcal{G}_2$ and $\mathcal{P}_1$ cohabit at some earlier iteration. In the first case, there necessarily exists $m \geq N_2$ such that $\mathcal{P}_2$ and $\mathcal{G}_1$ are gap regions relative to $\mathcal{S}_m$ and v_{m+1} belongs to $\mathcal{P}_2$, otherwise, $\mathcal{G}_1$ and $\mathcal{G}_2$ would not cohabit. By (4.28)*, the maximum hypervolume improvement to $\mathcal{S}_m$ of a vector of $\mathcal{G}_1$ and of a vector of $\mathcal{P}_2$ are at least $\frac{1}{2} \times (1-\epsilon) \times \lambda(\mathcal{G}_1)$ and at most $\frac{1}{2} \times (1+\epsilon) \times \lambda(\mathcal{P}_2)$, respectively. Since a vector of $\mathcal{P}_2$, v_{m+1}, maximizes the hypervolume improvement to $\mathcal{S}_m$, we have $\lambda(\mathcal{G}_1) \times \frac{1}{2} \times (1-\epsilon) \leq \lambda(\mathcal{P}_2) \times \frac{1}{2} \times (1+\epsilon)$. Since $\lambda(\mathcal{P}_2)$ is smaller than $4 \times (1+\epsilon)$ times the area of its child $\lambda(\mathcal{G}_2)$ by* (4.29)*, this inequality implies* (4.30)*. In the second case, $\mathcal{P}_2$ is filled before $\mathcal{P}_1$. Thus,*

there exists $m \geq N_2$ such that $\mathcal{P}_1$ and $\mathcal{P}_2$ cohabit at iteration m and v_{m+1} belongs to $\mathcal{P}_2$. Since the area of $\mathcal{P}_1$ is larger than the one of its child $\mathcal{G}_1$, the hypervolume improvement of v_{m+1} to $\mathcal{S}_m$ is still larger than $\frac{1}{2} \times (1-\epsilon) \times \lambda(\mathcal{G}_1)$. The rest of the argumentation remains valid.

We now have all the results needed to analyze the asymptotic impact of doubling the number of points in the greedy set. To prove the following asymptotic upper bound, we rely on similar arguments as for its nonasymptotic counterpart, Proposition 11. The previous lemma guarantees that the impact of doubling the number of points in the greedy set is asymptotically similar to the impact of passing from n points to $2n+1$.

Proposition 12 *Let $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ be a greedy set sequence relative to valid reference point r. If the Pareto front is described by a bilipschitz function f with a Hölder continuous derivative, then for all $\epsilon > 0$, we have for n large enough*

$$\frac{HV_r(PF_f) - HV_r(\mathcal{S}_{2n})}{HV_r(PF_f) - HV_r(\mathcal{S}_n)} \leq 1 - \frac{1}{2} \times (1-\epsilon) + \frac{1-\epsilon}{2 \times (n-1)} \ . \tag{4.31}$$

Proof *Fix $\epsilon > 0$. Fix n large enough to verify* (4.28) *and* (4.30) *for this particular ϵ.*

Let σ be a permutation of $[\![1, n+1]\!]$ such that the i-th gap region of $\mathcal{S}_n$ is filled by $v_{n+\sigma(i)}$ when it is filled before iteration $2n+1$. With this choice of permutation, $\mathcal{G}^r_{\mathcal{S}_n,i}$ is always a gap region of $\mathcal{S}_{n+\sigma(i)-1}$. Thus, $HVI_r(v_{n+\sigma(i)}, \mathcal{S}_{n+\sigma(i)-1})$ is larger than the maximum hypervolume improvement of a vector of $\mathcal{G}^r_{\mathcal{S}_n,i}$ to $\mathcal{S}_{n+\sigma(i)-1}$, which is larger than $\frac{1}{2} \times (1-\epsilon) \times \lambda(\mathcal{G}^r_{\mathcal{S}_n,i})$ by (4.28)*. It is equivalent to say that the hypervolume improvement at iteration $n+\sigma(i)$ is larger than $\frac{1}{2} \times (1-\epsilon) \times \lambda(\mathcal{G}^r_{\mathcal{S}_n,i})$. Summing over $i \in [\![1, n+1]\!]$, we obtain that the hypervolume improvement between iteration n and $2n+1$ is larger than the sum over i of $\frac{1}{2} \times (1-\epsilon) \times \lambda(\mathcal{G}^r_{\mathcal{S}_n,i})$, that is $\frac{1}{2} \times (1-\epsilon)$ times the optimality gap at iteration n.*

Now, we need to bound the hypervolume improvement at iteration $2n+1$, that is $HVI_r(v_{2n+1}, \mathcal{S}_{2n})$. It is smaller than $\frac{1}{2} \times (1+\epsilon) \times \max_{i\in[\![1,2n+1]\!]} \lambda(\mathcal{G}^r_{\mathcal{S}_{2n},i})$ by (4.10) *and* (4.28)*. Since the area of a gap region is smaller than the one of its parent, the maximum area of a gap region is lower at iteration $2n$ than at iteration n. The maximum area of one of the more than $n-1$ gap regions of $\mathcal{S}_n$ is itself smaller than $\frac{1}{n-1} \times \frac{4\times(1+\epsilon)^2}{1-\epsilon}$ times the optimality gap at iteration n by* (4.30)*.*

We conclude that the relation between the optimality gap at iteration $2n$ and at iteration n is smaller than $1 - \frac{1}{2} \times (1-\epsilon) + \frac{1-\epsilon}{2\times(n-1)}$.

We roughly follow the same approach to obtain the following asymptotic lower bound on the impact of doubling the number of points in the greedy set. Lemmas 10 and 11 are key to prove an upper bound on the hypervolume improvement at iteration k. They allow to prove that filling a gap region of $\mathcal{S}_n$ more than once gives, up to a factor $1+o(\epsilon)$, a lower hypervolume improvement than filling a gap region which was not filled. Indeed, the area of a descendant of a gap region of $\mathcal{S}_n$ is at most $\frac{1}{4} + o(\epsilon)$ times the area of its parent by Lemma 10, which is itself at most $4 + o(\epsilon)$ times the area of any other gap region of $\mathcal{S}_n$ by Lemma 11.

Proposition 13 *Let $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ be a greedy set sequence relative to a valid reference point r. If the Pareto front is desribed by a bilipschitz function f with a Hölder continuous derivative, then for all $\epsilon > 0$, we have for n large enough*

$$\frac{HV_r(PF_f) - HV_r(\mathcal{S}_{2n})}{HV_r(PF_f) - HV_r(\mathcal{S}_n)} \geq 1 - \frac{1}{2} \times \frac{(1+\epsilon)^3}{(1-\epsilon)^2} - \frac{1+\epsilon}{2\times(n-1)} \ . \tag{4.32}$$

Proof *Fix $\epsilon > 0$. Fix n large enough to verify (4.28), (4.29) and (4.30) for this particular ϵ. Let $\delta \in \{-1, 0, 1\}$ be such that $\mathcal{S}_n$ has $n + \delta$ non-empty gap regions. Let $i_0 := 1$ when the left extreme gap region is empty and $i_0 := 0$ otherwise.*

Let σ be a permutation of $[\![1, n + \delta]\!]$ such that the i-th non-empty gap region of $\mathcal{S}_n$, $\mathcal{G}^r_{\mathcal{S}_n,i_0+i}$, is filled by the vector $v_{n+\sigma(i)}$ when it is filled before iteration $2n + \delta$. We distinguish two cases. In the first case, $v_{n+\sigma(i)}$ is the child of the i-th non-empty gap region of $\mathcal{S}_n$, and consequently its hypervolume improvement to $\mathcal{S}_{n+\sigma(i)-1}$ is at most $\frac{1}{2} \times (1 + \epsilon) \times \lambda(\mathcal{G}^r_{\mathcal{S}_n,i_0+i})$ by (4.28). In the second case, $v_{n+\sigma(i)}$ belongs to $\mathcal{G}^r_{\mathcal{S}_n,i_0+j}$, the j-th non-empty gap region of $\mathcal{S}_n$, with $j \neq i$ and, by definition of σ, fills a descendant of this gap region not $\mathcal{G}^r_{\mathcal{S}_n,i_0+j}$ itself. By (4.28), the hypervolume improvement of $v_{n+\sigma(i)}$ to $\mathcal{S}_{n+\sigma(i)-1}$ is still at most $\frac{1}{2}\times(1+\epsilon)$ times the area of the gap region it fills. By (4.29), the area of a descendant of $\mathcal{G}^r_{\mathcal{S}_n,i_0+j}$ is smaller than $\frac{1}{4\times(1-\epsilon)}$ times the area of its ancestor. By (4.30), we also know that the area of the i-th non-empty gap region of $\mathcal{S}_n$ is at most $4 \times \frac{(1+\epsilon)^2}{1-\epsilon}$ times the area of any other gap region of $\mathcal{S}_n$, in particular its i-th non-empty gap region. We conclude that the hypervolume improvement of $v_{n+\sigma(i)}$ to $\mathcal{S}_{n+\sigma(i)-1}$ is smaller than $\frac{1}{2}\times\frac{(1+\epsilon)^3}{(1-\epsilon)^2}\times\lambda(\mathcal{G}^r_{\mathcal{S}_n,i_0+i})$. To summarize, since $1+\epsilon$ is smaller than $\frac{(1+\epsilon)^3}{(1-\epsilon)^2}$, the hypervolume improvement at any iteration $n+\sigma(i)$ is smaller than $\frac{1}{2} \times \frac{(1+\epsilon)^3}{(1-\epsilon)^2} \times \lambda(\mathcal{G}^r_{\mathcal{S}_n,i})$. Summing over $i \in [\![1, n + \delta]\!]$, the hypervolume improvement from iteration n to $2n + \delta$ is smaller than $\frac{1}{2} \times \frac{(1+\epsilon)^3}{(1-\epsilon)^2}$ times the sum over i of $\lambda(\mathcal{G}^r_{\mathcal{S}_n,i})$, that is the optimality gap at iteration n.

Now, it is left to prove an upper bound on $HV_r(\mathcal{S}_{2n}) - HV_r(\mathcal{S}_{2n+\delta})$. This quantity is maximal for $\delta = -1$, where it is simply the hypervolume improvement at iteration $2n$. As in the previous proof, it is smaller than $\frac{1+\epsilon}{2\times(n-1)}$ times the optimality gap at iteration n. Therefore, the relation between the optimality gap at iteration $2n$ and at iteration n is larger than $1 - \frac{1}{2} \times \frac{(1+\epsilon)^3}{(1-\epsilon)^2} - \frac{1+\epsilon}{2\times(n-1)}$.

We combine the lower and upper asymptotic bounds to obtain the following theorem.

Theorem 6 *Consider a bi-objective optimization problem and a greedy set sequence $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ relative to a valid reference point r. If the Pareto front is described by a bilipschitz function f with a Hölder continuous derivative, we have*

$$\frac{HV_r(PF_f) - HV_r(\mathcal{S}_{2n})}{HV_r(PF_f) - HV_r(\mathcal{S}_n)} \xrightarrow[n\to\infty]{} \frac{1}{2} \ . \tag{4.33}$$

Consequently, for such Pareto front and reference point and under Assumption 2 of perfect single-objective optimization, HV-ISOOMOO algorithms relative to r satisfy (4.33) where $\mathcal{S}_n$ is replaced by $\mathcal{I}_n$, the final incumbents Pareto front approximation at iteration n.

4.6 Conclusion

We prove that HV-ISOOMOO algorithms coupled with a perfect single-objective solver have a $O(1/n)$ convergence rate on convex Pareto fronts and a $O(1/n^c)$ convergence rate on bilipschitz Pareto fronts with $c \leq 1$ depending on the bilipschitz constants where n is the number of meta-iterations. Each meta-iteration corresponds to a single-objective optimization run. Both bounds are tight over the class of Pareto fronts and reference points considered. They are reached for affine Pareto fronts and reference points dominating the nadir point. On convex Pareto fronts, the convergence rate is exactly $\Theta(1/n)$, the fastest convergence rate achievable by bi-objective optimization algorithms Chapter 3.

We also investigate numerically the non-asymptotic speed of convergence of HV-ISOOMOO algorithms coupled with a perfect single-objective solver on some simple convex and concave Pareto fronts. The optimality gap of the final incumbents Pareto front approximation at meta-iteration n is close to the optimality gap of n-optimal distributions, that is the lowest optimality gap achievable by a Pareto front representation of cardinal n. The ratio between these two optimality gaps fluctuates periodically with respect to n. At the lowest, the optimality gap of the final incumbents Pareto front approximation is only a few percent larger than the optimality gap of n-optimal distributions, while at the largest, it is 23% larger in the worst case. Both of these numerical and theoretical results show that greedily adding points maximizing the hypervolume contribution as in HV-ISOOMOO algorithms is an effective way to quickly increase the hypervolume.

Finally, we prove that for bilipschitz Pareto fronts with a Hölder continuous derivative, doubling the number of meta-iterations divides the optimality gap by a factor which converges asymptotically to two. This asymptotic behavior resembles what we would observe with an affine Pareto front and a reference point dominating the nadir point. Yet, it does not guarantee a $\Theta(1/n)$ convergence rate. Both $\left(\frac{\log(n)}{n}\right)_{n\in\mathbb{N}^*}$ and $\left(\frac{1}{n\times\log(n)}\right)_{n\in\mathbb{N}^*}$ are examples of sequences verifying this property which do not have a $\Theta(1/n)$ convergence rate. The convergence rate on nonconvex Pareto fronts could theoretically be worse than $\Theta(1/n)$, but not better Chapter 3.

Convergence rates with respect to meta-iterations similar to those achieved under the assumption of perfect single-objective optimization may be observed in practice. We expect it to be the case for efficient implementations of HV-ISOOMOO and easy multi-objective problems, where the single-objective solver should return good approximations of global optima. Additionally, lower bounds on the speed of convergence may be directly derived for practical HV-ISOOMOO algorithms if a non-asymptotic lower bound on the speed of convergence towards a global optimum is known for the single-objective solver SOOPTIMIZER and if SOOPTIMIZER stops late enough. To do this, one could rely on the approach described in Lemma 7 with $k = 1$ and consider the vectors v_n^* towards which the single-objective solver converges instead of the greedy vectors v_n. The v_n^* are global optima of the true subproblems solved during meta-iterations.

We expect that the approach we use to prove a lower bound on the speed of con-

vergence of HV-ISOOMOO coupled with a perfect single-objective solver generalizes to any number of objectives. Indeed, the gap regions can still be defined using local nadir points for any number of objectives $m \geq 3$ but they are no longer disjoints. This implies that we need to consider the hypervolume improvement from iteration n to $n+1$ instead of $2n+1$. Lemma 7 details how to do so. On top of that, the proof of Proposition 11 only requires an upper bound on the number of gap regions and a lower bound on the normalized maximum hypervolume for some categories of Pareto fronts. It is known that for $m = 3$ and $m > 3$, there are respectively less than $2n + 1$ [28] and $O(n^{\lfloor \frac{m}{2} \rfloor})$ [54] gap regions associated to a set of n points. We conjecture that lower bounds on the normalized maximum hypervolume can be proven for $m \geq 3$ for convex Pareto fronts and for Pareto fronts with an explicit representation which is ($L_{\min}$,$L_{\max}$)-bilipschitz w.r.t. all variables. A possible generalization of Theorem 6 asymptotic insight is more open.

Acknowledgments

We would like to thank Dimo Brockhoff and Nikolaus Hansen for several helpful discussions, notably about the importance of decomposing the optimality gap and for the conjecture of (4.33).

CHAPTER 5

An HV-ISOOMOO algorithm based on CMA-ES: MO-CMA-ES-2

Joint work with Dimo Brockhoff, Nikolaus Hansen, Cheikh Toure and Anne Auger.

Abstract

We introduce a new bi-objective algorithm based on the HV-ISOOMOO framework which relies on the state-of-the-art evolutionary algorithm CMA-ES to solve subproblems: MO-CMA-ES-2. This new algorithm is aligned with previous works using the single-objective optimization algorithm CMA-ES as the basis of multi-objective optimization algorithms such as MO-CMA-ES [51], UP-MO-CMA-ES [59] and COMO-CMA-ES [74]. The MO-CMA-ES-2 algorithm can be seen as an extension of COMO-CMA-ES where the CMA-ES runs are done one after the other instead of simultaneously.

On the **doublesphere** problem, the optimality gap measured with the hypervolume converges in $\Theta(1/n)$, the optimal convergence speed, where n is the meta-iteration index. More precisely, the exact convergence speed is close to that of HV-ISOOMOO coupled with a perfect single-objective solver, and thus of the convergence speed of n-optimal distributions. The convergence speed with respect to function evaluations l is slower than $\Theta(1/l)$, but resembles $\Theta(1/l^{\alpha})$ with α close to 1. This shows that the theoretical upper bound on the convergence speed of $\Omega(1/l)$ associated with l-optimal distributions can be approached by real algorithms, even if it may not be reachable.

Moreover, the benchmark of MO-CMA-ES-2 using the COCO benchmarking platform shows state-of-the-art performance.

5.1 Introduction

In Chapter 4, we introduced HV-ISOOMOO (see Algorithm 1), a generic framework for building bi-objective optimization algorithms. We analyzed the performance of HV-ISOOMOO coupled with a perfect single-objective optimizer, which we call *ideal HV-ISOOMOO*. We proved that when the Pareto front is convex, the convergence speed with respect to the number of meta-iterations n of the ideal HV-ISOOMOO is $\Theta(1/n)$. This convergence speed cannot be surpassed by any algorithm on such problems, as stated in the Corollary 3 of Chapter 3. Experiments on convex and non-convex Pareto fronts provided additional evidence suggesting that HV-ISOOMOO algorithms could reach close to optimal convergence speed on easy problems. Indeed, on all chosen benchmark Pareto fronts, the exact convergence speed of the ideal HV-ISOOMOO with respect to the number of meta-iterations n appeared to be very similar to that of n-optimal distributions. In this chapter, we describe our attempt to build an efficient HV-ISOOMOO algorithm, that we call Multi-Objective Optimization Covariance Matrix Adaptation Evolution Strategy 2 (MO-CMA-ES-2).

The name MO-CMA-ES-2 stems from the use of the CMA-ES algorithm [46] to solve the single-objective optimization subproblems. The CMA-ES is a state-of-the-art black-box evolutionary algorithm which performs especially well on multimodal functions (with a large enough population size) [50] and on ill-conditioned functions (thanks to the covariance matrix adaptation) [9]. Conveniently, this algorithm also requires little or no parameter tuning. For a complete tutorial on CMA-ES, see [48].

Before MO-CMA-ES-2, there were already multi-objective optimization algorithms based on CMA-ES, including MO-CMA-ES [51], UP-MO-CMA-ES [59] and COMO-CMA-ES [74]. The MO-CMA-ES was developed in 2007 [51]. It is a variant of CMA-ES adapted to multi-objective problems which runs simultaneously several $(1+\lambda)$-CMA-ES algorithms. Like most MOEAs, it aims at providing a fixed size Pareto front approximation, consisting of as many solutions as there are $(1 + \lambda)$-CMA-ES running. The ranking of the solutions during the selection process involves two steps. In a first step, solutions are ranked with respect to domination: first, the non-dominated solutions, then the non-dominated solutions among those remaining, and so on. In a second step, solutions with the same domination rank are ranked according to either their crowding distance [31] or their hypervolume contribution. In 2010, a more efficient step size adaptation for MO-CMA-ES was introduced [76]. In 2016, an extension of MO-CMA-ES with a dynamic population size was designed, UP-MO-CMA-ES [59]. The population of UP-MO-CMA-ES contains all the non-dominated solutions evaluated so far, removing the need for dominated ranking. This population offers the benefit of providing an unbounded size approximation of the Pareto front. In 2019, the non-elitist algorithm COMO-CMA-ES was introduced [74]. Contrary to MO-CMA-ES, the CMA-ES involved use the non-elitist comma-selection. Comma-selection operates among offspring only, as opposed to plus-selection which operates among offspring and parents. A benchmark comparing MO-CMA-ES and COMO-CMA-ES can be found in [36].

The COMO-CMA-ES algorithm introduces the uncrowded hypervolume improve-

ment indicator (UHVI) as an alternative to the dual ranking of MO-CMA-ES. The UHVI is plainly equal to the hypervolume improvement (HVI) for non-dominated solutions. For dominated solutions, the UHVI is equal to minus the distance to the empirical Pareto front. We draw inspiration from COMO-CMA-ES and use the UHVI to build the objective function of the single-objective optimization problems in MO-CMA-ES-2. The MO-CMA-ES-2 algorithm can be seen as an extension of COMO-CMA-ES where CMA-ES are run one after the other instead of simultaneously, and thus can provide an unbounded size Pareto front approximation.

In practice, it is more convenient not to fix the number of solutions in the Pareto front approximation (PFA), and thus bound its quality, prior to running the algorithm. Oftentimes, one is interested in the best PFA that can be found in a given time. In such a case, algorithms able to provide an unbounded size PFA are more convenient. Typically, archiving of interesting evaluated solutions allows what would be bounded size algorithms to provide an unbounded size PFA [70]. However, this can only provide an approximation of unbounded quality in the case where the entire Pareto front is explored. It is not the case for the COMO-CMA-ES algorithm, as its successful runs converge to a n-optimal distribution.

The rest of the chapter is organized as follows. In Section 5.2, we outline how the CMA-ES algorithm operates, and we formally define the UHVI indicator. In Section 5.3, we introduce the MO-CMA-ES-2 algorithm. In Section 5.4, we investigate the performance of MO-CMA-ES-2. We benchmark this new algorithm using the platform COCO and compare it with three algorithms also based on CMA-ES: COMO-CMA-ES, HMO-CMA-ES and UP-MO-CMA-ES. In Section 5.5, we investigate the convergence speed of MO-CMA-ES-2 with greater detail on the **doublesphere** problem. Finally, we summarize the results and provide new research perspective in Section 5.6.

5.2 Background

Basic concepts related to multi-objective optimization and hypervolume have already been introduced in the previous chapters. We recall that we consider the minimization of a bi-objective function F over the search space $\Omega \subset \mathbb{R}^d$.

In this section, we present the CMA-ES algorithm and provide more details about the uncrowded hypervolume improvement.

5.2.1 The Single-Objective Optimization Algorithm CMA-ES

The CMA-ES acronym stands for Covariance Matrix Adaptation Evolution Strategy. Evolution strategies are evolutionary algorithms dedicated to solving continuous optimization problems [10]. They are stochastic optimization algorithms that evolve a population of candidate solutions by mutation, recombination, and selection. The CMA-ES is an adaptive algorithm where candidate solutions are sampled from a multivariate normal distribution parametrized by a mean vector and a covariance matrix of the form $\sigma^2 C$ (where σ is a positive parameter called step size and C is a positive definite ma-

trix). Both the mean, the step size and the matrix C vary depending on the iteration index. The incumbent solution of CMA-ES is the mean, which ideally converges towards the optimum of the single-objective function h. The matrix C encodes a metric on the search space. On convex-quadratic functions, the matrix C converges asymptotically to the inverse of the Hessian of h, up to a constant factor [47]. This means that for such functions, the level sets of the Gaussian distribution and the single-objective function h are asymptotically similar in shape.

We use the following definition of the state of an algorithm, which was taken from [53]:

Definition 6 *The state of the algorithm at a given point in its execution is the collection of values of all variables contained in the algorithm at that point.*

We denote the state of the algorithm by Θ. It contains the mean vector `mean`, the step size σ and the matrix C as well as two *evolution paths*. The evolution paths are vectors used to adapt the step size σ and the matrix C. Contrary to other classic definitions for which the state only contains parameters that are modified in the course of the algorithm, we also include in the state the fixed parameters of CMA-ES such as the learning rate for the matrix C, the population size λ and the weights used for the update of the mean.

Overall, the algorithm loops over three main steps, described in Algorithm 2. First, given the current state of the algorithm, λ candidate solutions, the *offspring*, are sampled in the search space (line 1). Then, the h values of the candidate solutions are computed (line 2). We refer to that operation as the evaluation of the candidate solutions. In this context, the objective function is often referred to as the *fitness function*. Finally, the state of the algorithm is updated using the information gathered on h through the evaluation of the candidate solutions (line 3).

The update of the state only takes into account the rank of the solutions when ordered by their fitness function values. This feature makes CMA-ES invariant under order-preserving transformations of the fitness function h such as the multiplication of h by a scaling factor $\alpha > 0$. The CMA-ES is also invariant under the rotation of the search space [47].

Algorithm 2 CMA-ES main steps

1: $x_1, \ldots, x_\lambda \leftarrow \text{sample}_\lambda(\Theta)$ # sample λ offspring
2: fitness_values $\leftarrow (h(x_1), \ldots, h(x_\lambda))$ # compute fitness values
3: $\Theta \leftarrow \text{update}(\Theta, (x_1, \ldots, x_\lambda), \text{fitness\_values})$ # update CMA-ES state

5.2.2 Uncrowded Hypervolume Improvement (UHVI)

Our algorithm MO-CMA-ES-2 is based on the *Uncrowded HyperVolume Improvement* (UHVI) fitness function that has been introduced in [74]. The UHVI is a modification of the hypervolume improvement designed to remove flatness. We briefly discussed it in Chapter 4 and define it extensively now.

Let u be an objective vector, S a set of objective vectors, and r a reference point. If u dominates r and is not dominated by S, then the *uncrowded hypervolume improvement* with respect to r of the vector u to the set S is $\mathrm{UHVI}_r(u, S) = \mathrm{HVI}_r(u, S)$. In this case, the uncrowded hypervolume improvement is always positive. An example is provided by v_1 in Figure 5.1.

Otherwise, $\mathrm{UHVI}_r(u, S) = -d(u, \mathrm{EPF}_{S,r})$ with $\mathrm{EPF}_{S,r}$ being the *empirical Pareto front* of the set S relative to r and d denoting a distance. In this case, the uncrowded hypervolume improvement is always negative. Given a reference point r and a set S of vectors of $\mathbb{R}^k$, the *empirical Pareto front* (the bold blue line in Figure 5.1) denoted by $\mathrm{EPF}_{S,r}$ is the boundary of the region of the objective space dominating the reference point r and dominated by an element of S. Here, we use the distance $d : u, S' \mapsto \inf_{v \in S'} \|u - v\|_2$. Examples are provided by v_2 and v_3 in Figure 5.1.

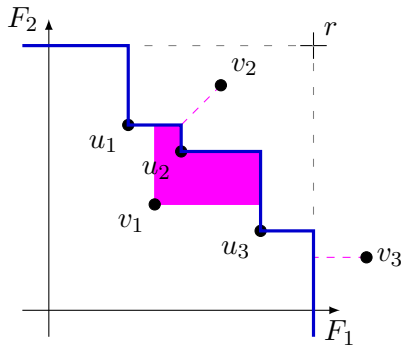


Figure 5.1: Illustration of the uncrowded hypervolume improvement with respect to a set $S := \{u_1, u_2, u_3\}$ and of the empirical Pareto front $\mathrm{EPF}_{S,r}$ (—): $\mathrm{UHVI}_r(v_1, S)$ is the area of the ■ region while $\mathrm{UHVI}_r(v_2, S)$ and $\mathrm{UHVI}_r(v_3, S)$ are equal to minus the length of the associated - - lines.

5.3 The MO-CMA-ES-2 algorithm

In this section, we introduce a new bi-objective optimization algorithm, that we call Multi-Objective Optimization Covariance Matrix Adaptation Evolution Strategy 2 (MO-CMA-ES-2).

5.3.1 Overview

The MO-CMA-ES-2 algorithm follows the HV-ISOOMOO framework defined in Chapter 4. At each meta-iteration, a version of CMA-ES with additional stopping criteria is run to find a solution with a large uncrowded hypervolume improvement with respect to the current Pareto front approximation, see Figure 5.2. More formally, the criterion used to define the objective functions of the single-objective subproblems is $J : \mathcal{I}, X \mapsto \mathrm{UHVI}_r(F(X), \mathcal{I})$ where X is a search vector and $\mathcal{I}$ is a set of objective

vectors. In practice, the vectors X are candidate solutions while $\mathcal{I}$ is the image in the objective space of the set of the final incumbents of previous runs of CMA-ES, i.e. the final means of the sampling distribution. The reference point r is a parameter of MO-CMA-ES-2.

The rationale behind the HV-ISOOMOO framework is that the set of final incumbents of the CMA-ES runs will hopefully provide a good approximation of the Pareto front with respect to the hypervolume. The image of the non-dominated subset of the final incumbents of the runs of the single-objective optimization solver forms what we called the *final incumbents Pareto front approximation* (FI-PFA) in Section 4.3.1. When the single-objective solver is CMA-ES, we call it *final means Pareto front approximation* (FM-PFA) and denote it by $\mathcal{M}_{\text{obj}}$. After p runs of CMA-ES, the final means Pareto front approximation provides a p'-points approximation of the Pareto front, with $p' = p$ when no final mean is dominated and $p' \leq p$ in the general case.

During the CMA-ES runs, many solutions are evaluated besides the final means, including solutions that provide additional information on the shape of the Pareto front. The image of the non-dominated subset of all solutions evaluated so far provides another Pareto front approximation, which we call *archive Pareto front approximation* (A-PFA). The A-PFA dominates the final means Pareto front approximation. The archiving does not influence the behavior of MO-CMA-ES-2.

In Figure 5.2, we represent a simplified abstraction of MO-CMA-ES-2, which does not take into account archiving.

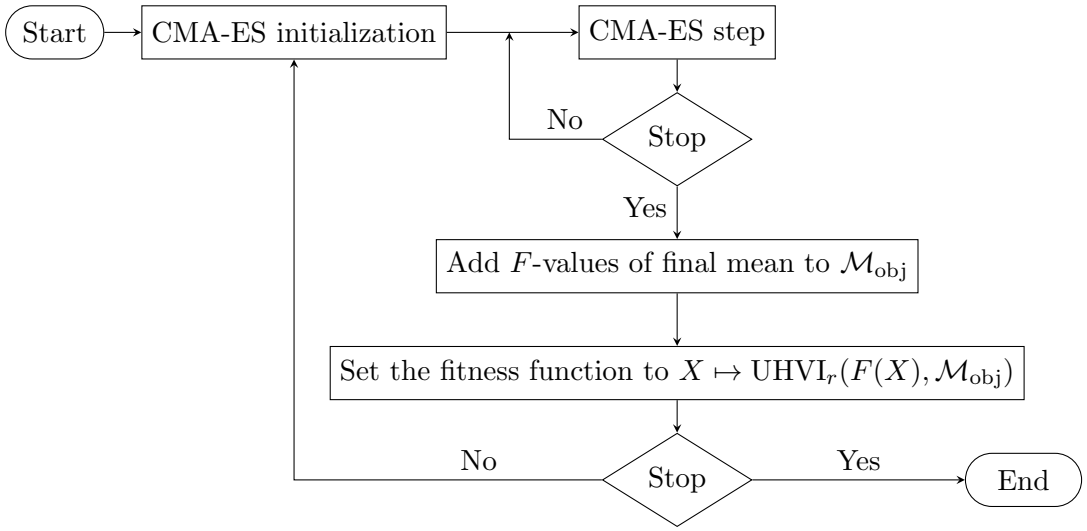


Figure 5.2: Simplified view of MO-CMA-ES-2

5.3.2 Characteristics of CMA-ES runs

While CMA-ES does not require parameter tuning, we still need to set both the stopping criteria and the initial state.

Stopping criteria

We want the stopping criteria to become more and more constraining in a way that is adapted to the evolution of a MO-CMA-ES-2 run. Indeed, it is desirable for the precision of the final means of the CMA-ES runs to increase as the FM-PFA grows and becomes denser.

To this end, we introduce a new stopping criterion: `tolfunrel`. The variable `median0` stores the median of the fitness values of the offspring the first time the empirical Pareto front dominates less than half of the offspring. The `tolfunrel` stopping criterion is triggered when the fitness variation among the offspring is strictly inferior to the `tolfunrel` factor times the difference between the median of the fitness values of the offspring and `median0`. In practice, we set the `tolfunrel` factor to 0.1. Setting the `tolfunrel` factor to the smaller value of 0.01 instead appeared to have almost no impact on the algorithm performance.

We also dynamically adapt the CMA-ES default `tolx` criterion through the MO-CMA-ES-2 run. The `tolx` criterion relates to the search space and detects when the standard deviation of the sampling gaussian distribution of all the search space variables are small. We set the `tolx` criterion precision to 10^{-4} at the beginning and update it as follows. The `tolx` precision of a CMA-ES run is 5% of the smallest final `tolx` value of the earlier CMA-ES runs that verify the following condition: their final mean dominates the reference point r and is not dominated by any of the other CMA-ES final means. The initial `tolx` criterion precision only impact the behavior of MO-CMA-ES-2 when the first CMA-ES run stops due to the triggering of the `tolx` stopping criterion.

In addition, we want to prevent the stopping criterion `tolfacupx` from being triggered. Indeed, this criterion is triggered by a significant increase of the step size, which is expected behavior at the beginning of `best_chv` runs. For this reason, we set the `tolfacupx` criterion precision to 10^{15}. Other than that, we set all other stopping criteria of CMA-ES with default values.

Initialization

We initialize the CMA-ES runs in two different ways, that we call `random` and `best_chv`.

For the `random` initialization, we rely on the default initialization of CMA-ES, which does not assume much knowledge of the optimization problem. The CMA-ES algorithm only requires an initial point x_0 and an initial step size σ_0; the initial covariance matrix is set by default to the identity matrix. The initial step size σ_0 is the same for all `random` initialization. Typically, we want to set it large enough to explore a large part of the search space at the first iteration. The initial mean x_0 is sampled uniformly at random in the search space, which is assumed to be bounded, before each `random` initialization. The `random` initialization drive is exploration, which is necessary to leave the neighbourhood of local optima and discover more promising regions of the search space.

However, starting the search process from zero at the beginning of each meta-iteration can be inefficient. After the first run of CMA-ES (which relies necessarily on `random` initialization), we are in a position to exploit the data of earlier CMA-ES runs to speed

up the search process. This is done by the `best_chv` initialization.

The `best_chv` initialization starts by picking a state among the final states of earlier CMA-ES runs as described in Algorithm 3.

Algorithm 3 `best_chv` picking function

Argument: Θ **(history of the final states of CMA-ES runs)**
function PICK_STATE(Θ)
  $\mathcal{M}_{\text{obj}} \leftarrow \{\text{F}(\Theta.\text{mean}) \colon \Theta \in \Theta\}$
  $\text{PF}_{\text{cut}} \leftarrow \{\Theta \colon \Theta \in \Theta,\ \mathcal{M}_{\text{obj}} \not\preceq F(\Theta.\text{mean}),\ \text{F}(\Theta.\text{mean}) \preceq r\}$
  # Special case: no final mean dominates the reference point
  if $\text{PF}_{\text{cut}} = \emptyset$ **then**
    idx = random(1, ..., $|\Theta|$)
    return Θ_{idx} # a state is picked uniformly at random
  end if
  # We iterate over the states sorted by decreasing UHVI
  for Θ in Θ.sorted($\Theta \mapsto \text{UHVI}_r(F(\Theta.\text{mean}), \mathcal{M}_{\text{obj}})$) **do**
    # Case where the mean of the state is an extreme vector
    if $F(\Theta.\text{mean})$ is an extreme vector **then**
      if $\{F(\Theta.\text{mean}) : \Theta \in PF_{\text{cut}}\}$ only contains extreme vectors **then**
        return Θ
      else
        return Θ with probability $1/|\Theta|$
      end if
    else
      # Case where the mean of the state is not an extreme vector
      return Θ
    end if
  end for
end function

Roughly speaking, we pick the state whose final mean has the largest contributing hypervolume with respect to the FM-PFA. When no final mean dominates the reference point r, a state is picked at random. We handle differently the case where the final mean with the largest contributing hypervolume has the largest or smallest value among the final means for one objective. The rationale is to avoid picking extremes frequently. Then, the sampling distribution and the options are initialized by copying those of the picked state using the `copy_light` method of CMA-ES. Other state variables like the evolution paths are initialized following CMA-ES default.

We alternate between `random` and `best_chv` initialization. The function evaluations budget is shared equally betwen the two kinds of initialization. The kind of initialization used for a new CMA-ES run is the one for which previous CMA-ES runs initialized this way have done the fewest function evaluations so far. Since `random` runs are longer, there is a greater proportion of CMA-ES runs initialized using `best_chv`. This guarantees that

using both of these restarts, we cannot lose more than a factor 2 in convergence speed compared to using only the more efficient of the two. Here, we only introduce two kinds of initialization, but more advanced MO-CMA-ES-2 versions could use other initializations.

5.3.3 Full overview of MO-CMA-ES-2

We depict the outline of MO-CMA-ES-2 algorithm in Algorithm 4. In the interest of readability, we do not specify the initialization and the stopping criteria details presented in Section 5.3.2.

Algorithm 4 MO-CMA-ES-2 Algorithm

Given: σ_0 **(initial step size),** r **(reference point)**
Initialization: $\Theta =$ **empty list (history of the final states of CMA-ES runs),** $\mathcal{M}_{\text{obj}} =$ **empty set (history of the objective values of the final means of CMA-ES runs),** $\mathcal{R} =$ **empty list (history of the CMA-ES initialization methods)**
while MO-CMA-ES-2 stopping criterion not met **do**
    # Initialization
    restart_method$\leftarrow$ pick_restart($\mathcal{R}$)
    $\Theta \leftarrow$ restart_method(Θ, σ_0) # initialize the CMA-ES state
    Θ.tolx $\leftarrow$ compute_tolx(Θ)
    Θ.median0 $\leftarrow$ None
    # Run of CMA-ES
    while CMA-ES stopping criterion(Θ) not met **do**
        # CMA-ES step
        $x_1, \ldots, x_\lambda \leftarrow \text{sample}_\lambda(\Theta)$
        fitness_values $\leftarrow \left(\text{UHVI}_r(F(x_i), \mathcal{M}_{\text{obj}})\ i = 1, \ldots, \lambda\right)$
        $\Theta \leftarrow$ update($\Theta, (x_1, \ldots, x_\lambda)$, fitness_values)
        # Initialize median0 the first time the EPF dominates $\leq 1/2$ of the offspring
        if Θ.median0 is None and median(fitness_values) < 0 **then**
            Θ.median0 $\leftarrow$ median(fitness_values)
        end if
    end while
    # Archiving
    $\mathcal{R}$.append(restart_method)
    $\mathcal{M}_{\text{obj}} \leftarrow \mathcal{M}_{\text{obj}} \cup \{f(\Theta.\text{mean})\}$
    Θ.append(Θ)
end while

5.4 Benchmarking of MO-CMA-ES-2

We implemented MO-CMA-ES-2 for bi-objective problems. Our implementation is based on the COMO-CMA-ES code available at https://github.com/CMA-ES/pycomocma, but is not yet released. In this section, we evaluate the performance of this implementation on the `bbob-biobj` test suite of the COCO platform and compare it to other bi-objective optimization algorithms.

5.4.1 The COCO platform and the `bbob-biobj` test suite

We rely on the benchmarking platform COCO [49], available at https://github.com/numbbo/coco. This platform provides numerous test problems and tools for recording and post-processing data. The experiments were conducted using COCO version 2.3.1 and the post-processing using version 2.6.2. We use the `bbob-biobj` test suite, which consists of bi-objective optimization problems for which the search space is $[-5, 5]^d$.

The `bbob-biobj` test suite contains 55 raw problems, which are described in [21]. From each raw problem, 15 particular instances are built using basic transformations such as the translation of the optimal solution in the search space or the translation of one or both objective functions. These transformations are designed to add variability without changing the important properties of the optimization problem. Each raw test problem is available in dimensions 2, 3, 5, 10, 20 and 40.

The raw bi-objective functions $F = (F_1, F_2)$ are a combination of two single-objective functions. These single-objective functions vary in terms of separability, conditioning, structure, and multimodality. They are gathered in five groups: unimodal and separable, unimodal with low or moderate conditioning, unimodal with high-conditioning, multimodal with global structure and multimodal with weak global structure. The bi-objective functions can therefore be separated into fifteen categories, depending on the category of the two objective functions they are comprised of.

The COCO platform assesses the quality of the archive Pareto front approximation (A-PFA), the unbounded archive of all solutions evaluated so far, depending on the *runtime* defined as the number of function evaluations. The quality of the A-PFA is measured using a quality indicator that is equal to the hypervolume normalized by the area of the region of interest as long as the A-PFA dominates the nadir point [49]. The platform registers the runtime when the quality of the A-PFA reaches fixed precision targets. The 58 precision targets are positioned in relation to a reference quality at the following distances: $-10^{-4}, -10^{-4.1}, \ldots, -10^{-5}, 0, 10^{-}5, 10^{-4.9}, \ldots, 10^{0}$ [22]. This reference quality is the hypervolume of a Pareto front approximation obtained by running several multi-objective algorithms (NSGA-II, SMS-EMOA, MOEA/D, RM-MEDA, MO-CMA-ES, ...) on the test problems and recording all evaluated solution.

The COCO platform provides plots in lin-log scale of empirical cumulative distribution functions (ECDF). An ECDF assigns to a runtime the proportion of the targets that have been reached at this runtime. On the x-axis is the runtime normalized by the search space dimension. The data is always aggregated over ten instances of a problem, and sometimes over different problems for the same search space dimension. The more

targets that are reached at a fixed runtime, the better. Conversely, the sooner a fixed proportion of targets is reached, the better. The horizontal gap between the ECDFs of two different algorithms informs on the multiplicative speed-up between the two.

5.4.2 Benchmark parameters

We set the initial step size σ_0 to 2, which corresponds to 10% of the search space size. The remaining of the settings are chosen as in Section 5.3.2. We run MO-CMA-ES-2 for $5 \times 10^4 d$ function evaluations, where d is the dimension of the search space.

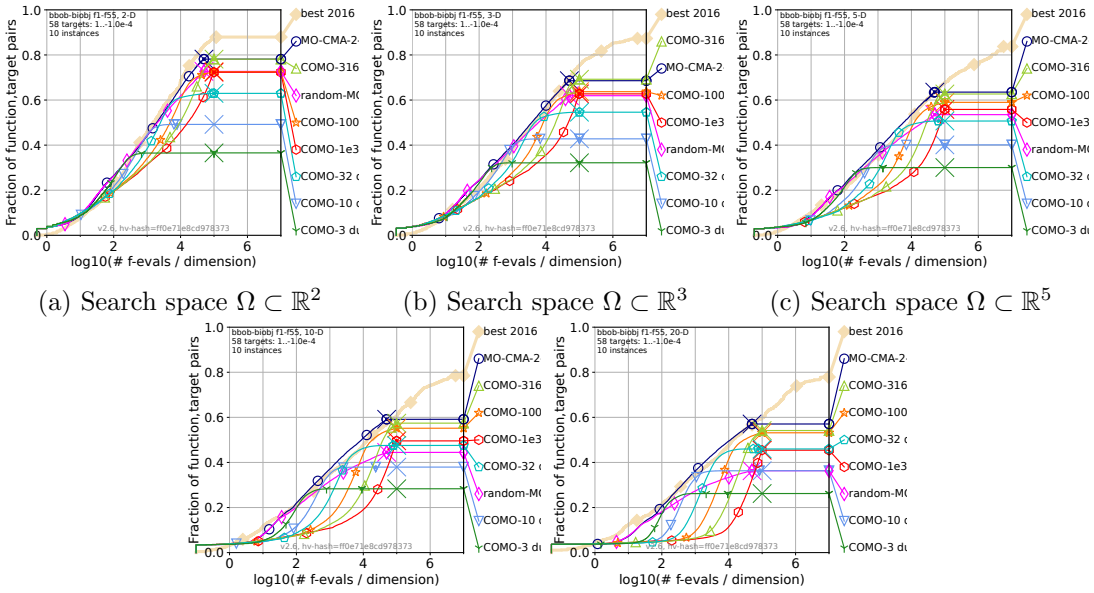
5.4.3 Comparison between COMO-CMA-ES and MO-CMA-ES-2

We compare MO-CMA-ES-2 with the COMO-CMA-ES algorithm which has already been benchmarked using the COCO platform for a budget of $10^5 d$ function evaluations [36]. We reuse the public benchmark data. Like MO-CMA-ES-2, COMO-CMA-ES is based on runs of the single-objective optimization algorithm CMA-ES. Unlike MO-CMA-ES-2, its goal is to provide a Pareto front approximation of finite size $p \in \mathbb{N}^*$. The COMO-CMA-ES does p simultaneous runs of CMA-ES. It runs through all the CMA-ES in a random order, one iteration at a time. After having done an iteration for each CMA-ES run, it starts again in a new random order. Ideally, the set of the p means of the CMA-ES runs converges to a p-optimal distribution. The objective function of a CMA-ES run is the uncrowded hypervolume improvement with respect to the set of the means of the other CMA-ES runs. Benchmark data is available for COMO-CMA-ES for the following numbers of CMA-ES: 3, 10, 100, 312 and 1000.

On top of that, we also compare the standard MO-CMA-ES-2 with a modified version of MO-CMA-ES-2 with `random` runs only. This comparison highlights the contribution of the `best_chv` runs to the performance of the standard MO-CMA-ES-2.

In Figure 5.3, we show the ECDFs aggregated over all test problems. Each search space dimension has its figure. We show the ECDFs for each test problem individually for search space dimension 10 in Figure 5.4 and for search space dimensions 2, 3, 5 and 20 in Appendix C.

The MO-CMA-ES-2 algorithm has been built with the intention of providing an extension of COMO-CMA-ES able to achieve good performance for any budget without fitting any algorithm parameter to the budget. For a fixed number of simultaneous CMA-ES runs, the budget range for which COMO-CMA-ES achieves competitive performance is limited, as can be seen in Figure 5.3. Therefore, COMO-CMA-ES requires picking the number of CMA-ES runs wisely to achieve good performance for a certain budget. If the number of CMA-ES runs is too large, the algorithm runs out of budget before converging to the Pareto front. If the number of CMA-ES runs is too small, only a portion of the budget is helpful. Indeed, after some time, the COMO-CMA-ES reaches a point where the quality of the Pareto front approximation provided by all solutions evaluated so far stops improving. The MO-CMA-ES-2 performs well for all budget. Aggregated over all test problems (see Figure 5.3), the standard MO-CMA-ES-2 performs as well or slightly better than COMO-CMA-ES with adequate number of CMA-ES runs. The



(a) Search space $\Omega \subset \mathbb{R}^2$ (b) Search space $\Omega \subset \mathbb{R}^3$ (c) Search space $\Omega \subset \mathbb{R}^5$

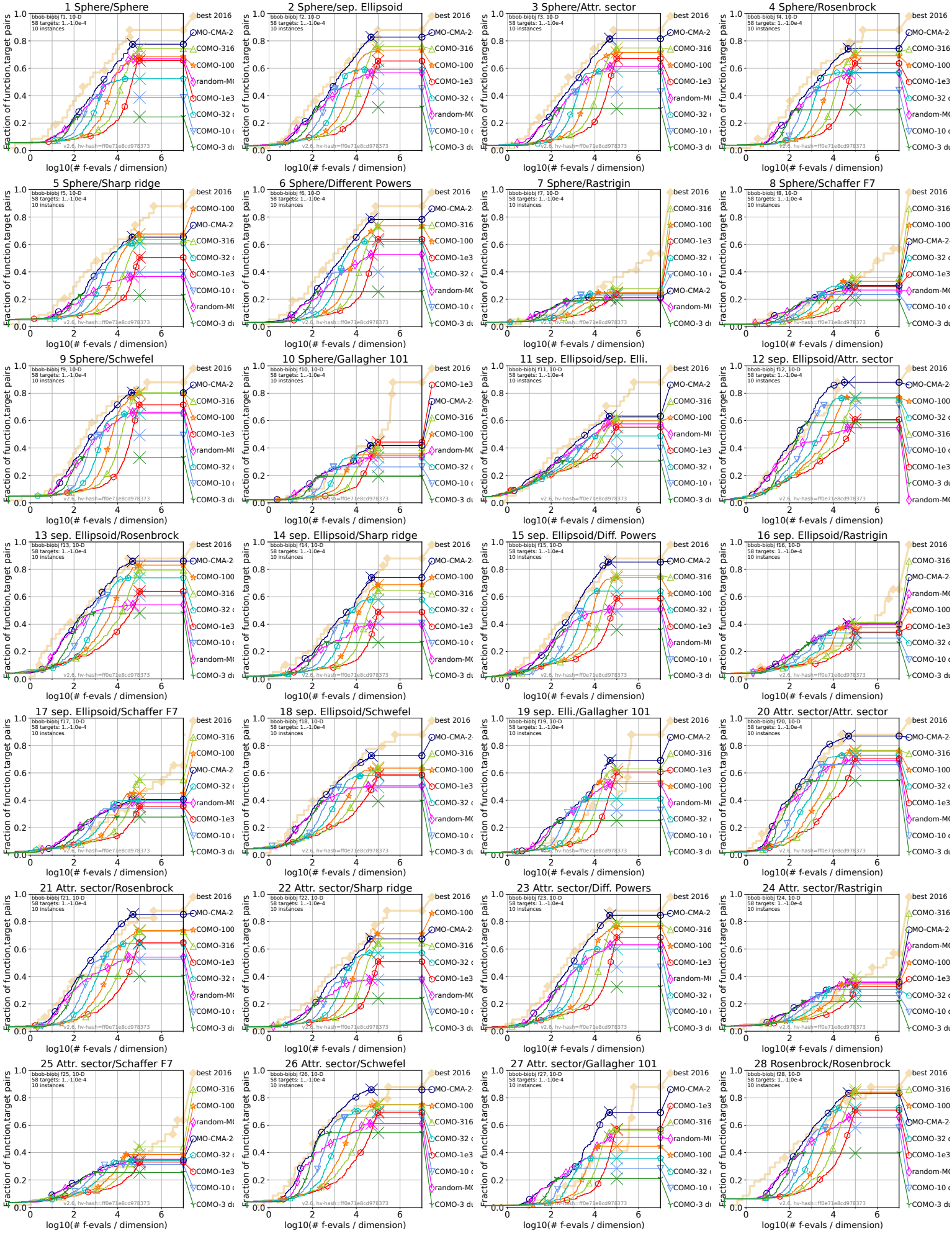
(d) Search space $\Omega \subset \mathbb{R}^{10}$ (e) Search space $\Omega \subset \mathbb{R}^{20}$

Figure 5.3: ECDFs of COMO-CMA-ES, the standard version of MO-CMA-ES-2 and MO-CMA-ES-2 with `random` runs only, aggregated over all `bbob-biobj` test problems for search space dimensions 2, 3, 5, 10 and 20.

advantage of MO-CMA-ES-2 slightly increases with budget and search space dimension. When looking at each test problem individually (see Figure 5.4), the picture is more mixed. When both objective functions are unimodal, MO-CMA-ES-2 performs as well as or better than COMO-CMA-ES. But, when one of the objective function is multimodal and has global structure (either the Schaffer or the Rastrigin function), with some exceptions when combined with the Sharp Ridge function, COMO-CMA-ES matches or outperforms MO-CMA-ES-2 for search space dimension up to 10. The MO-CMA-ES-2 often shows a loss of performance when one of the objective function is the Rastrigin function. In single-objective optimization, the Rastrigin function is known to require a large population size to be minimized successfully [50]. The performance of MO-CMA-ES-2 and COMO-CMA-ES with adequate number of CMA-ES, although not identical, are similar on most test problems. There are test problems on which MO-CMA-ES-2 performs significantly better for large budget, such as the problems 27, 34 and 55 in dimension 10, see Figure 5.4.

In Figure 5.3, we see that the `best_chv` runs are essential for MO-CMA-ES-2 to perform as well as COMO-CMA-ES, especially for large search space dimensions. The version of MO-CMA-ES-2 with `random` initialization only has poorer performance. When looking at the test problems individually, the standard MO-CMA-ES-2 appears to match or outperform the `random` version on most problems. There are a few exceptions for small search space dimensions (2 and 3), which all involve either the Rastrigin or the

Schaffer function, the two multimodal objective functions with global structure of the benchmark. Even then, the advantage of the `random` version is small. On the other hand, on many functions, the standard MO-CMA-ES-2 outperforms the `random` version by a large margin.


1 Sphere/Sphere
2 Sphere/sep. Ellipsoid
3 Sphere/Attr. sector
4 Sphere/Rosenbrock
5 Sphere/Sharp ridge
6 Sphere/Different Powers
7 Sphere/Rastrigin
8 Sphere/Schaffer F7
9 Sphere/Schwefel
10 Sphere/Gallagher 101
11 sep. Ellipsoid/sep. Elli.
12 sep. Ellipsoid/Attr. sector
13 sep. Ellipsoid/Rosenbrock
14 sep. Ellipsoid/Sharp ridge
15 sep. Ellipsoid/Diff. Powers
16 sep. Ellipsoid/Rastrigin
17 sep. Ellipsoid/Schaffer F7
18 sep. Ellipsoid/Schwefel
19 sep. Elli./Gallagher 101
20 Attr. sector/Attr. sector
21 Attr. sector/Rosenbrock
22 Attr. sector/Sharp ridge
23 Attr. sector/Diff. Powers
24 Attr. sector/Rastrigin
25 Attr. sector/Schaffer F7
26 Attr. sector/Schwefel
27 Attr. sector/Gallagher 101
28 Rosenbrock/Rosenbrock
Fraction of function,target pairs
log10(# f-evals / dimension)
best 2016
MO-CMA-2
COMO-316
COMO-100
COMO-1e3
COMO-32
COMO-10
COMO-3
random-MO

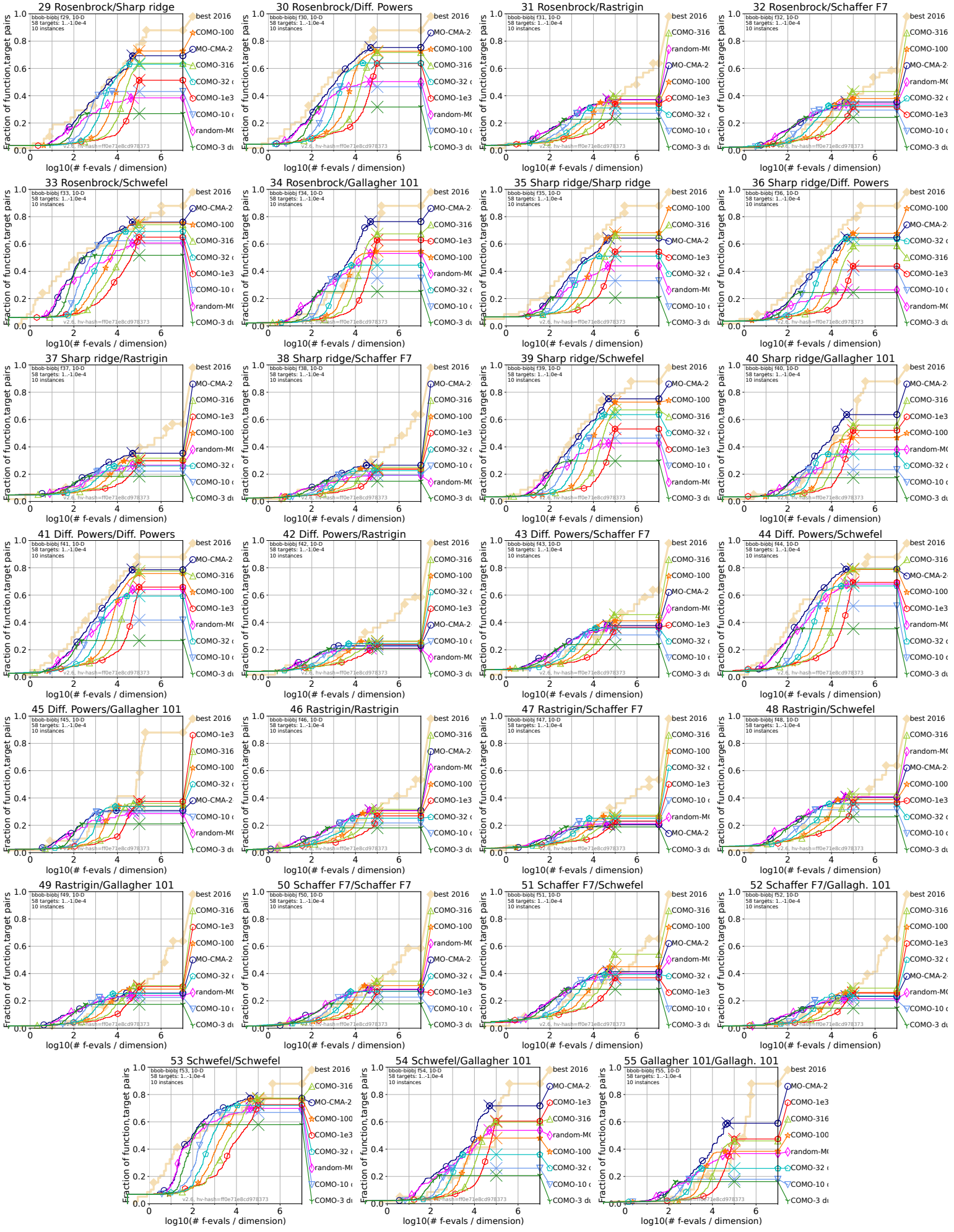


Figure 5.4: ECDFs of COMO-CMA-ES, the standard MO-CMA-ES-2 and a modified version of MO-CMA-ES-2 with `random` runs only, for all `bbob-biobj` test problems and for search space dimension 10.

5.4.4 Comparison between MO-CMA-ES-2, UP-MO-CMA-ES and HMO-CMA-ES

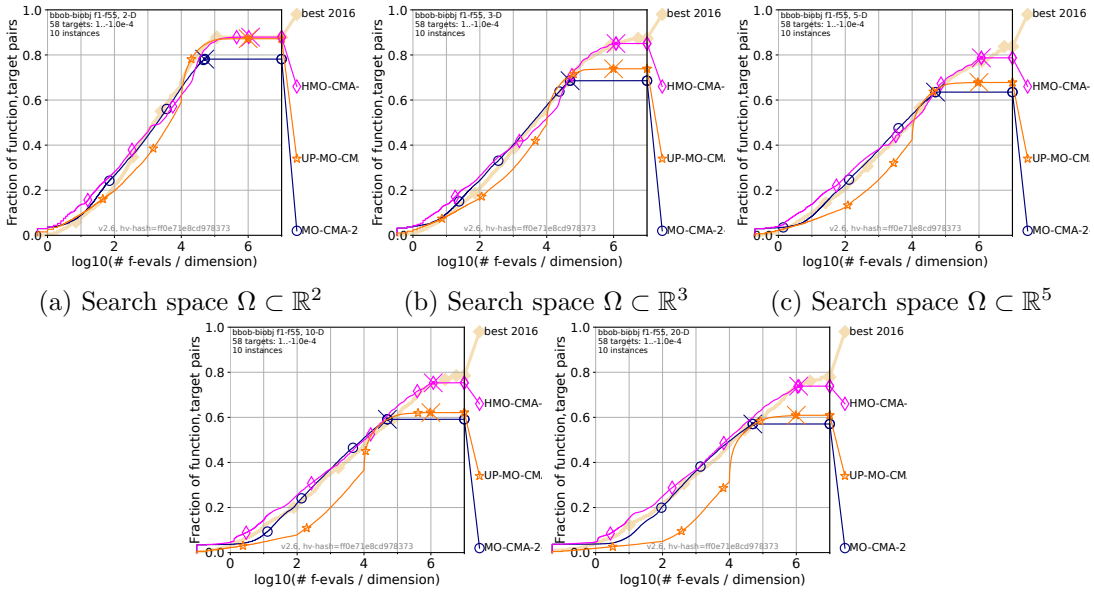
We compare MO-CMA-ES-2 with two bi-objective algorithms based on CMA-ES which, like MO-CMA-ES-2, aim at finding a Pareto front approximation of unbounded size: HMO-CMA-ES [61] and UP-MO-CMA-ES [59]. These two algorithms already have been benchmarked with COCO [61, 59]. We reuse the public data. The HMO-CMA-ES algorithm is also the algorithm with the best overall performance on the `bbob-biobj` test suite among all those that have been benchmarked up to 2021. We did not ascertain whether HMO-CMA-ES has been outperformed in 2022. We provide the ECDFs aggregated over all test problems in Figure 5.5. Each search space dimension has its figure. We also provide the ECDFs for each test problem individually for search space dimension 10 in Figure 5.6 and for the other search space dimensions (2, 3, 5 and 20) in Appendix C.

The HMO-CMA-ES algorithm is a hybrid of BOBYQA and four algorithms based on CMA-ES which is designed to perform well for any budget. A run of HMO-CMA-ES can be decomposed in four phases. It starts by running BOBYQA for $10d$ function evaluations to find rough approximations of minimizers of scalarizations of the form $\alpha F_1/F_1(0) + (1-\alpha)F_2/F_2(0)$ with $0 < \alpha < 1$. During the three phases that follow, two variants of MO-CMA-ES and a variant of CMA-ES are run, on the original function and on a linear aggregation of the objective functions, respectively. All three involve an increase of the population size. The third phase begins after $1000d$ function evaluations. It is only from this third phase onwards that algorithms involving restarts are run. Finally, the fourth phase starts after $20000d$ function evaluations.

In Figure 5.5, we see that MO-CMA-ES-2 achieves similar performance to HMO-CMA-ES, except for small budget (up to $10^2 d$ function evaluations) where HMO-CMA-ES performs consistently better. It is often the case that on a single problem, each algorithm outperforms the other, for different budgets. The HMO-CMA-ES is generally better for small budget while MO-CMA-ES-2 is better for medium budget, and HMO-CMA-ES catches up later.

The HMO-CMA-ES generally outperforms MO-CMA-ES-2 when both objective functions are unimodal. The magnitude of HMO-CMA-ES's advantage increases with the search space dimension. On the other hand, MO-CMA-ES-2 often outperforms HMO-CMA-ES for large budget when both functions are multimodal, with the exception of problems that involve the Schweffer function.

The Unbounded Population Multi-Objective Covariance Matrix Adaptation Evolution Strategy (UP-MO-CMA-ES) algorithm is a multi-objective evolutionary algorithm. Unlike most MOEAs, its population size is not fixed: the population is formed of all the non-dominated solutions evaluated so far. Except for extreme solutions, the probability for a member of the population to be selected to be a parent at some iteration of UP-MO-CMA-ES is proportional to its hypervolume contribution to the population. The covariance matrix adaptation is based on the relative positions in the search space of the two neighbors of the parent in the objective space among the population. The benchmarked algorithm first runs multiple MO-CMA-ES algorithms in parallel with a



(a) Search space $\Omega \subset \mathbb{R}^2$ (b) Search space $\Omega \subset \mathbb{R}^3$ (c) Search space $\Omega \subset \mathbb{R}^5$

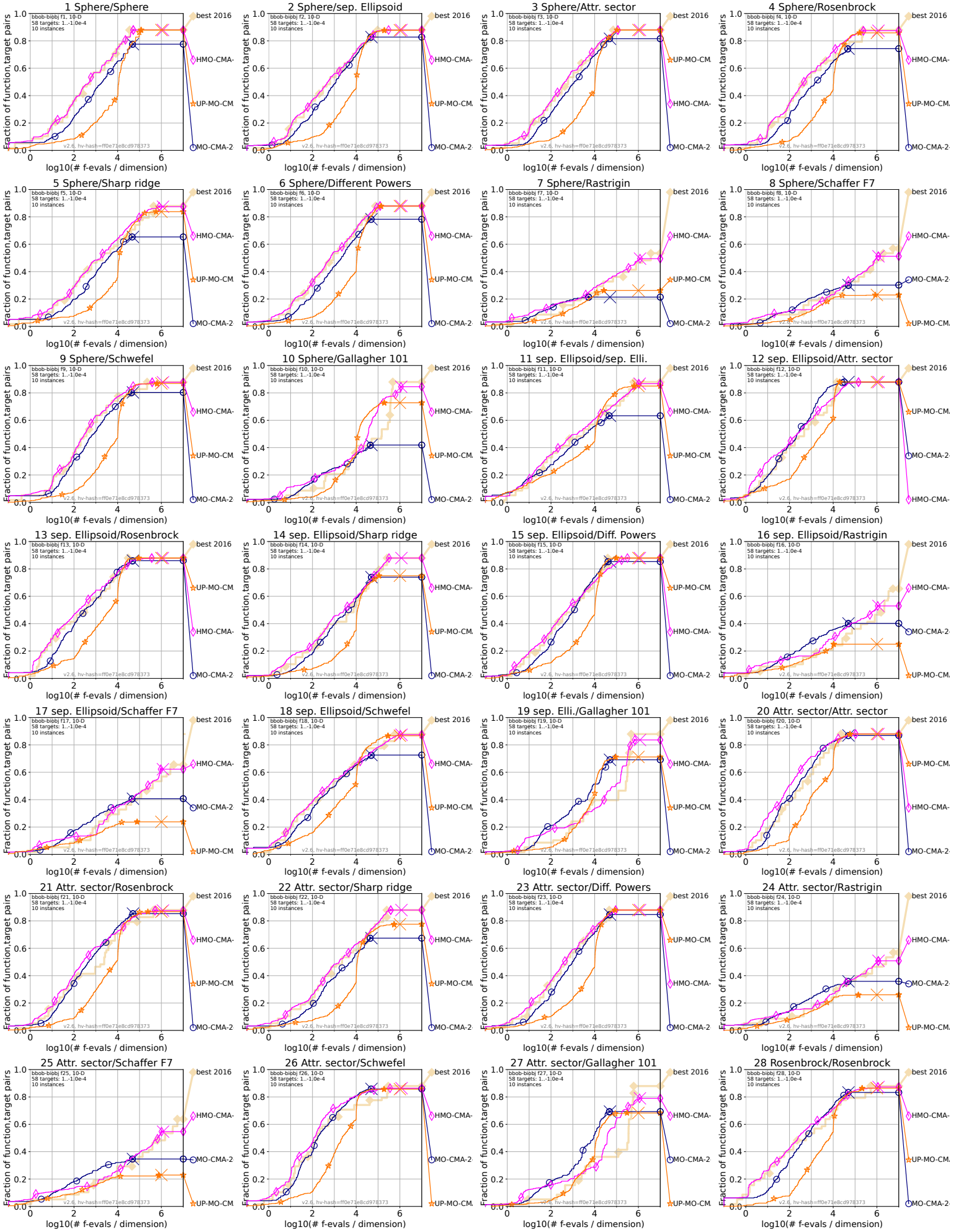
(d) Search space $\Omega \subset \mathbb{R}^{10}$ (e) Search space $\Omega \subset \mathbb{R}^{20}$

Figure 5.5: ECDFs of UP-MO-CMA-ES, HMO-CMA-ES and MO-CMA-ES-2 aggregated over all `bbob-biobj` test problems. Each search space dimension has its figure.

budget of $10^4 d$ function evaluations before running UP-MO-CMA-ES. The role of these MO-CMA-ES runs is to explore the search space. After $10^4 d$ function evaluations, the populations of the MO-CMA-ES are merged. The non-dominated solutions form the initial population of the UP-MO-CMA-ES run.

Over all test problems, with very few exceptions, MO-CMA-ES-2 achieves better performance than UP-MO-CMA-ES for any budget. This is reflected in the aggregated plots provided in Figure 5.5, as well as for each problem individually in Figure 5.6. After the end of the exploration phase, at $10^4 d$ function evaluations, UP-MO-CMA-ES starts to catch up and succeeds quickly. How UP-MO-CMA-ES and MO-CMA-ES-2 compare for large budget is unclear. Indeed, the exploration phase of the UP-MO-CMA-ES benchmark takes place up to $10^4 d$ function evaluations while the budget of the MO-CMA-ES benchmark is only $5 \times 10^4 d$ function evaluations. For a small budget, smaller than 10 times the search space dimension, UP-MO-CMA-ES does sometimes perform slightly better.

The gap between the ECDFs of MO-CMA-ES-2 and UP-MO-CMA-ES decreases with the budget and is almost closed for the maximum budget of $5 \times 10^4 d$ function evaluations, with few exceptions. Test problems where this gap does not close all involve either the Schaffer or the Rastrigin function, both highly multimodal functions.


1 Sphere/Sphere
2 Sphere/sep. Ellipsoid
3 Sphere/Attr. sector
4 Sphere/Rosenbrock
5 Sphere/Sharp ridge
6 Sphere/Different Powers
7 Sphere/Rastrigin
8 Sphere/Schaffer F7
9 Sphere/Schwefel
10 Sphere/Gallagher 101
11 sep. Ellipsoid/sep. Elli.
12 sep. Ellipsoid/Attr. sector
13 sep. Ellipsoid/Rosenbrock
14 sep. Ellipsoid/Sharp ridge
15 sep. Ellipsoid/Diff. Powers
16 sep. Ellipsoid/Rastrigin
17 sep. Ellipsoid/Schaffer F7
18 sep. Ellipsoid/Schwefel
19 sep. Elli./Gallagher 101
20 Attr. sector/Attr. sector
21 Attr. sector/Rosenbrock
22 Attr. sector/Sharp ridge
23 Attr. sector/Diff. Powers
24 Attr. sector/Rastrigin
25 Attr. sector/Schaffer F7
26 Attr. sector/Schwefel
27 Attr. sector/Gallagher 101
28 Rosenbrock/Rosenbrock
Fraction of function,target pairs
log10(# f-evals / dimension)
best 2016
HMO-CMA-
UP-MO-CM
MO-CMA-2

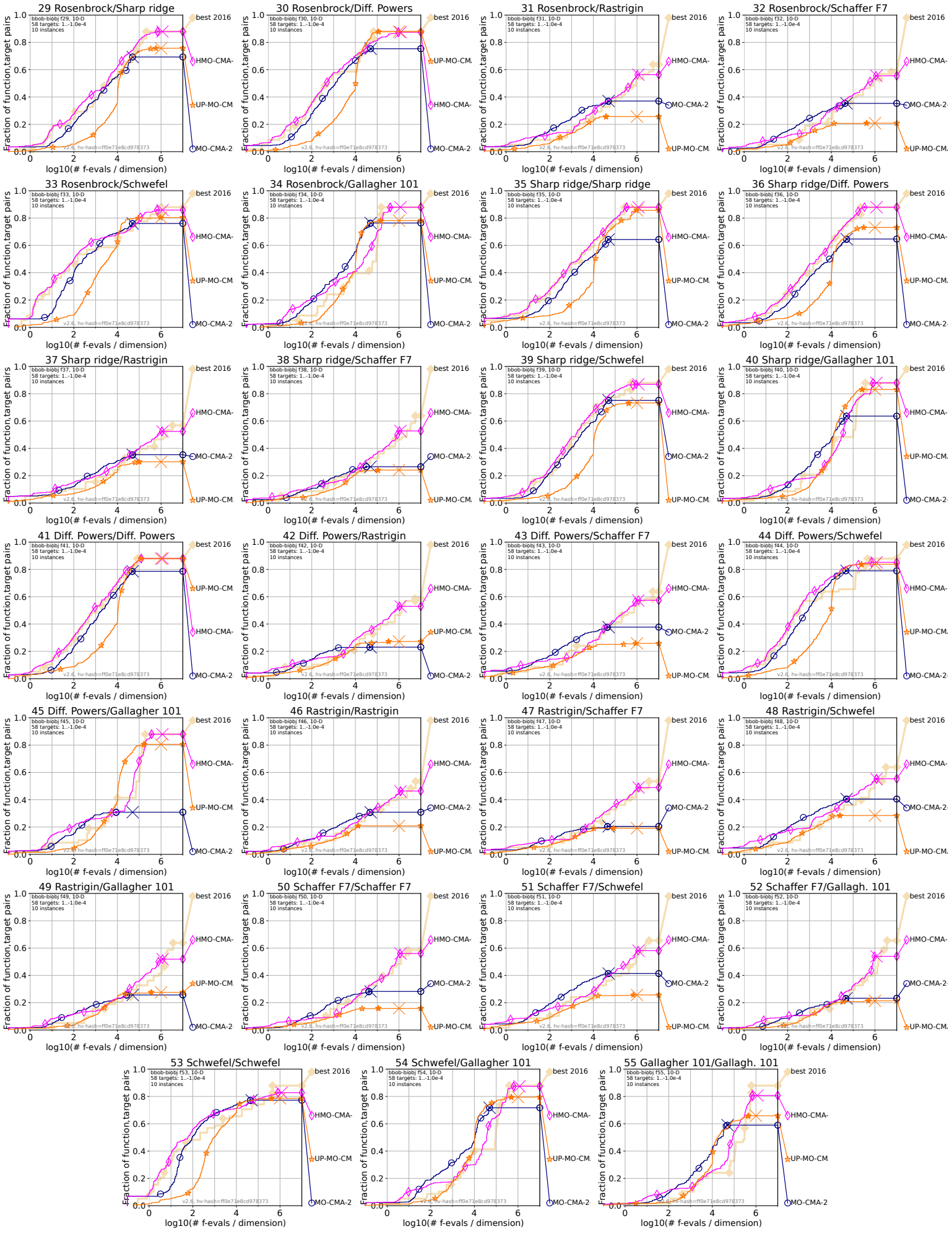


Figure 5.6: ECDFs of COMO-CMA-ES, the standard MO-CMA-ES-2 and a modified version of MO-CMA-ES-2 with `random` runs only, for all `bbob-biobj` test problems and for search space dimension 10.

5.4.5 Time

We investigate the dependence of the time taken by the benchmark on the budget factor. The budget of a benchmark experiment is equal to the budget factor times the search space dimension d. The following experiments were run on a server with 36 cores and 3.0 GhZ frequency.

We conduct simplified benchmark experiments, still relying on the COCO tools, for various budgets factors between 1000 and 90000. Each benchmark experiment runs MO-CMA-ES-2 on the first instance of the well-conditioned quadratic problem, **doublesphere** for search space dimensions 2, 3, 5, 10, 20 and 40. In Figure 5.7, we plot in log-log scale the time taken by the benchmark against the budget factor. We also fit a linear model to the evolution of the logarithm of the time taken by the benchmark with the logarithm of the budget factor from a budget factor of 10^4 and plot the associated line as a red dashed line. For large budget, the time (in seconds) evolves like 2.5×10^{-8} times the budget factor to the power 2.7.

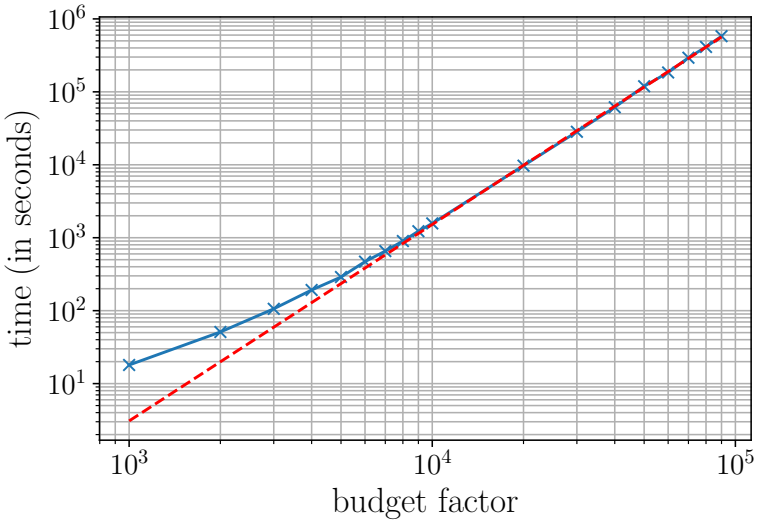


Figure 5.7: Dependence of the time taken by the simplified benchmarking on the function evaluations budget.

We also investigate the time per function evaluation for the small budget of $1000d$ using the same server. In this case, the benchmark is done for every instance of every problem. The time per function evaluation of the benchmark of MO-CMA-ES-2 is comprised between 180 and 220 microseconds depending on the search space dimension. The experiments mostly used one core. This is almost one order of magnitude larger than values reported for HMO-CMA-ES in [61], when run on one core with frequency 3.50GHz, which were comprised between 36 and 51 microseconds.

5.5 Convergence speed on the doublesphere problem

In this section, we empirically investigate the convergence speed of MO-CMA-ES-2 on the easiest problem of the `bbob-biobj` test suite, the **doublesphere** problem. Both objective functions of the **doublesphere** problem are so-called Sphere functions, i.e. functions of the form $F_i(X) = \|X - X_i^*\|_2^2 + F_i^*$ $(i = 1, 2)$, with $X_i^* \in \Omega$ and $F_i^* \in \mathbb{R}$. Sphere functions are unimodal, convex and $\mathcal{C}^\infty$, and their level sets are spheres. The Pareto front of the **doublesphere** problem is, up to an affine transformation, the **doublesphere** Pareto front, one of the benchmark Pareto fronts used in the previous chapters, see Figures 3.7 and 4.2.

We run MO-CMA-ES-2 on the first instance of the **doublesphere** problem, for various dimensions of the search space: 5, 10, 20, and 40. Two different versions of MO-CMA-ES-2 are run: the standard version, where the budget is equally distributed between `random` and `best_chv` runs and a modified version with `best_chv` runs only. Unless otherwise specified, we refer to the standard version. Due to the random sampling step of CMA-ES, the MO-CMA-ES-2 algorithm is not deterministic. We run MO-CMA-ES-2 five times for each search space dimension and for both the standard and the modified versions of the algorithm to get an idea of the performance variability.

The coordinates of the reference point are set to the largest objective values of interest of the **doublesphere** problem, which are given by COCO via the attribute `largest_fvalues_of_interest`. The initial step size σ_0 is set to 2, that is 10% of the search space size. The remaining of the settings are chosen as in Section 5.3.2.

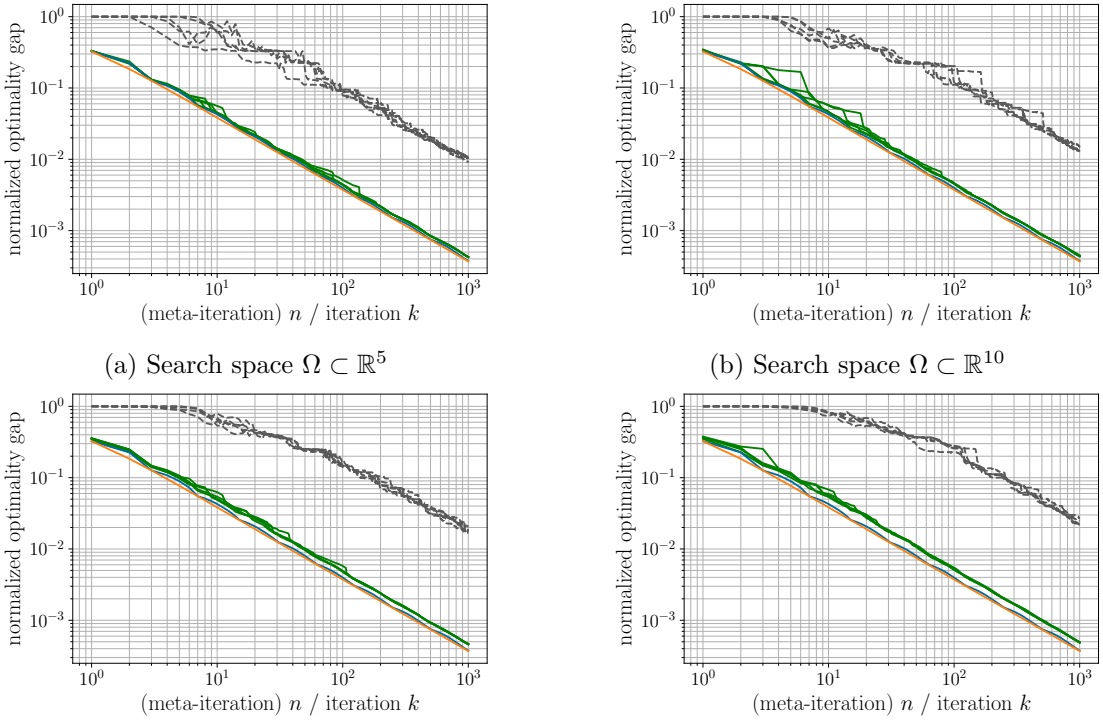
5.5.1 Dependence of the convergence speed on the number of meta-iterations

We investigate how close MO-CMA-ES-2 comes to achieving the same convergence speed with respect to meta-iterations as n-optimal distributions and HV-ISOOMOO coupled with a perfect single-objective solver. The coupling of the HV-ISOOMOO framework with a perfect single-objective solver is called *ideal HV-ISOOMOO* for short.

We define the *extended final means Pareto front approximation* as the non-dominated subset of the union between the final means Pareto front approximation and the image of the mean of the current CMA-ES run. Compared to the final means Pareto front approximation, which is only updated at the end of a meta-iteration, its extension allows to see the contribution of each CMA-ES iteration to the quality of the Pareto front approximation.

In Figure 5.8, we represent the normalized optimality gap of the extended FM-PFA of MO-CMA-ES-2 and ideal HV-ISOOMOO against meta-iterations as well as that of n-optimal distributions against n. While the convergence speed with respect to iterations is not the focus here, we also represent the normalized optimality gap of the extended FM-PFA of MO-CMA-ES-2 against iterations. The plotting is done in log-log scale. Each search space dimension has its figure.

We observe in Figure 5.8 that the normalized optimality gap of the extended FM-PFA of MO-CMA-ES-2 and ideal HV-ISOOMOO at meta-iteration n, as well as the



(a) Search space $\Omega \subset \mathbb{R}^5$

(b) Search space $\Omega \subset \mathbb{R}^{10}$

(c) Search space $\Omega \subset \mathbb{R}^{20}$

(d) Search space $\Omega \subset \mathbb{R}^{40}$

Figure 5.8: Convergence speed comparison between the standard MO-CMA-ES-2, the ideal HV-ISOOMOO and n-optimal distributions. We represent the normalized optimality gaps of the extended FM-PFA of five runs of the standard MO-CMA-ES-2 against meta-iterations — and iterations - -, of the ideal HV-ISOOMOO against meta-iterations —, and of n-optimal distributions against n — for search space dimensions 5, 10, 20 and 40.

normalized optimality gap of n-optimal distributions are close. On average, over the five MO-CMA-ES-2 runs and over meta-iterations, the normalized optimality gap of the extended FM-PFA of MO-CMA-ES-2 is 32%, 37%, 45% and 54% larger than that of the ideal HV-ISOOMOO for search space dimensions 5, 10, 20 and 40 respectively. From a perfect single-objective solver to a real one, less than a factor two has been lost.

After a warm-up phase, the curve of the normalized optimality gap against iterations becomes close to a straight line. But this curve slowly moves away from the other curves, which implies that the convergence speed is slower than $O(1/k)$, with k being the number of iterations.

Other observations can be made on these convergence graphs. At the very beginning, there is a short phase where the normalized optimality gap of the extended FM-PFA of MO-CMA-ES-2 is equal to 1. At this point, the mean of the first CMA-ES run has not yet dominated the reference point. This phase never lasts more than 10 iterations.

We can also observe a slight decrease of the convergence speed with respect to meta-iterations when the search space dimension grows. This phenomenon does not seem to have an obvious explanation. An hypothesis would be that this decrease in convergence speed is linked with a decrease of the quality of the solutions returned by the CMA-ES runs. It remains to be investigated whether a loss of precision actually occurs.

The normalized optimality gap of the extended FM-PFA of MO-CMA-ES after n meta-iterations is optimal and converges approximatively like $1/n$ multiplied by a constant on the available data. The MO-CMA-ES-2 succeeds in achieving a convergence speed with respect to meta-iterations similar to that of ideal HV-ISOOMOO. However, the convergence speed with respect to iterations and number of function evaluations proves to be slower than $O(1/k)$, with k being the iteration index. We investigate more precisely the convergence speed with respect to iterations in the next section, see Figure 5.9.

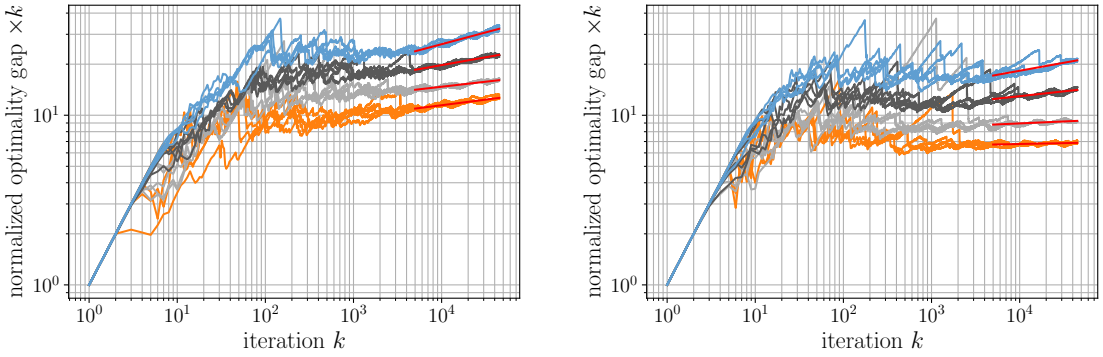
5.5.2 Dependence of the convergence speed on the number of iterations

The number of iterations is a more accurate measure of algorithmic cost than the number of meta-iterations. Indeed, it is proportional to the number of function evaluations. The number of function evaluations per iteration is equal to the number of sampled offspring, the population size. We use the default population size of CMA-ES, $4 + 3 \times \lfloor \ln(d) \rfloor$ [48], that only depends on the search space dimension d. For search space dimensions 5, 10, 20 and 40, the population size is equal to 8, 10, 12 and 15, respectively.

In Figure 5.9, we represent the normalized optimality gap of the extended FM-PFA multiplied by the iteration index k for both MO-CMA-ES-2 versions, against iterations for five algorithm runs. This representation highlights how the convergence speed relates to $1/k$. The curve having a horizontal asymptote would mean that the normalized optimality gap is equivalent to a constant times $1/k$. The oscillation of the curve around a horizontal line would mean that the normalized optimality gap evolves as $\Theta(1/k)$. A rising curve would mean that the convergence speed of the normalized optimality gap of the extended FM-PFA of MO-CMA-ES-2 is slower than $O(1/k)$. All search space dimensions are gathered in the same plot. The plotting is again done in log-log scale. On the left, we have the standard MO-CMA-ES-2. On the right, we have the `best_chv` runs only version.

The curves roughly oscillate around a line from a certain point onwards. We fit a linear model to the evolution of the logarithm of the mean over the five MO-CMA-ES-2 runs of the normalized optimality of the extended FM-PFA at iteration k as a function of the logarithm of k for $k = 5000$ to 45000. This is done for each MO-CMA-ES-2 version and each search space dimension. The corresponding affine functions are plotted as red lines.

First, we examine the Figure 5.9(a), which corresponds to the standard MO-CMA-ES-2. We see again the warm-up phase, which lasts for less than ten iterations. Afterwards, the curves keep on going up steeply for a period that lasts until the hundred or so iterations. During the last phase, the curves oscillate around a line that is less steep but still not horizontal. The oscillations become more regular and less wide-ranging after



(a) Standard MO-CMA-ES-2

(b) MO-CMA-ES-2 with only `best_chv` initialization

Figure 5.9: The normalized optimality gap of the extended FM-PFA of MO-CMA-ES-2 times the iteration index k against iterations for search space dimensions of the search space 5 —, 10 —, 20 — and 40 —.

the 5000th iteration or so. This confirms that the convergence speed of the extended FM-PFA of the standard MO-CMA-ES-2 towards the Pareto front is slower than $O(1/k)$ with k being the iteration index. We also observe that the number of iterations needed to reach a fixed accuracy increases with the search space dimension.

The coefficients of determination `R2` of the linear models are large. They start at 0.88 for search space dimension 5 and increases as the search space dimension grows, up to 0.99 in dimension 40. The linear models translate into models of the evolution of the normalized optimality gap of the extended FM-PFA as a function of k of the form C/k^α with $0 < \alpha < 1$. These models are $6.24/k^{0.93}$, $8.53/k^{0.94}$, $8.43/k^{0.91}$ and $7.3/k^{0.86}$ for search space dimensions 5, 10, 20 and 40, respectively. There is no trend in the evolution of either the slope or the multiplicative factor of the model with the search space dimension. The oscillations and the noise make it difficult to ensure that the convergence speed is really $\Theta(k^\alpha)$ with $\alpha < 1$.

The convergence speed of the normalized optimality gap of the extended final means Pareto front approximation of the standard MO-CMA-ES-2 is slower than $O(1/k)$. The evidence suggests that the normalized optimality gap speed of convergence resembles $\Theta(1/k^\alpha)$ with $\alpha < 1$. Such convergence speeds are asymptotic, and cannot be rigorously established using these empirical data. An obvious explanation is that as the precision required for the CMA-ES runs increases, `random` runs need more iterations to reach it. We do not observe a clear increase of the number of iterations per `best_chv` run throughout the MO-CMA-ES-2 runs.

We now examine the right-hand plot of Figure 5.9, which corresponds to a version of MO-CMA-ES-2 with `best_chv` runs only. As the Pareto front of the **doublesphere** problem is convex, the `best_chv` runs are able to solve the optimization problem without the support of the `random` runs for exploration.

We can see a warm-up phase followed by a steep increase at the beginning similar to

what we observe for standard MO-CMA-ES-2. The steep increase stops a bit earlier, after roughly 50 iterations. After that, the iteration index times the normalized optimality gap no longer increases significantly. It seems to be decreasing for a while, though the trend is unclear. There are more variations between runs than for the standard version of MO-CMA-ES-2. In the end, the curves are almost flat for search space dimensions 5 and 10 but are visibly increasing for dimensions 20 and 40. At the last stage, from 5000 to 45000 iterations, we observe that the normalized optimality gap of `best_chv` MO-CMA-ES-2 is lower than the one of the standard MO-CMA-ES-2.

The coefficients of determination `R2` of the linear models are 0.25, 0.49, 0.77 and 0.94 for search space dimensions 5, 10, 20 and 40 respectively. The coefficients of the linear models associated to dimensions 5, 10 and 20 should be considered with more caution, especially for dimension 5. However, these small `R2` values do not invalidate the model pertinence. The coefficient `R2` measures the proportion of the variation in y-values predicted by the model. For search space dimension 5, we observe that the line around which the curves oscillate is almost horizontal and that oscillations contribute more than the rise of this line to the variations in y-values. This phenomenon also occurs for search space dimension 10 and 20, but to a lesser degree. This observation explains the small `R2` values. The models of the evolution of the normalized optimality gap of the extended FM-PFA of the `best_chv` MO-CMA-ES-2 as a function of the number of iterations k are $6.17/k^{0.99}$, $7.24/k^{0.98}$, $7.68/k^{0.94}$ and $7.71/k^{0.91}$ for search space dimensions 5, 10, 20 and 40, respectively. These models all have a larger superscript coefficient than their equivalent for the standard MO-CMA-ES-2.

The `best_chv` only MO-CMA-ES-2 version performs better for intermediate function evaluations budget than the standard MO-CMA-ES-2 on the **doublesphere** problem. However, it is not enough to achieve a $O(1/k)$ convergence speed with respect to the number of iterations. The algorithm comes very close to a $O(1/k)$ convergence speed for search space dimension 5, and moves away from it for larger search space dimensions.

5.6 Conclusion

In this chapter, we introduce an instantiation of HV-ISOOMOO based on the single-objective algorithm CMA-ES and the UHVI$_r$ indicator: MO-CMA-ES-2. The CMA-ES runs are initialized in two ways that share the function evaluations budget evenly: a random initialization, `random`, that explores the search space from the ground up, and an initialization near promising regions, `best_chv`, that promotes efficiency.

We investigate the performance of MO-CMA-ES-2 using the COCO benchmarking platform up to $5 \times 10^4 d$ function evaluations. This benchmarking platform is designed for black-box optimization algorithms and relies on the hypervolume to measure the quality of sets of solutions.

The MO-CMA-ES-2 algorithm has been built as an anytime extension of COMO-CMA-ES. This goal has been successfully achieved. The MO-CMA-ES-2 shows overall similar performance as COMO-CMA-ES, but does not require to adapt any algorithm parameter to the budget. Its performance is also similar to that of HMO-CMA-ES,

a hybrid of various algorithms that had the best anytime performance on the COCO platform so far. We also compare the MO-CMA-ES-2 algorithm with the unbounded population variant of MO-CMA-ES (UP-MO-CMA-ES) combined with an exploration phase that consists in running simultaneous multiple MO-CMA-ES for $10^4 d$ function evaluations in total [59]. The MO-CMA-ES-2 performs significantly better during the exploration phase, before UP-MO-CMA-ES catches up. The later trend is unclear due to limited available data.

The MO-CMA-ES-2 performs well for a wide range of problems and budget with no parameter setting except for the initial step size. Thanks to the covariance matrix adaptation, it handles well ill-conditioned objective functions. It struggles to solve some problems involving multimodal objective functions, in particular the Rastrigin function. Adding a mechanism for increasing the population size of the CMA-ES run through the MO-CMA-ES-2 run could be of help on such problems. The main flaw of MO-CMA-ES-2 appears to be its computation time. For large budget, the computation time of a simplified benchmark of MO-CMA-ES-2 evolves as the budget to the power of 2.7. It also appears to be almost one order of magnitude larger than that of HMO-CMA-ES for small budget.

The convergence speed of HV-ISOOMOO with respect to meta-iterations investigated in Chapter 4 proves to be an ideal that can be approached. The convergence speed of the extended FM-PFA of MO-CMA-ES-2 with respect to meta-iterations on the well-conditioned convex quadratic **doublesphere** problem is $\Theta(1/n)$. The optimality gap of MO-CMA-ES-2 after n meta-iterations is close to that of n-optimal distributions and HV-ISOOMOO after n meta-iterations. More precisely, its optimality gap is on average less than twice as large.

However, the convergence speed of the extended FM-PFA of MO-CMA-ES-2 with respect to the number of iterations k on the **doublesphere** problem is slower than $O(1/k)$. Only initializing the CMA-ES near promising regions using `best_chv` initialization improves the performance on this particular problem, but not enough to reach a $O(1/k)$ convergence speed. The MO-CMA-ES-2 algorithm displays convergence speed resembling $\Theta(1/k^\alpha)$ for α as large as 0.99 (for search space dimension 5 and `best_chv` runs only).

The archive Pareto front approximation, the set of all solutions evaluated so far, should converge faster to the Pareto front than the extended final means Pareto front approximation whose convergence speed is examined here. It remains to be investigated how close the convergence speed of the A-PFA with respect to the number of function evaluations l comes to $O(1/l)$. Empirical cumulative distributive functions are indicative of the convergence speed. Therefore, the algorithms with the potential to be closer to a convergence speed of $O(1/l)$ than MO-CMA-ES-2 are the algorithms that perform better than MO-CMA-ES-2 on the **doublesphere** problem when compared using the ECDFs provided by the COCO platform. One such algorithm is HMO-CMA-ES.

CHAPTER 6

Conclusion

6.1 Summary

The common thread of this Ph.D. thesis has been the following question:

> How fast can a bi-objective optimization algorithm converge to the entire Pareto front ?

We proved a theoretical upper bound on the convergence speed of any bi-objective algorithm with respect to the number of function evaluations and provided evidence in favor of the thesis that this upper bound is optimal.

In Chapter 3, we examined the convergence speed of n-optimal distributions, the sets of n solutions with the largest hypervolume. We proved that the optimality gap of n-optimal distributions evolves as $\Omega(1/n)$ for a wide class of Pareto fronts. This implies that the convergence speed with respect to the number of function evaluations l of any bi-objective optimization algorithm is also $\Omega(1/l)$. We also provide evidence that this kind of convergence speed is not restricted to large l- or n-values. We indeed observe empirically on multiple Pareto fronts an evolution of the optimality gap of n-optimal distributions like a constant times $1/n$ from $n = 10$.

In Chapter 4, we exhibited sequences of solutions $(v_n)_{n\in\mathbb{N}^*}$ such that the optimality gap of the set $S_n := \{v_i, i = 1, \ldots, n\}$ is empirically close to that of a n-optimal distribution for various Pareto fronts, including non-convex ones. We call the sequences $(v_n)_{n\in\mathbb{N}^*}$ and $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ *greedy sequences* and *greedy set sequences* respectively because these sequences are built by way of greedy maximization of the hypervolume. Among other theoretical results, we prove rigorously that the convergence speed of a greedy set sequence $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ is $\Theta(1/n)$ when the Pareto front is convex.

Approximations of greedy sequences can be computed by multi-objective optimization algorithms. We introduce an algorithm framework HV-ISOOMOO that builds a

Pareto front approximation by iteratively finding a solution that increases the hypervolume of the Pareto front approximation as much as possible when added. Such solution is found by solving a single-objective optimization problem. In the ideal case where the single-objective solver always provides global optima of single-objective subproblems, the sequence of solutions returned by the single-objective solver form a greedy sequence.

In Chapter 5, we design and implement a real HV-ISOOMOO algorithm: MO-CMA-ES-2. The convergence speed of the final means Pareto front approximation of MO-CMA-ES-2 with respect to the number of meta-iterations n is $\Theta(1/n)$. Furthermore, the optimality gap of MO-CMA-ES-2 after n meta-iterations is close to that of n-optimal distribution. However, the convergence speed of the extended final means Pareto front approximation of MO-CMA-ES-2 with respect to the number of function evaluations l is slower than $O(1/l)$. More precisely, the convergence speed resembles $\Theta(1/l^\alpha)$ with α as large as 0.99 for search space dimension 5 and with only the efficient `best_chv` initialization of CMA-ES.

We examined in this exposition contributions that provide some answers to the main question addressed in this manuscript. In Chapter 4, we prove two additional theoretical results on the convergence speed of greedy set sequences. In Chapter 5, we also compare the performance of MO-CMA-ES-2 with that of other bi-objective optimization algorithms on the COCO benchmarking platform. This algorithm showed state-of-the-art performance when benchmarked and succeeds in providing an anytime extension of the COMO-CMA-ES algorithm. A more comprehensive summary of the results is available in the introduction.

6.2 Perspectives

A straigthforward approach to question addressed in this manuscript is simply to investigate empirically the convergence speed of existing bi-objective optimization algorithms on various problems. We did that in Chapter 5 with MO-CMA-ES-2 on the **doublesphere** problem. The next step would be to investigate the convergence speed of state-of-the-art bi-objective optimization algorithms, starting with those based on the hypervolume indicator including HMO-CMA-ES, both on the **doublesphere** and on more difficult problems. There may already be bi-objective optimization algorithms that reach the $\Theta(1/l)$ optimal convergence rate with respect to the number of function evaluations l. Such a study would yield a better understanding of the scope of problems where a $\Theta(1/l^\alpha)$ convergence speed is reachable and with what α. Closer to what we did, one could also investigate the convergence speed of the optimality gap of the archive of MO-CMA-ES-2 with respect to function evaluations on the **doublesphere** problem.

We focus in this manuscript on a particular definition of the convergence speed that is based on the hypervolume indicator and the number of function evaluations. The convergence speed measured with the multiplicative ϵ-indicator also has a $O(1/l)$ upper bound [19]. We conjecture that for many other set-quality indicators used as a measure of progress, there is a sublinear upper-bound on the convergence speed with respect to the number of function evaluations. A generalization of our results could take the form

of a sublinear upper-bound that holds under some assumptions on the Pareto front and the set-quality indicator used in the convergence speed definition.

A major limitation of the scope of this manuscript is the focus on bi-objective optimization algorithms and problems. The MO-CMA-ES-2 algorithm has been developed as a bi-objective optimization algorithm. This is not the case of the HV-ISOOMOO framework which could be extended as a multi-objective algorithm framework without any modification. Many of our theoretical results may be generalizable to a number of objectives $m \geq 3$ by relying on more complex mathematical tools and analysis. We conjecture that the sublinear upper bounds on the convergence speed of n-optimal distributions derived in Chapter 3 generalize to $m \geq 3$ with the optimal upper bound depending on m. In the conclusion of Chapter 4, we describe the steps and missing bricks that would be involved in the proof of a generalization of the lower bounds on the convergence speed of ideal HV-ISOOMOO. We also discuss whether or not the other theoretical results are generalizable.

APPENDIX A

Normalized areas of the gap regions relative to an hypervolume maximizer

The goal of this section is to prove bounds on the normalized areas of the gap regions $\mathcal{G}^{u^*}_{\text{left}}$ and $\mathcal{G}^{u^*}_{\text{right}}$ relative to an hypervolume maximizer u^* (see the lefthand plot of Figure A.1) in the case of a bilipschitz Pareto front and of a reference point r dominating the nadir point. These bounds are stated in Lemma 12. The proof relies on the bounds on the normalized maximum hypervolume proven in Section 4.4.2 and the following lower and upper bounds on the relation between $\lambda(\mathcal{G}^{u^*}_{\text{left}})$ and $\lambda(\mathcal{G}^{u^*}_{\text{right}})$.

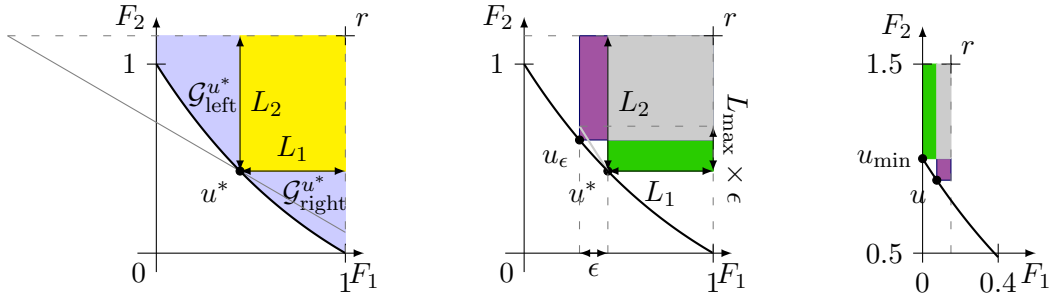


Figure A.1: Illustration of elements of the proofs of Proposition 14 in the case $r_1 \leq x_{\max}$ (left and middle) and of Lemma 13 (right). The Pareto front is described by $f(x) = \frac{e}{e-1} \times e^{-x} + 1 - \frac{e}{e-1}$ for $x \in [0, 1]$. Left: the gap regions $\mathcal{G}^{u^*}_{\text{left}}$ and $\mathcal{G}^{u^*}_{\text{right}}$ with a segment of slope $-L_{\min}$ passing through u^*. Middle: the hypervolume improvements $\text{HVI}_r(u^*, u_\epsilon)$ (▬) and $\text{HVI}_r(u_\epsilon, u^*)$ (▬) . Right: the hypervolume improvement $\text{HVI}_r(u_{\min}, u)$ (▬) and its counterpart $\text{HVI}_r(u, u_{\min})$ (▬) where u is a vector of the Pareto front which dominates r.

Proposition 14 *We assume that the Pareto front is described by a ($L_{\min}$,$L_{\max}$)-bilipschitz function f. Let u^* be a non-extreme vector of the Pareto front which maximizes the hypervolume with respect to a valid reference point r. If $r_1 \leq x_{\max}$, we have $\lambda(\mathcal{G}^{u^*}_{right}) \geq \frac{L^2_{\min}}{L^2_{\max}} \times \lambda(\mathcal{G}^{u^*}_{left})$. If $r_2 \leq f(x_{\min})$, we have $\lambda(\mathcal{G}^{u^*}_{left}) \geq \frac{L^2_{\min}}{L^2_{\max}} \times \lambda(\mathcal{G}^{u^*}_{right})$.*

Proof *We consider the case where $r_1 \leq x_{\max}$. Let x^* be the first coordinate of u^*. We denote $L_1 := r_1 - x^*$ and $L_2 := r_2 - f(x^*)$ the lengths of the sides of the rectangle $\mathcal{D}^r_{u^*}$. For all $x, y \in [x_{\min}, x_{\max}]$, we have $|f(x) - f(y)| \geq L_{\min} \times |x - y|$. Additionally, since $r_1 \leq x_{\max}$, the segment $[x^*, x^* + L_1]$ is included in $[x_{\min}, x_{\max}]$. As a consequence, the section of the Pareto front on the right of u^* dominates the segment between u^* and $u^* + L_1 \times (1, -L_{\min})$, see the lefthand plot of Figure A.1. Therefore, $\lambda(\mathcal{G}^{u^*}_{right})$ is larger than the area of the region of the objective space dominated by this segment, not dominated by u^* and dominating r, that is $\frac{1}{2} \times L_{\min} \times L_1^2$. For all $x, y \in [x_{\min}, x_{\max}]$, we also have $|f(x) - f(y)| \leq L_{\max} \times |x - y|$. Therefore, the part of the Pareto front on the left of u^* is dominated by the segment between u^* and $u^* + L_2 \times (-\frac{1}{L_{\min}}, 1)$, and $\lambda(\mathcal{G}^{u^*}_{left})$ is smaller than $\frac{1}{2} \times \frac{1}{L_{\min}} \times L_2^2$. We have yet to prove a lower bound on $\frac{L_1}{L_2}$. The vector u^* being different from $u_{\min}$, for $\epsilon > 0$ small enough, the vector $u_\epsilon := (x^* - \epsilon, f(x^* - \epsilon))$ belongs to the Pareto front. As we can see in the middle plot of Figure A.1, $HVI_r(u^*, u_\epsilon)$ is smaller than $L_1 \times L_{\max} \times \epsilon$ and $HVI_r(u_\epsilon, u^*)$ is larger than $\epsilon \times (L_2 - \epsilon \times L_{\max})$. Additionally, u^* being an hypervolume maximizer, $HVI_r(u^*, u_\epsilon)$ is larger than $HVI_r(u_\epsilon, u^*)$, and thus $L_1 \times L_{\max} \geq L_2 - \epsilon \times L_{\max}$ for all $\epsilon > 0$. Taking the limit of this inequality when $\epsilon \to 0$, we obtain that $L_1 \times L_{\max} \geq L_2$. Combining the bounds on $\lambda(\mathcal{G}^{u^*}_{left})$ and $\lambda(\mathcal{G}^{u^*}_{right})$ with the lower-bound on $\frac{L_1}{L_2}$, we obtain the desired lower bound on $\lambda(\mathcal{G}^{u^*}_{right})$. We can obtain the symmetric inequality when $r_2 \geq f(x_{\min})$ by following the same approach.*

In particular, when f is bilipschitz and r dominates the nadir point, both bounds hold. We now prove the desired bounds on the normalized area of the gap regions $\mathcal{G}^{u^*}_{\text{left}}$ and $\mathcal{G}^{u^*}_{\text{right}}$.

Lemma 12 *Let u^* be a vector which maximizes the hypervolume with respect to a valid reference point r. If the Pareto front is described by a ($L_{\min}$,$L_{\max}$)-bilipschitz function f and the reference point r dominates the nadir point, both $\lambda(\mathcal{G}^{u^*}_{left})$ and $\lambda(\mathcal{G}^{u^*}_{right})$ are between $(1 - \frac{1}{2} \times \frac{L_{\max}}{L_{\min}})/(1 + \frac{L^2_{\max}}{L^2_{\min}})$ and $(1 - \frac{1}{2} \times \frac{L_{\min}}{L_{\max}})/(1 + \frac{L^2_{\min}}{L^2_{\max}})$.*

Proof *Let note arbitrarily $\mathcal{G}_1$ and $\mathcal{G}_2$ the two gap regions of the set $S = \{u^*\}$. By Proposition 14, $\lambda(\mathcal{G}_2)$ is between $\frac{L^2_{\min}}{L^2_{\max}} \times \lambda(\mathcal{G}_1)$ and $\frac{L^2_{\max}}{L^2_{\min}} \times \lambda(\mathcal{G}_1)$. Additionally, by Propositions 9 and 10, the normalized maximum hypervolume $\max_{u \in PF_f} HV_r(u)$ over $HV_r(PF_f)$ is between $\frac{1}{2} \times \frac{L_{\min}}{L_{\max}}$ and $\frac{1}{2} \times \frac{L_{\max}}{L_{\min}}$. These bounds can be transformed into bounds on $HV_r(PF_f) - \max_{u \in PF_f} HV_r(u)$, that is $\lambda(\mathcal{G}_1) + \lambda(\mathcal{G}_2)$. As a consequence, $\lambda(\mathcal{G}_1)$ is between $(1 - \frac{1}{2} \times \frac{L_{\max}}{L_{\min}}) \times HV_r(PF_f) - \frac{L^2_{\max}}{L^2_{\min}} \times \lambda(\mathcal{G}_1)$ and $(1 - \frac{1}{2} \times \frac{L_{\min}}{L_{\max}}) \times HV_r(PF_f) - \frac{L^2_{\min}}{L^2_{\max}} \lambda(\mathcal{G}_1)$. Moving all the $\lambda(\mathcal{G}_1)$ terms on the same side and re-normalizing this side, we obtain the desired bounds for $\mathcal{G}_1$, which can be chosen to be either $\mathcal{G}^{u^*}_{left}$ or $\mathcal{G}^{u^*}_{right}$.*

APPENDIX B

The nadir point is dominated by all the r_i^n corresponding to non-empty gap regions for n large

We show in this section that for bilipschitz Pareto fronts, the nadir point is dominated by all the local nadir points r_i^n corresponding to non-empty gap regions, for n large enough. This result is stated in Proposition 15 and used in Section 4.5.2. It is equivalent to prove that the extreme vectors which dominate the reference point belong to the greedy set for n large enough.

First, we prove in the next proposition that if $r_1 > x_{\max}$ (resp. $r_2 > f(x_{\min})$), then for r_2 (resp. r_1) close enough to $f(x_{\max})$ (resp. $x_{\min}$) the extreme vector $u_{\max}$ (resp. $u_{\min}$) is the only hypervolume maximizer, see the righthand plot of Figure A.1. There are similar statements in [20] for the set of μ points maximizing the hypervolume, but they only apply to $\mu \geq 2$.

Lemma 13 *We assume that the Pareto front is described by a function f which is ($L_{\min}$,$L_{\max}$)-bilipschitz and that the reference point r is valid. If $r_1 > x_{\max}$ and $f(x_{\max}) < r_2 < f(x_{\max}) + L_{\min} \times (r_1 - x_{\max})$, the right extreme of the Pareto front $u_{\max}$ is the only maximizer of $HV_r(.)$. Additionally, if $r_2 > f(x_{\min})$ and $x_{\min} < r_1 < x_{\min} + \frac{r_2 - f(x_{\min})}{L_{\max}}$, the vector $u_{\min} = (x_{\min}, f(x_{\min}))$ is the only maximizer of $HV_r(.)$.*

Proof *This proof is illustrated in the righthand plot of Figure A.1. Let r be a reference point such that $r_2 > f(x_{\min})$ and $x_{\min} < r_1 < x_{\min} + \frac{r_2 - f(x_{\min})}{L_{\max}}$. Let $u = (x, f(x)) \neq u_{\min}$ be a vector of the Pareto front which dominates r. The hypervolume improvement of $u_{\min}$ to $\{u\}$ is $(r_2 - f(x_{\min})) \times (x - x_{\min})$. The hypervolume improvement of u to $\{u_{\min}\}$ is $(f(x_{\min}) - f(x)) \times (r_1 - x)$, which is smaller than $L_{\max} \times (x - x_{\min}) \times (r_1 - x_{\min})$ since u dominates r and f is ($L_{\min}$,$L_{\max}$)-bilipschitz. Since we assume that $L_{\max} \times (r_1 - x_{\min}) < r_2 - f(x_{\min})$, the upper bound on $HVI_r(u, u_{\min})$ is strictly smaller than $HVI_r(u_{\min}, u)$. Therefore, the hypervolume of $u_{\min}$ is strictly larger than the one of u. We conclude that*

$u_{\min}$ is the unique hypervolume maximizer. The symmetric result can be obtained with the same approach.

It is left to prove that when $r_1 > x_{\max}$ (resp. $r_2 > f(x_{\min})$), the second coordinate of r_{n+1}^n (resp. the first coordinate of r_0^n) indeed converge to $f(x_{\max})$ (resp. $x_{\min}$). It is a straightforward consequence of Lemma 8. Therefore, we are able to conclude.

Proposition 15 *We assume that the Pareto front is described by a bilipschitz function f. Let $(\mathcal{S}_n)_{n\in\mathbb{N}^*}$ be a greedy set sequence relative to a valid reference point r. For n large enough, every local nadir point r_i^n corresponding to a non-empty gap region $\mathcal{G}_{\mathcal{S}_n,i}^r$ dominates the nadir point.*

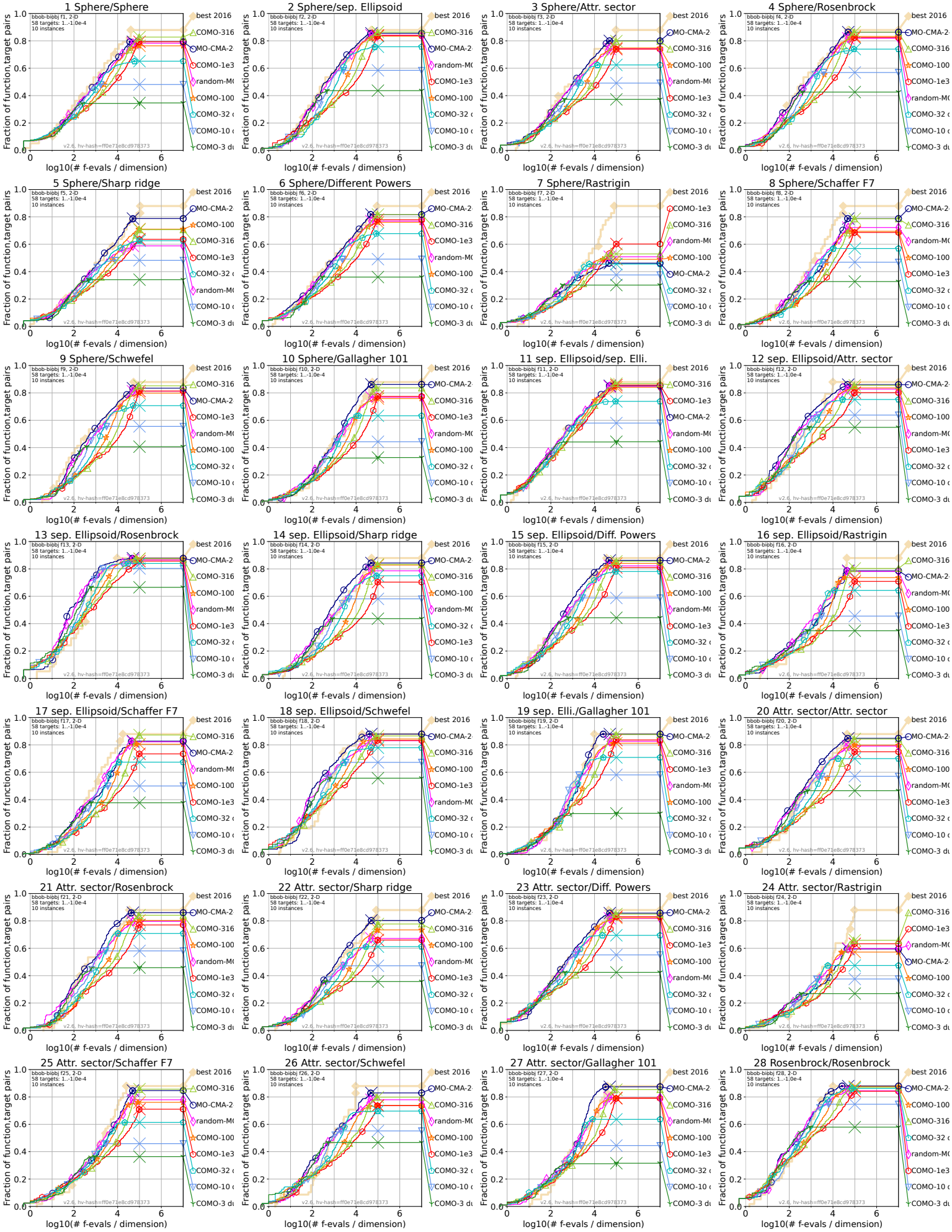
Proof *By Lemma 8, $w_{n,r}^n$ converges to $x_{\max}$, and thus the right extreme local nadir point $r_{n+1}^n := (r_1, f(w_{n,r}^n))$ converges to $(r_1, f(x_{\max}))$ by continuity of f. Therefore, if r_1 is strictly larger than $x_{\max}$, then there exists N such that for all $n \geq N$, r_{n+1}^n verifies the assumptions on the reference point of Lemma 13 which guarantee that $u_{\max}$ is the unique maximizer of $HV_r(.)$ over the right extreme gap region $\mathcal{G}_{\mathcal{S}_n,n+1}^r$. Let assume that $u_{\max}$ does not belong to $\mathcal{S}_n$. Then, $w_{N,r}^N \neq x_{\max}$, and since $w_{n,r}^n$ converges to $x_{\max}$, the left extreme gap region $\mathcal{G}_{\mathcal{S}_n,i}^r$ is necessarily filled at some later iteration. When the right extreme gap region is filled, $u_{\max}$, the unique minimizer of $HV_r(.)$ over this gap region, is added to the greedy set. To summarize, if $r_1 > x_{\max}$, then for n large enough $\mathcal{S}_n$ contains $u_{\max}$, and thus the right extreme gap region is empty. We can prove with the same approach that for $r_2 > f(x_{\min})$, $\mathcal{S}_n$ contains $u_{\min}$ for n large enough.*

At any iteration, the non-extreme local nadir points dominate the nadir point. Additionally, we proved that either $r_1 < x_{\max}$ (resp. $r_2 < f(x_{\min})$), and thus the left (resp. right) extreme local nadir point dominates the nadir point or for n large enough, the left (resp. right) extreme gap region is empty.

APPENDIX C

Additional ECDFs

We provide additional ECDFs of the standard MO-CMA-ES-2, the `random` only version of MO-CMA-ES-2, various COMO-CMA-ES, UP-MO-CMA-ES and HMO-CMA-ES. These are the equivalent of Figure 5.6 and Figure 5.4, that were for search space dimension 10, for the following search space dimensions: 2, 3, 5 and 20.


1 Sphere/Sphere
2 Sphere/sep. Ellipsoid
3 Sphere/Attr. sector
4 Sphere/Rosenbrock
5 Sphere/Sharp ridge
6 Sphere/Different Powers
7 Sphere/Rastrigin
8 Sphere/Schaffer F7
9 Sphere/Schwefel
10 Sphere/Gallagher 101
11 sep. Ellipsoid/sep. Elli.
12 sep. Ellipsoid/Attr. sector
13 sep. Ellipsoid/Rosenbrock
14 sep. Ellipsoid/Sharp ridge
15 sep. Ellipsoid/Diff. Powers
16 sep. Ellipsoid/Rastrigin
17 sep. Ellipsoid/Schaffer F7
18 sep. Ellipsoid/Schwefel
19 sep. Elli./Gallagher 101
20 Attr. sector/Attr. sector
21 Attr. sector/Rosenbrock
22 Attr. sector/Sharp ridge
23 Attr. sector/Diff. Powers
24 Attr. sector/Rastrigin
25 Attr. sector/Schaffer F7
26 Attr. sector/Schwefel
27 Attr. sector/Gallagher 101
28 Rosenbrock/Rosenbrock
Fraction of function,target pairs
log10(# f-evals / dimension)
best 2016
COMO-316
MO-CMA-2
COMO-1e3
random-M(
COMO-100
COMO-32 (
COMO-10 (
COMO-3 (

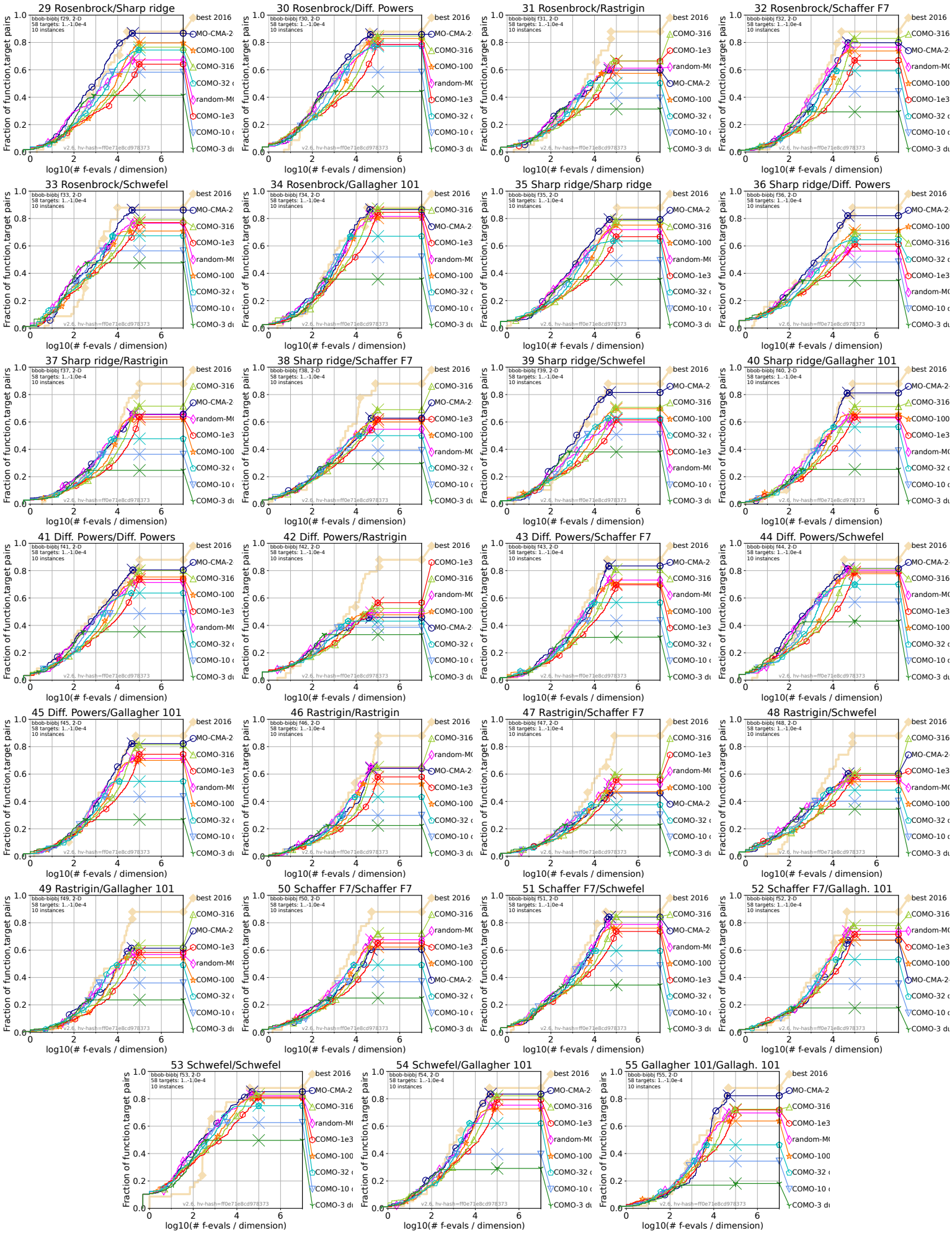
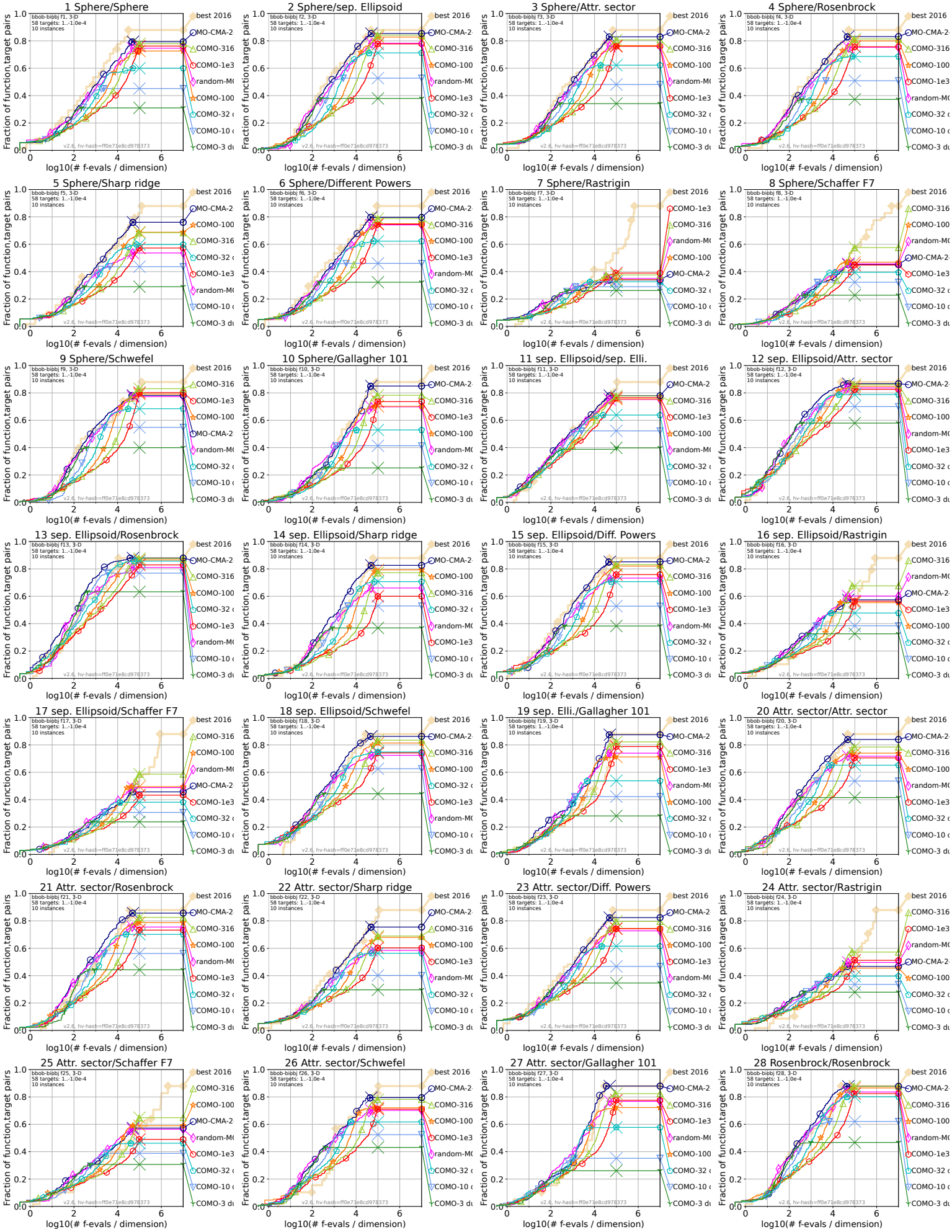


Figure C.1: ECDFs of COMO-CMA-ES, the standard MO-CMA-ES-2 and a modified version of MO-CMA-ES-2 with `random` runs only, for all `bbob-biobj` test problems and for search space dimension 2.


1 Sphere/Sphere
2 Sphere/sep. Ellipsoid
3 Sphere/Attr. sector
4 Sphere/Rosenbrock
5 Sphere/Sharp ridge
6 Sphere/Different Powers
7 Sphere/Rastrigin
8 Sphere/Schaffer F7
9 Sphere/Schwefel
10 Sphere/Gallagher 101
11 sep. Ellipsoid/sep. Elli.
12 sep. Ellipsoid/Attr. sector
13 sep. Ellipsoid/Rosenbrock
14 sep. Ellipsoid/Sharp ridge
15 sep. Ellipsoid/Diff. Powers
16 sep. Ellipsoid/Rastrigin
17 sep. Ellipsoid/Schaffer F7
18 sep. Ellipsoid/Schwefel
19 sep. Elli./Gallagher 101
20 Attr. sector/Attr. sector
21 Attr. sector/Rosenbrock
22 Attr. sector/Sharp ridge
23 Attr. sector/Diff. Powers
24 Attr. sector/Rastrigin
25 Attr. sector/Schaffer F7
26 Attr. sector/Schwefel
27 Attr. sector/Gallagher 101
28 Rosenbrock/Rosenbrock
Fraction of function,target pairs
log10(# f-evals / dimension)
best 2016
MO-CMA-2
COMO-316
COMO-100
COMO-1e3
random-MO
COMO-32
COMO-10
COMO-3

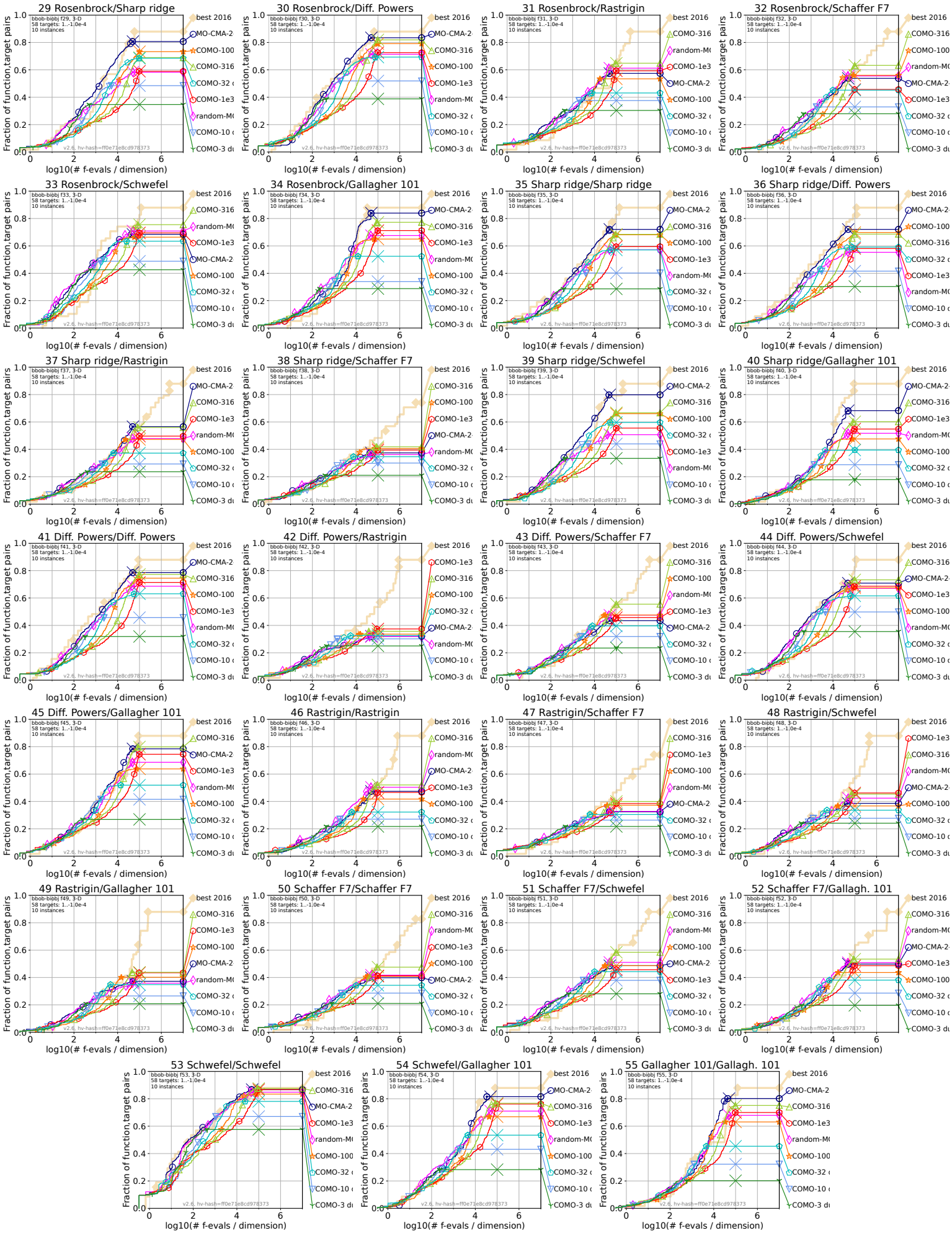
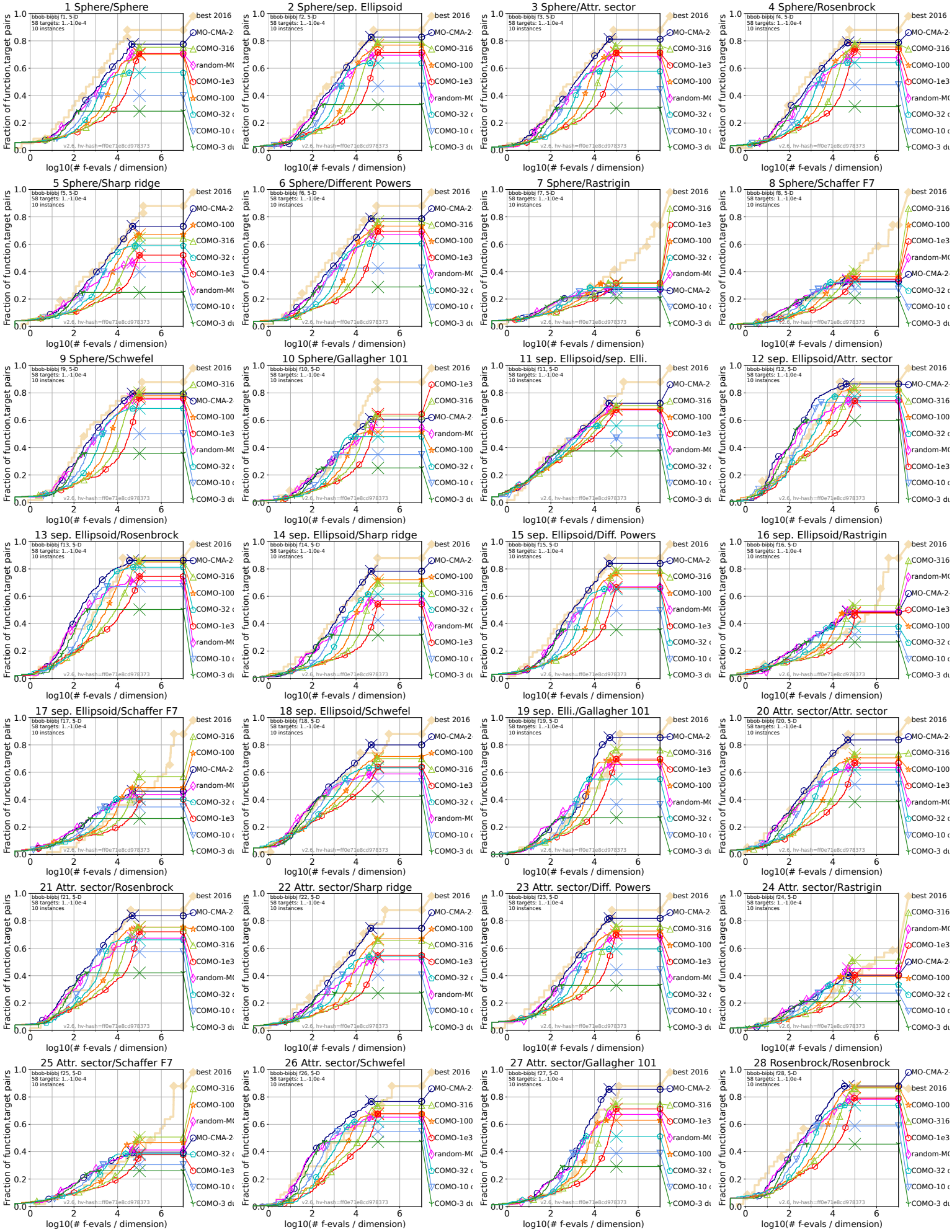


Figure C.2: ECDFs of COMO-CMA-ES, the standard MO-CMA-ES-2 and a modified version of MO-CMA-ES-2 with `random` runs only, for all `bbob-biobj` test problems and for search space dimension 3.


1 Sphere/Sphere
2 Sphere/sep. Ellipsoid
3 Sphere/Attr. sector
4 Sphere/Rosenbrock
5 Sphere/Sharp ridge
6 Sphere/Different Powers
7 Sphere/Rastrigin
8 Sphere/Schaffer F7
9 Sphere/Schwefel
10 Sphere/Gallagher 101
11 sep. Ellipsoid/sep. Elli.
12 sep. Ellipsoid/Attr. sector
13 sep. Ellipsoid/Rosenbrock
14 sep. Ellipsoid/Sharp ridge
15 sep. Ellipsoid/Diff. Powers
16 sep. Ellipsoid/Rastrigin
17 sep. Ellipsoid/Schaffer F7
18 sep. Ellipsoid/Schwefel
19 sep. Elli./Gallagher 101
20 Attr. sector/Attr. sector
21 Attr. sector/Rosenbrock
22 Attr. sector/Sharp ridge
23 Attr. sector/Diff. Powers
24 Attr. sector/Rastrigin
25 Attr. sector/Schaffer F7
26 Attr. sector/Schwefel
27 Attr. sector/Gallagher 101
28 Rosenbrock/Rosenbrock
Fraction of function,target pairs
log10(# f-evals / dimension)
best 2016
MO-CMA-2
COMO-316
COMO-100
COMO-1e3
random-M
COMO-32
COMO-10
COMO-3

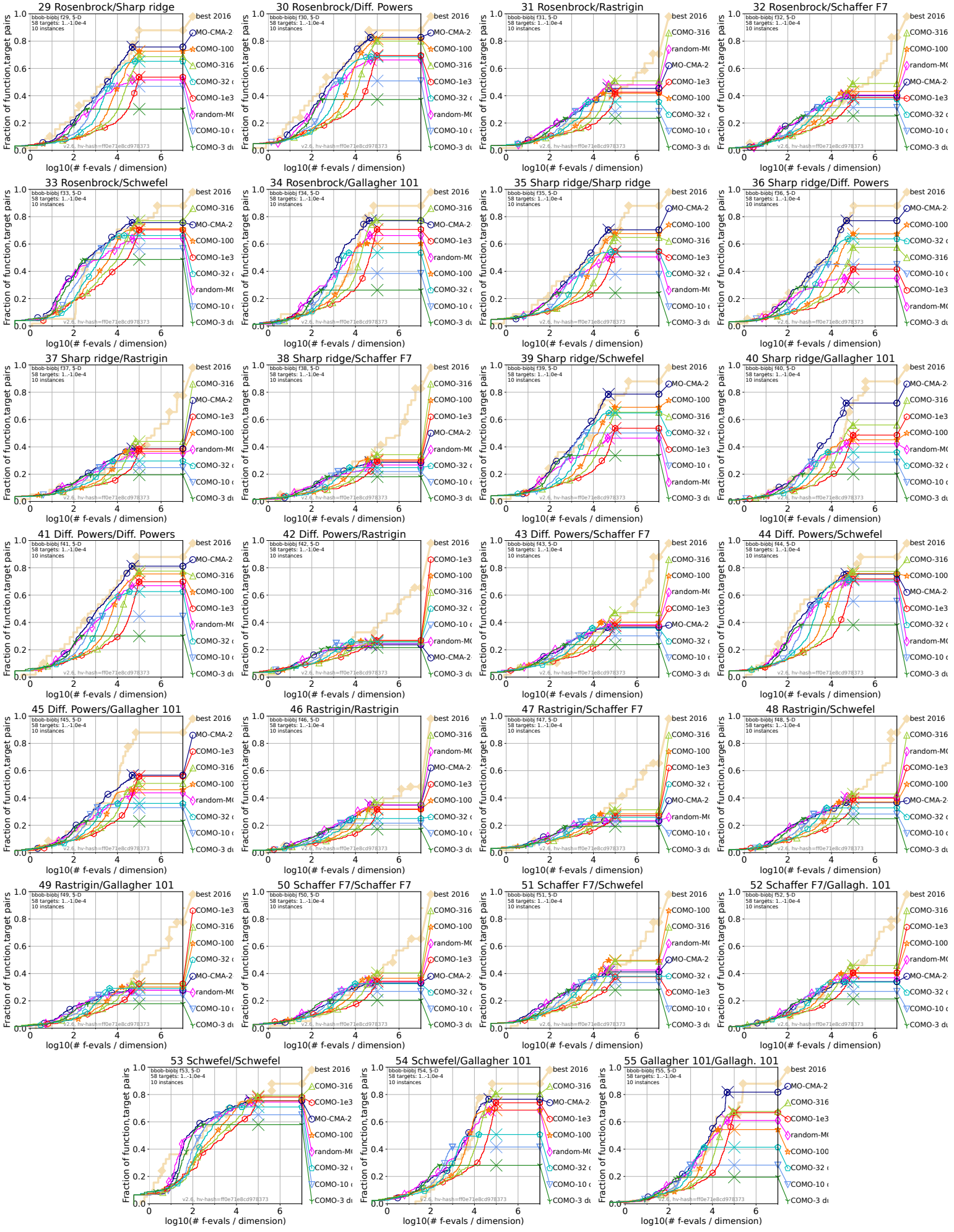
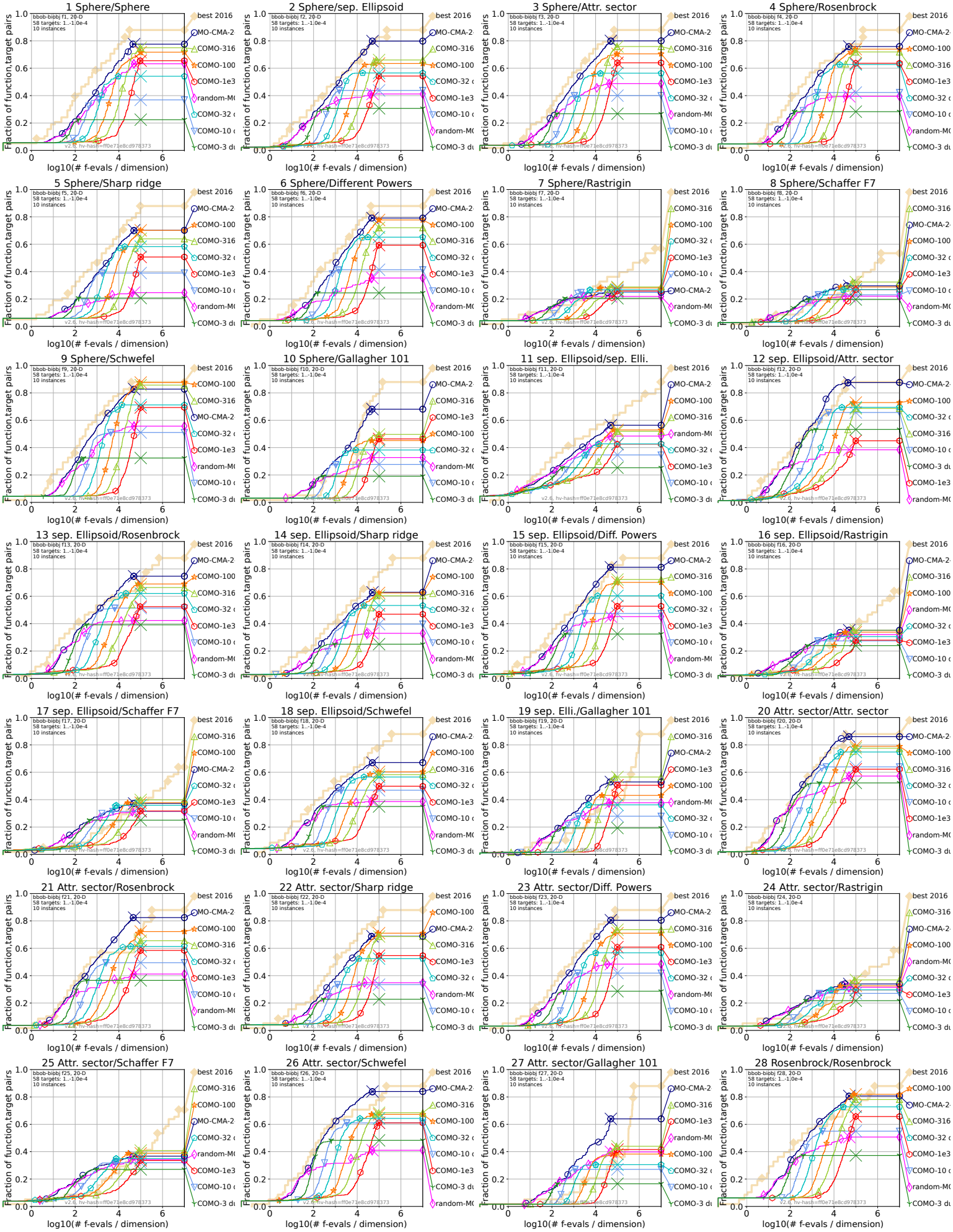


Figure C.3: ECDFs of COMO-CMA-ES, the standard MO-CMA-ES-2 and a modified version of MO-CMA-ES-2 with `random` runs only, for all `bbob-biobj` test problems and for search space dimension 5.


1 Sphere/Sphere
2 Sphere/sep. Ellipsoid
3 Sphere/Attr. sector
4 Sphere/Rosenbrock
5 Sphere/Sharp ridge
6 Sphere/Different Powers
7 Sphere/Rastrigin
8 Sphere/Schaffer F7
9 Sphere/Schwefel
10 Sphere/Gallagher 101
11 sep. Ellipsoid/sep. Elli.
12 sep. Ellipsoid/Attr. sector
13 sep. Ellipsoid/Rosenbrock
14 sep. Ellipsoid/Sharp ridge
15 sep. Ellipsoid/Diff. Powers
16 sep. Ellipsoid/Rastrigin
17 sep. Ellipsoid/Schaffer F7
18 sep. Ellipsoid/Schwefel
19 sep. Elli./Gallagher 101
20 Attr. sector/Attr. sector
21 Attr. sector/Rosenbrock
22 Attr. sector/Sharp ridge
23 Attr. sector/Diff. Powers
24 Attr. sector/Rastrigin
25 Attr. sector/Schaffer F7
26 Attr. sector/Schwefel
27 Attr. sector/Gallagher 101
28 Rosenbrock/Rosenbrock
Fraction of function,target pairs
log10(# f-evals / dimension)
best 2016
MO-CMA-2
COMO-316
COMO-100
COMO-1e3
COMO-32
COMO-10
COMO-3
random-M

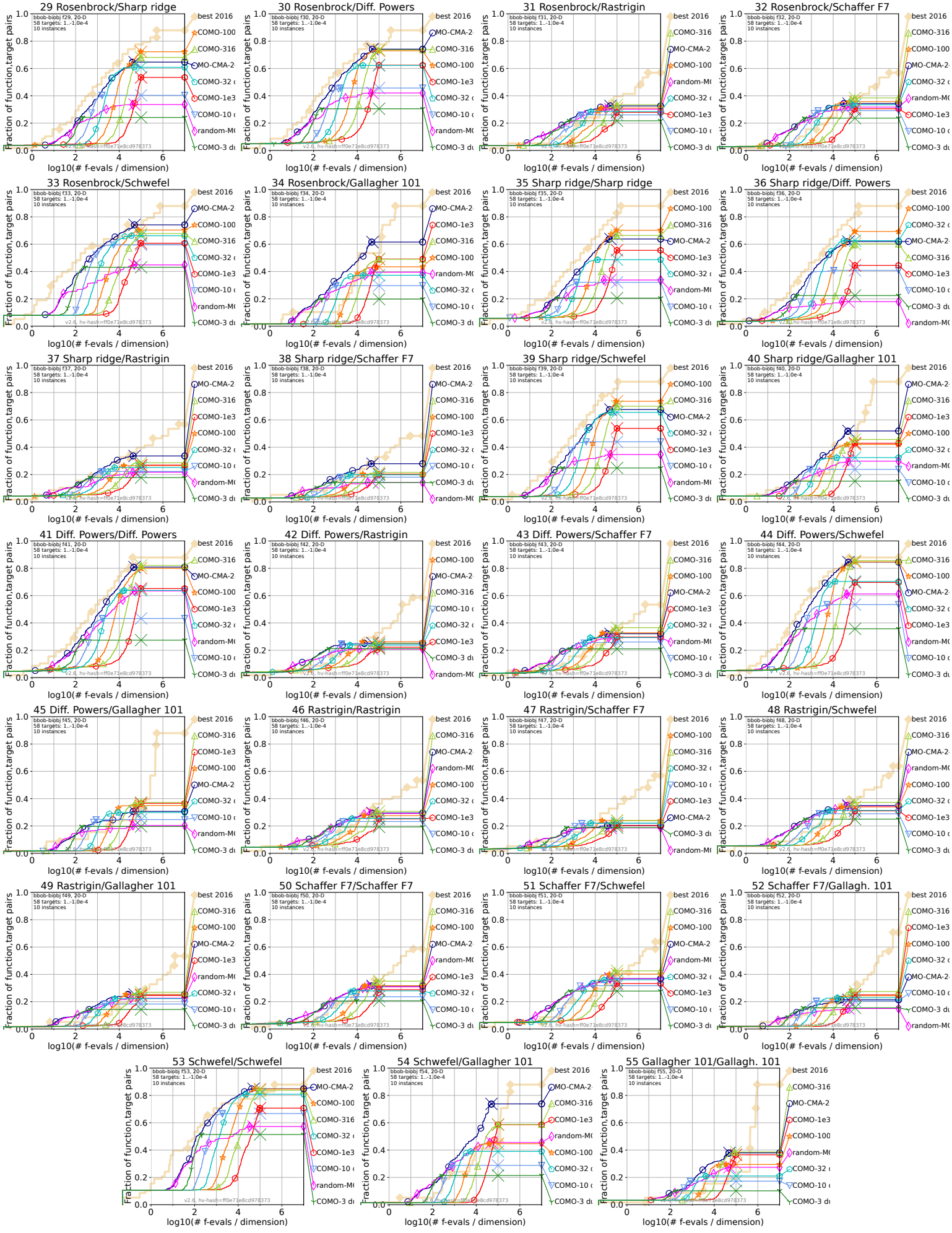
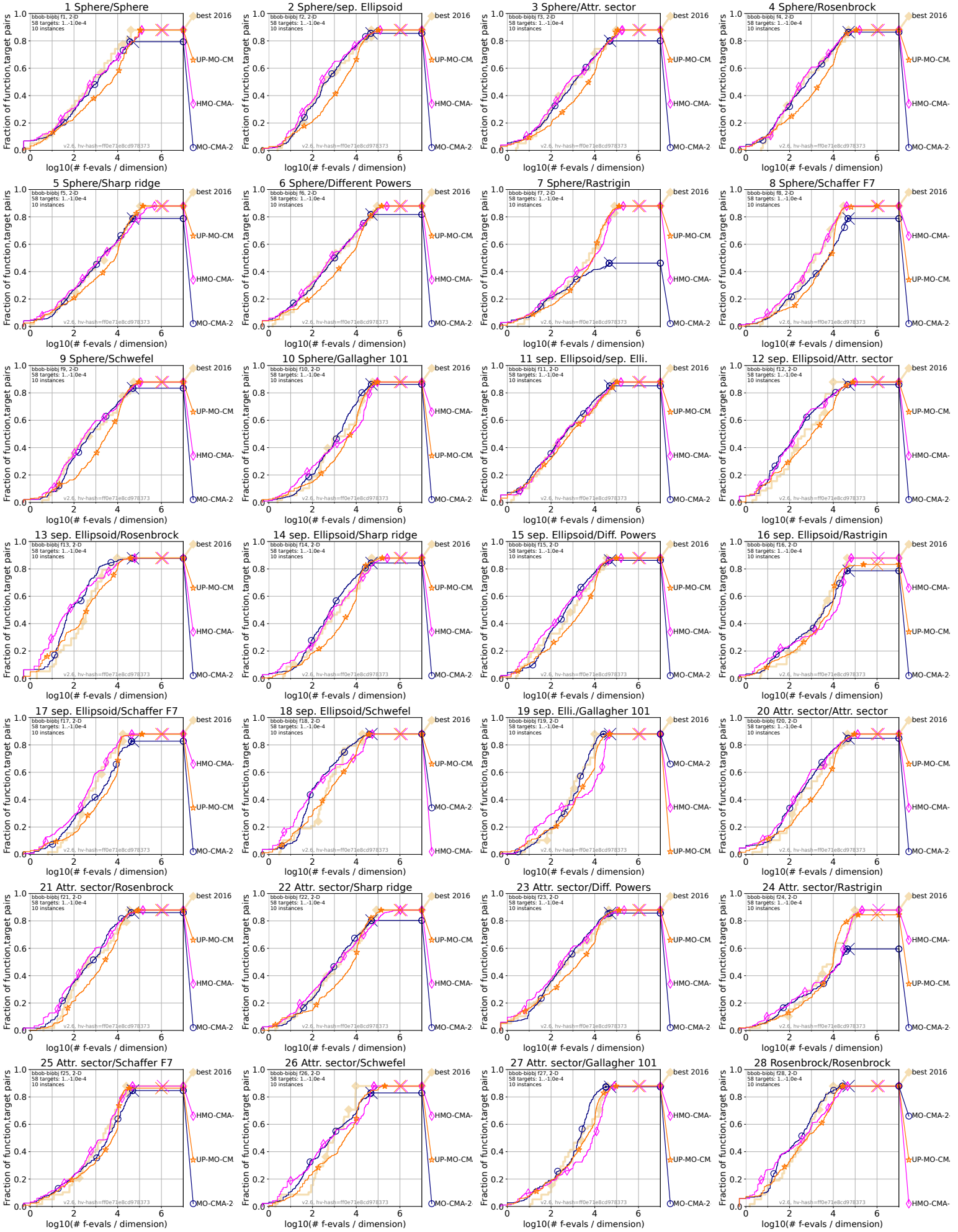


Figure C.4: ECDFs of COMO-CMA-ES, the standard MO-CMA-ES-2 and a modified version of MO-CMA-ES-2 with `random` runs only, for all `bbob-biobj` test problems and for search space dimension 20.


1 Sphere/Sphere
2 Sphere/sep. Ellipsoid
3 Sphere/Attr. sector
4 Sphere/Rosenbrock
5 Sphere/Sharp ridge
6 Sphere/Different Powers
7 Sphere/Rastrigin
8 Sphere/Schaffer F7
9 Sphere/Schwefel
10 Sphere/Gallagher 101
11 sep. Ellipsoid/sep. Elli.
12 sep. Ellipsoid/Attr. sector
13 sep. Ellipsoid/Rosenbrock
14 sep. Ellipsoid/Sharp ridge
15 sep. Ellipsoid/Diff. Powers
16 sep. Ellipsoid/Rastrigin
17 sep. Ellipsoid/Schaffer F7
18 sep. Ellipsoid/Schwefel
19 sep. Elli./Gallagher 101
20 Attr. sector/Attr. sector
21 Attr. sector/Rosenbrock
22 Attr. sector/Sharp ridge
23 Attr. sector/Diff. Powers
24 Attr. sector/Rastrigin
25 Attr. sector/Schaffer F7
26 Attr. sector/Schwefel
27 Attr. sector/Gallagher 101
28 Rosenbrock/Rosenbrock
Fraction of function,target pairs
log10(# f-evals / dimension)
best 2016
UP-MO-CMA-
HMO-CMA-
MO-CMA-2

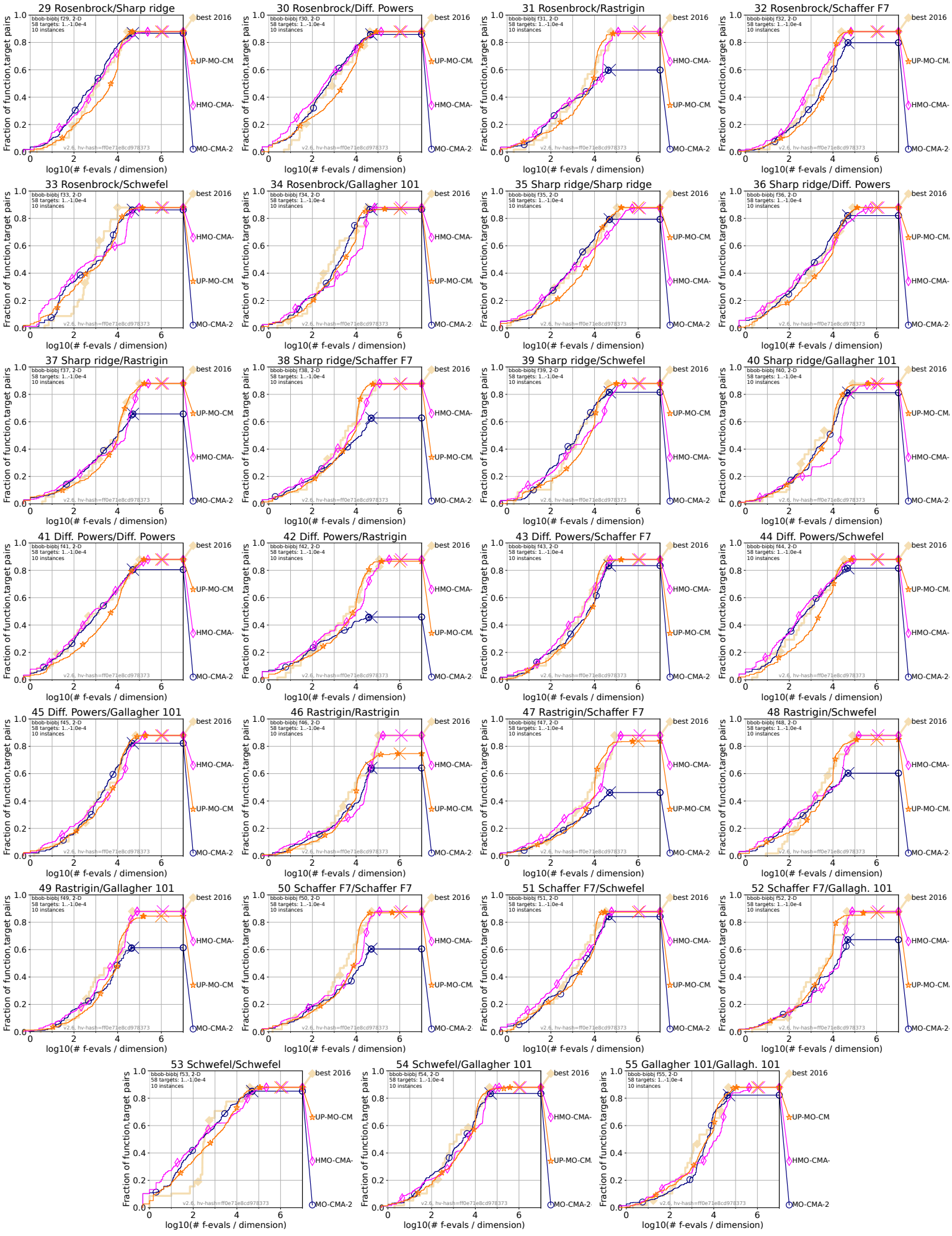
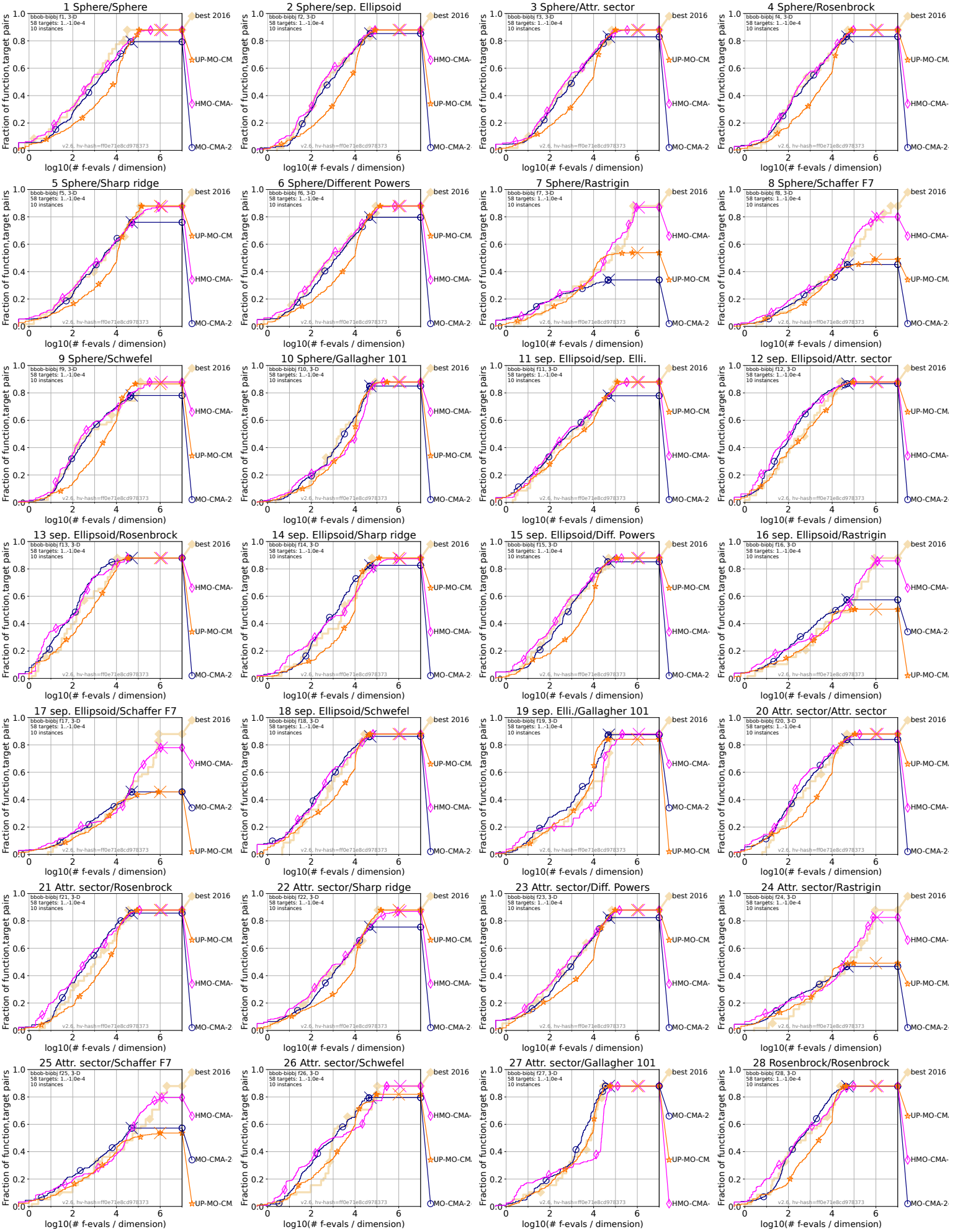


Figure C.5: ECDFs of COMO-CMA-ES, the standard MO-CMA-ES-2 and a modified version of MO-CMA-ES-2 with `random` runs only, for all `bbob-biobj` test problems and for search space dimension 2.


1 Sphere/Sphere
2 Sphere/sep. Ellipsoid
3 Sphere/Attr. sector
4 Sphere/Rosenbrock
5 Sphere/Sharp ridge
6 Sphere/Different Powers
7 Sphere/Rastrigin
8 Sphere/Schaffer F7
9 Sphere/Schwefel
10 Sphere/Gallagher 101
11 sep. Ellipsoid/sep. Elli.
12 sep. Ellipsoid/Attr. sector
13 sep. Ellipsoid/Rosenbrock
14 sep. Ellipsoid/Sharp ridge
15 sep. Ellipsoid/Diff. Powers
16 sep. Ellipsoid/Rastrigin
17 sep. Ellipsoid/Schaffer F7
18 sep. Ellipsoid/Schwefel
19 sep. Elli./Gallagher 101
20 Attr. sector/Attr. sector
21 Attr. sector/Rosenbrock
22 Attr. sector/Sharp ridge
23 Attr. sector/Diff. Powers
24 Attr. sector/Rastrigin
25 Attr. sector/Schaffer F7
26 Attr. sector/Schwefel
27 Attr. sector/Gallagher 101
28 Rosenbrock/Rosenbrock
Fraction of function,target pairs
log10(# f-evals / dimension)
best 2016
UP-MO-CMA-
HMO-CMA-
MO-CMA-2

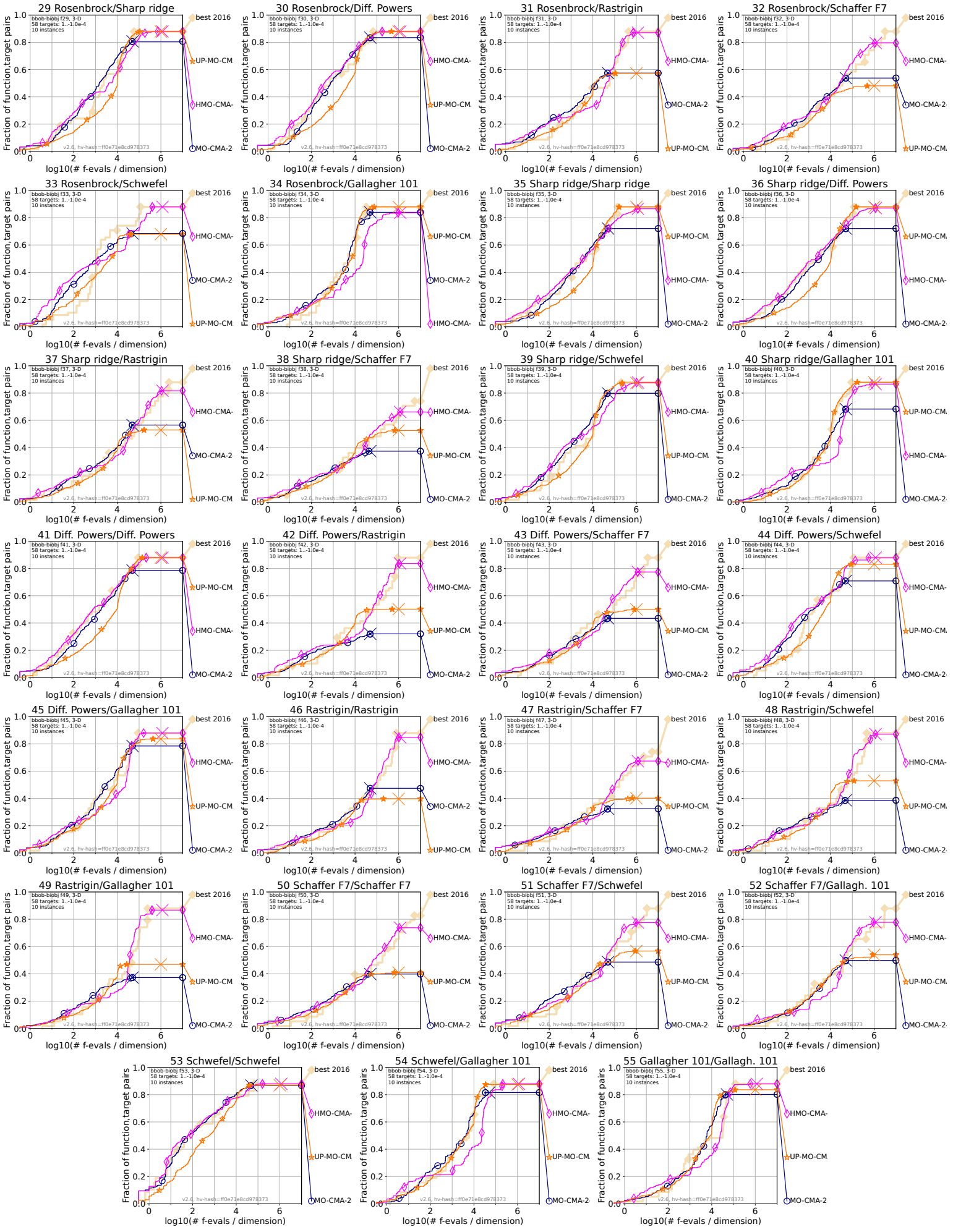
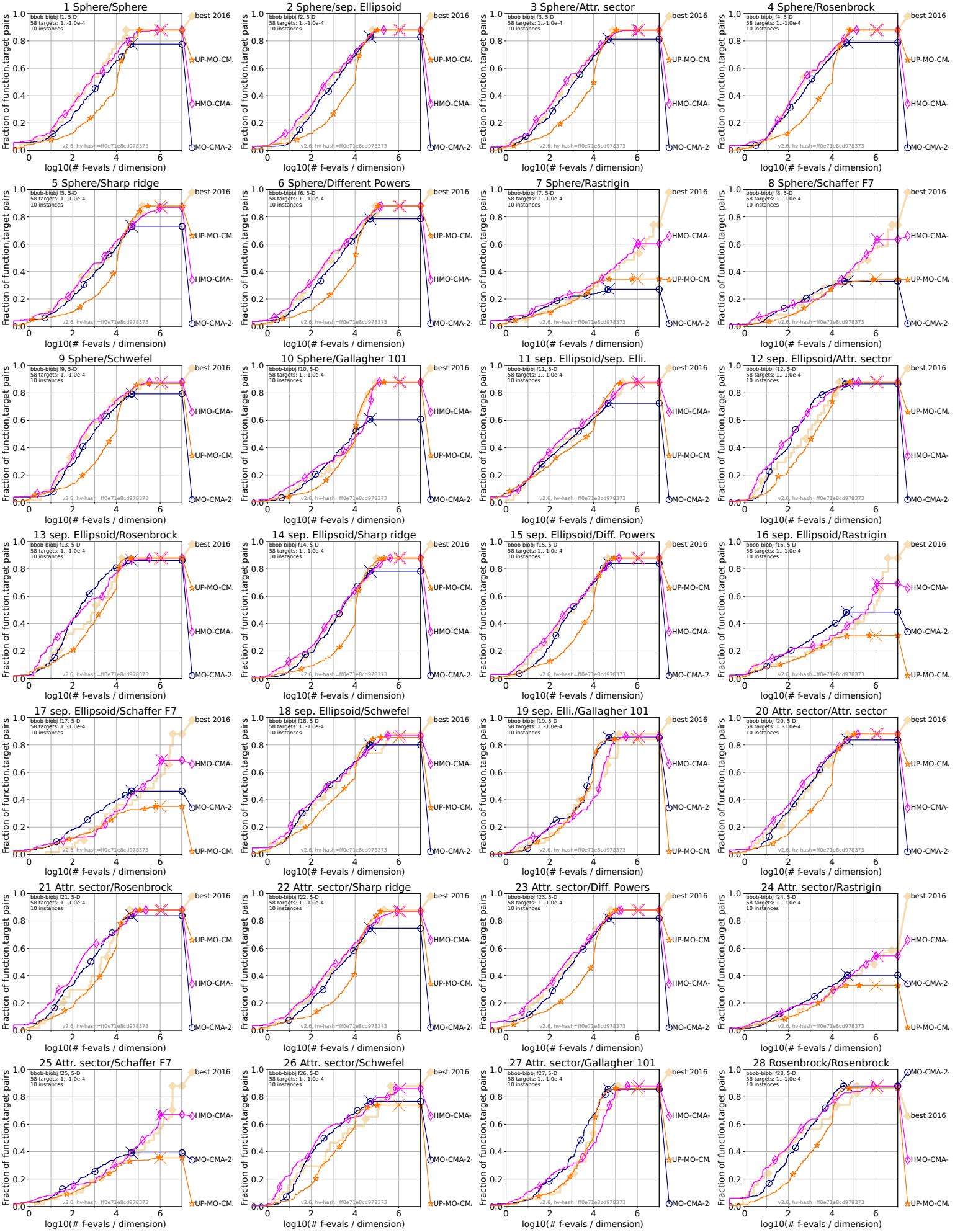


Figure C.6: ECDFs of COMO-CMA-ES, the standard MO-CMA-ES-2 and a modified version of MO-CMA-ES-2 with `random` runs only, for all `bbob-biobj` test problems and for search space dimension 3.


1 Sphere/Sphere
2 Sphere/sep. Ellipsoid
3 Sphere/Attr. sector
4 Sphere/Rosenbrock
5 Sphere/Sharp ridge
6 Sphere/Different Powers
7 Sphere/Rastrigin
8 Sphere/Schaffer F7
9 Sphere/Schwefel
10 Sphere/Gallagher 101
11 sep. Ellipsoid/sep. Elli.
12 sep. Ellipsoid/Attr. sector
13 sep. Ellipsoid/Rosenbrock
14 sep. Ellipsoid/Sharp ridge
15 sep. Ellipsoid/Diff. Powers
16 sep. Ellipsoid/Rastrigin
17 sep. Ellipsoid/Schaffer F7
18 sep. Ellipsoid/Schwefel
19 sep. Elli./Gallagher 101
20 Attr. sector/Attr. sector
21 Attr. sector/Rosenbrock
22 Attr. sector/Sharp ridge
23 Attr. sector/Diff. Powers
24 Attr. sector/Rastrigin
25 Attr. sector/Schaffer F7
26 Attr. sector/Schwefel
27 Attr. sector/Gallagher 101
28 Rosenbrock/Rosenbrock
Fraction of function,target pairs
log10(# f-evals / dimension)
best 2016
UP-MO-CMA-
HMO-CMA-
MO-CMA-2

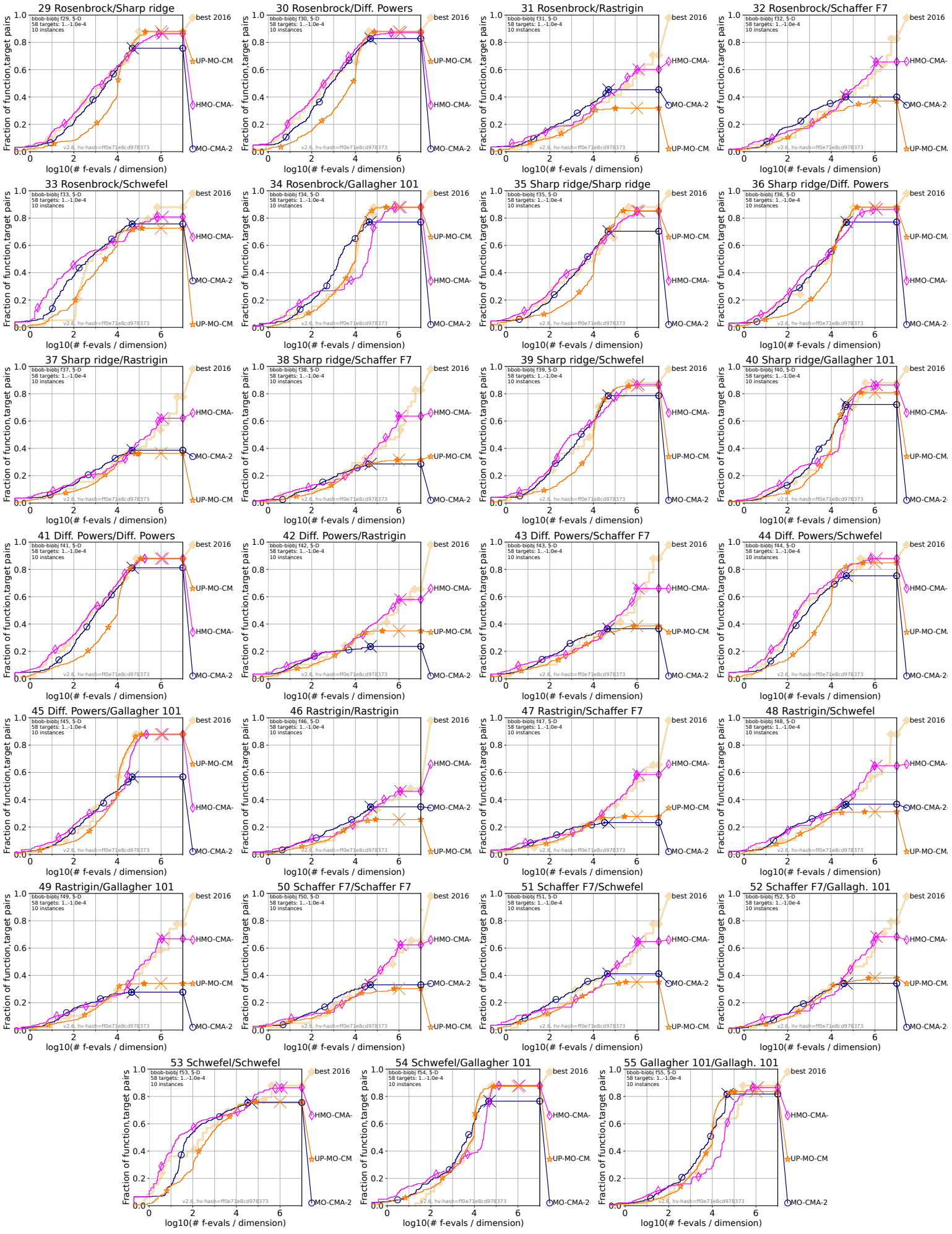
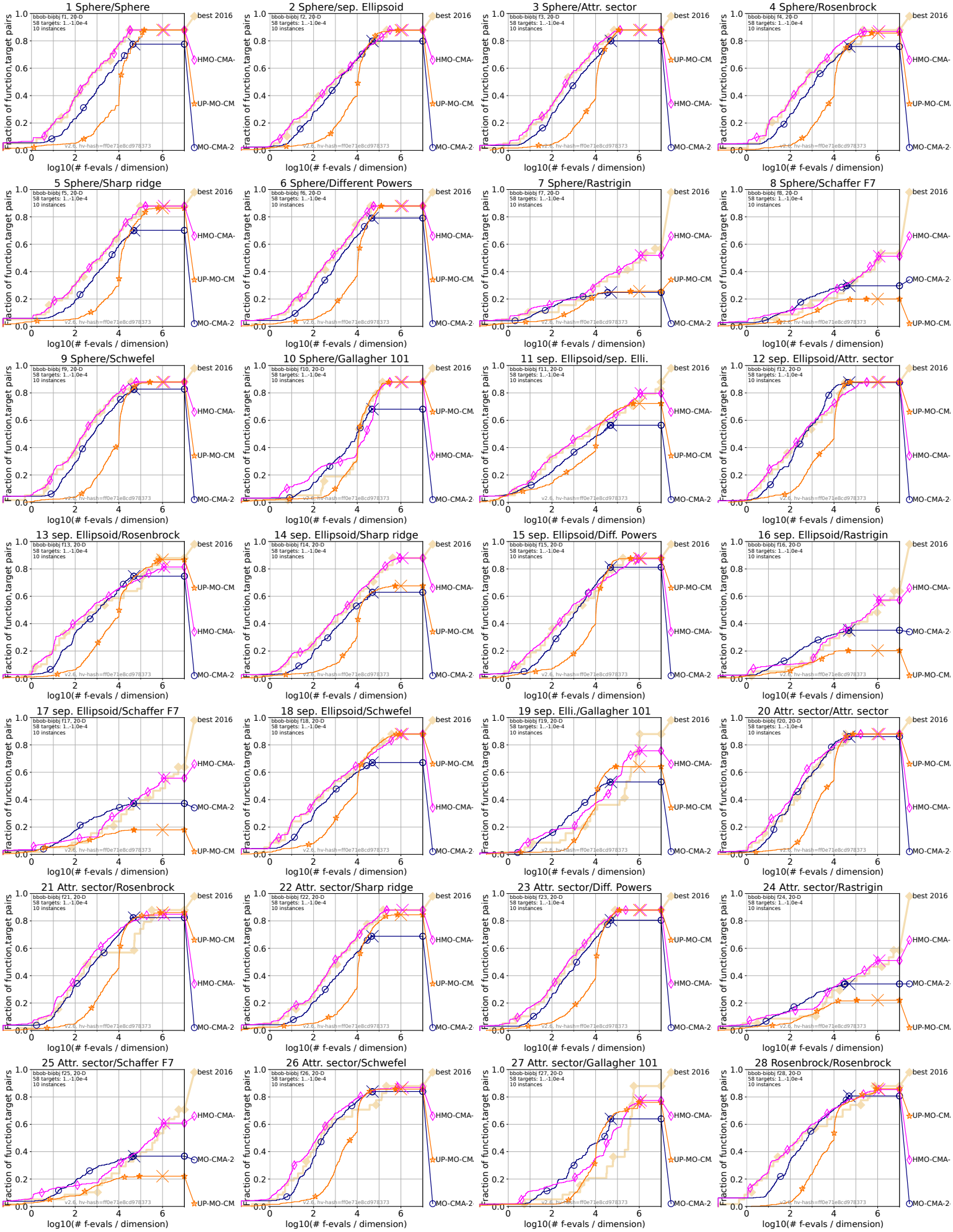


Figure C.7: ECDFs of COMO-CMA-ES, the standard MO-CMA-ES-2 and a modified version of MO-CMA-ES-2 with `random` runs only, for all `bbob-biobj` test problems and for search space dimension 5.


1 Sphere/Sphere
2 Sphere/sep. Ellipsoid
3 Sphere/Attr. sector
4 Sphere/Rosenbrock
5 Sphere/Sharp ridge
6 Sphere/Different Powers
7 Sphere/Rastrigin
8 Sphere/Schaffer F7
9 Sphere/Schwefel
10 Sphere/Gallagher 101
11 sep. Ellipsoid/sep. Elli.
12 sep. Ellipsoid/Attr. sector
13 sep. Ellipsoid/Rosenbrock
14 sep. Ellipsoid/Sharp ridge
15 sep. Ellipsoid/Diff. Powers
16 sep. Ellipsoid/Rastrigin
17 sep. Ellipsoid/Schaffer F7
18 sep. Ellipsoid/Schwefel
19 sep. Elli./Gallagher 101
20 Attr. sector/Attr. sector
21 Attr. sector/Rosenbrock
22 Attr. sector/Sharp ridge
23 Attr. sector/Diff. Powers
24 Attr. sector/Rastrigin
25 Attr. sector/Schaffer F7
26 Attr. sector/Schwefel
27 Attr. sector/Gallagher 101
28 Rosenbrock/Rosenbrock
Fraction of function,target pairs
log10(# f-evals / dimension)
best 2016
HMO-CMA-
UP-MO-CM
MO-CMA-2

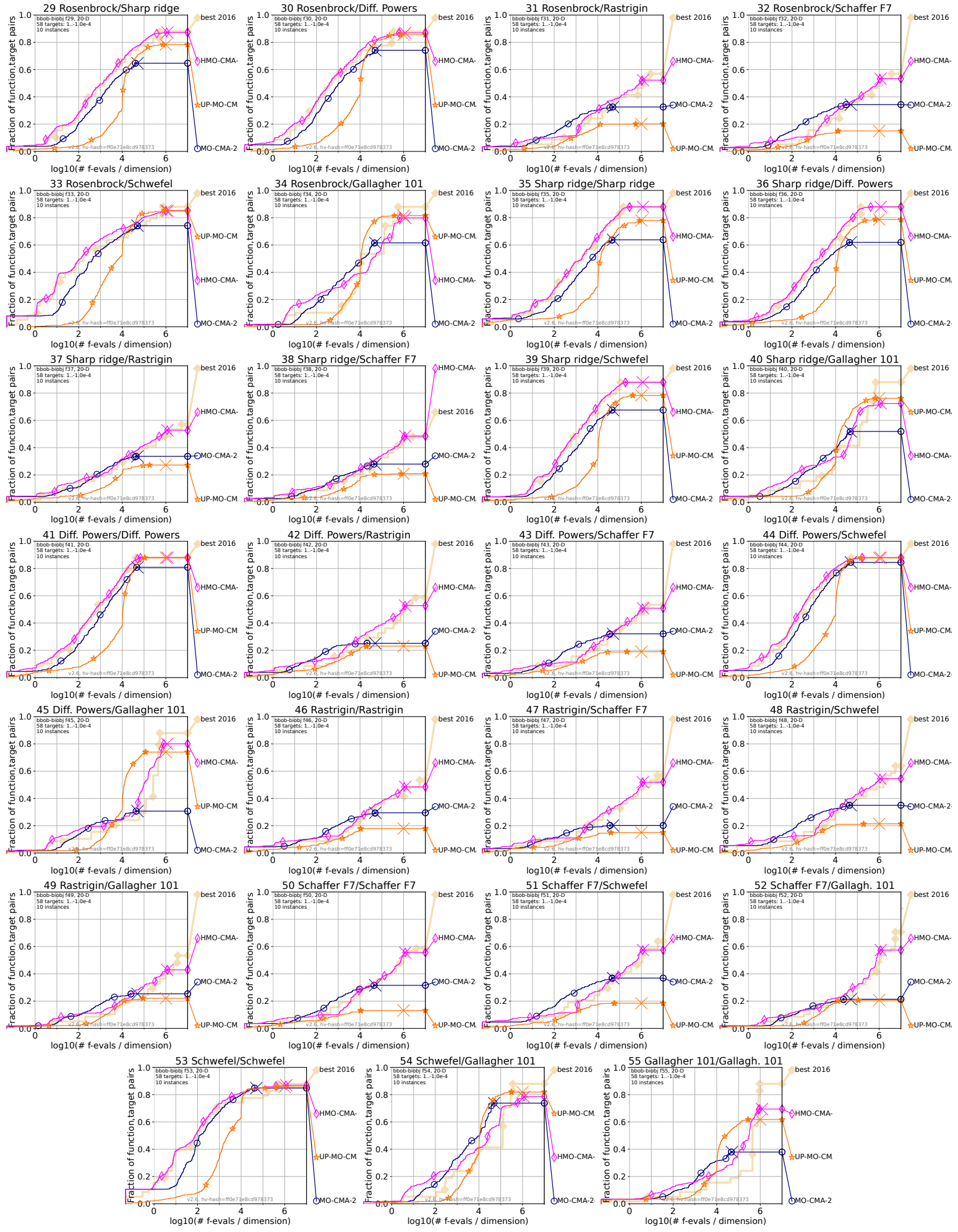


Figure C.8: ECDFs of COMO-CMA-ES, the standard MO-CMA-ES-2 and a modified version of MO-CMA-ES-2 with `random` runs only, for all `bbob-biobj` test problems and for search space dimension 20.

Bibliography

[1] Youhei Akimoto, Anne Auger, Tobias Glasmachers, and Daiki Morinaga. "Global Linear Convergence of Evolution Strategies on More than Smooth Strongly Convex Functions". In: *SIAM Journal on Optimization* 32.2 (June 2022), pp. 1402–1429. ISSN: 1052-6234. DOI: 10.1137/20M1373815.

[2] Charles Audet, Jean Bigeon, Dominique Cartier, Sébastien Le Digabel, and Ludovic Salomon. "Performance Indicators in Multiobjective Optimization". In: *European Journal of Operational Research* 292.2 (July 2021), pp. 397–422. ISSN: 03772217. DOI: 10.1016/j.ejor.2020.11.016.

[3] Charles Audet, Gilles Savard, and Walid Zghal. "Multiobjective Optimization Through a Series of Single-Objective Formulations". In: *SIAM Journal on Optimization* 19.1 (Jan. 2008), pp. 188–210. ISSN: 1052-6234. DOI: 10.1137/060677513.

[4] Charles Audet and Hare Warren. *Derivative-Free and Blackbox Optimization*. Springer, 2017.

[5] A. Auger and N. Hansen. "Theory of Evolution Strategies: A New Perspective". In: *Theory of Randomized Search Heuristics: Foundations and Recent Developments*. Ed. by A. Auger and B. Doerr. Singapore: World Scientific Publishing, 2011. Chap. 10, pp. 289–325.

[6] Anne Auger, Johannes Bader, and Dimo Brockhoff. "Theoretically investigating optimal μ-distributions for the hypervolume indicator: first results for three objectives". In: *Proceedings of the 11th international conference on Parallel problem solving from nature: Part I*. PPSN'10. Berlin, Heidelberg: Springer-Verlag, Sept. 2010, pp. 586–596. ISBN: 978-3-642-15843-8.

[7] Anne Auger, Johannes Bader, Dimo Brockhoff, and Eckart Zitzler. "Hypervolume-Based Multiobjective Optimization: Theoretical Foundations and Practical Implications". In: *Theoretical Computer Science* 425 (Mar. 2012), pp. 75–103. ISSN: 0304-3975. DOI: 10.1016/j.tcs.2011.03.012.

[8] Anne Auger, Johannes Bader, Dimo Brockhoff, and Eckart Zitzler. "Theory of the Hypervolume Indicator: Optimal μ-Distributions and the Choice of the Reference Point". In: *Proceedings of the Tenth ACM SIGEVO Workshop on Foundations of Genetic Algorithms*. FOGA '09. New York, NY, USA: Association for Computing Machinery, Jan. 2009, pp. 87–102. ISBN: 978-1-60558-414-0. DOI: 10.1145/1527125.1527138.

[9] Anne Auger, Nikolaus Hansen, Jorge Perez Zerpa, Raymond Ros, and Marc Schoenauer. "Empirical Comparisons of Several Derivative Free Optimization Algorithms". In: *9e Colloque National En Calcul Des Structures, CSMA*. Giens, France, May 2009, p. 12.

[10] Thomas Bäck. *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. New York: Oxford University Press, 1996. ISBN: 978-0-19-509971-3. DOI: 10.1093/oso/9780195099713.001.0001.

[11] Johannes Bader and Eckart Zitzler. "HypE: An Algorithm for Fast Hypervolume-Based Many-Objective Optimization". In: *Evolutionary Computation* 19.1 (Mar. 2011), pp. 45–76. ISSN: 1063-6560. DOI: 10.1162/EVCO_a_00009.

[12] Pauline Bennet, Carola Doerr, Antoine Moreau, Jeremy Rapin, Fabien Teytaud, and Olivier Teytaud. "Nevergrad: Black-Box Optimization Platform". In: *ACM SIGEVOlution* 14.1 (Apr. 2021), pp. 8–15. ISSN: 1931-8499. DOI: 10.1145/3460310.3460312.

[13] Nicola Beume, Carlos M. Fonseca, Manuel Lopez-Ibanez, LuÍs Paquete, and Jan Vahrenhold. "On the Complexity of Computing the Hypervolume Indicator". In: *IEEE Transactions on Evolutionary Computation* 13.5 (Oct. 2009), pp. 1075–1082. ISSN: 1941-0026. DOI: 10.1109/TEVC.2009.2015575.

[14] Nicola Beume, Marco Laumanns, and Günter Rudolph. "Convergence Rates of (1+1) Evolutionary Multiobjective Optimization Algorithms". In: *Parallel Problem Solving from Nature, PPSN XI*. Ed. by Robert Schaefer, Carlos Cotta, Joanna Kołodziej, and Günter Rudolph. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2010, pp. 597–606. ISBN: 978-3-642-15844-5. DOI: 10.1007/978-3-642-15844-5_60.

[15] Nicola Beume, Boris Naujoks, and Michael Emmerich. "SMS-EMOA: Multiobjective Selection Based on Dominated Hypervolume". In: *European Journal of Operational Research* 181.3 (Sept. 2007), pp. 1653–1669. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2006.08.008.

[16] Jean Bigeon, Sébastien Le Digabel, and Ludovic Salomon. "DMulti-MADS: Mesh Adaptive Direct Multisearch for Bound-Constrained Blackbox Multiobjective Optimization". In: *Computational Optimization and Applications* 79.2 (June 2021), pp. 301–338. ISSN: 1573-2894. DOI: 10.1007/s10589-021-00272-9.

[17] Karl Bringmann and Tobias Friedrich. "Approximating the Volume of Unions and Intersections of High-Dimensional Geometric Objects". In: *Computational Geometry* 43.6 (Aug. 2010), pp. 601–610. ISSN: 0925-7721. DOI: 10.1016/j.comgeo.2010.03.004.

[18] Karl Bringmann and Tobias Friedrich. "The Maximum Hypervolume Set Yields Near-Optimal Approximation". In: *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation - GECCO '10*. Portland, Oregon, USA: ACM Press, 2010, p. 511. ISBN: 978-1-4503-0072-8. DOI: 10.1145/1830483.1830576.

[19] Karl Bringmann and Tobias Friedrich. "Tight Bounds for the Approximation Ratio of the Hypervolume Indicator". In: *Parallel Problem Solving from Nature, PPSN XI*. Ed. by Robert Schaefer, Carlos Cotta, Joanna Kołodziej, and Günter Rudolph. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2010, pp. 607–616. ISBN: 978-3-642-15844-5. DOI: 10.1007/978-3-642-15844-5_61.

[20] Dimo Brockhoff. "Optimal μ-distributions for the hypervolume indicator for problems with linear bi-objective fronts: exact and exhaustive results". In: *Proceedings of the 8th international conference on Simulated evolution and learning*. SEAL'10. Berlin, Heidelberg: Springer-Verlag, Dec. 2010, pp. 24–34. ISBN: 978-3-642-17297-7.

[21] Dimo Brockhoff, Tea Tusar, Anne Auger, and Nikolaus Hansen. *Using Well-Understood Single-Objective Functions in Multiobjective Black-Box Optimization Test Suites*. Jan. 2019. arXiv: 1604.00359 [cs].

[22] Dimo Brockhoff, Tea Tušar, Dejan Tušar, Tobias Wagner, Nikolaus Hansen, and Anne Auger. *Biobjective Performance Assessment with the COCO Platform*. May 2016. arXiv: 1605.01746 [cs].

[23] Ma. Guadalupe Castillo Tapia and Carlos A. Coello Coello. "Applications of Multi-Objective Evolutionary Algorithms in Economics and Finance: A Survey". In: *2007 IEEE Congress on Evolutionary Computation*. Sept. 2007, pp. 532–539. DOI: 10.1109/CEC.2007.4424516.

[24] Timothy M. Chan. "Klee's Measure Problem Made Easy: 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, FOCS 2013". In: *Proceedings - 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, FOCS 2013*. Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS (2013), pp. 410–419. ISSN: 9780769551357. DOI: 10.1109/FOCS.2013.51.

[25] Morris R Cohen and I. E Drabkin. *A Source Book In Greek Science*. Cambridge: Harvard University Press, 1958.

[26] A. L. Custódio and J. F. A. Madeira. "MultiGLODS: Global and Local Multiobjective Optimization Using Direct Search". In: *Journal of Global Optimization* 72.2 (Oct. 2018), pp. 323–345. ISSN: 1573-2916. DOI: 10.1007/s10898-018-0618-1.

[27] A. L. Custódio, J. F. A. Madeira, A. I. F. Vaz, and L. N. Vicente. "Direct Multisearch for Multiobjective Optimization". In: *SIAM Journal on Optimization* 21.3 (July 2011), pp. 1109–1140. ISSN: 1052-6234, 1095-7189. DOI: 10.1137/10079731X.

[28] Kerstin Dächert and Kathrin Klamroth. "A Linear Bound on the Number of Scalarizations Needed to Solve Discrete Tricriteria Optimization Problems". In: *Journal of Global Optimization* 61.4 (Apr. 2015), pp. 643–676. ISSN: 1573-2916. DOI: 10.1007/s10898-014-0205-z.

[29] Kerstin Dächert and Katrin Teichert. "An Improved Hyperboxing Algorithm for Calculating a Pareto Front Representation". In: *arXiv:2003.14249* (2020). arXiv: 2003.14249.

[30] Etienne de Klerk, François Glineur, and Adrien B. Taylor. "On the Worst-Case Complexity of the Gradient Method with Exact Line Search for Smooth Strongly Convex Functions". In: *Optimization Letters* 11.7 (Oct. 2017), pp. 1185–1199. ISSN: 1862-4480. DOI: 10.1007/s11590-016-1087-4.

[31] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. "A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II". In: *IEEE Transactions on Evolutionary Computation* 6.2 (Apr. 2002), pp. 182–197. ISSN: 1941-0026. DOI: 10.1109/4235.996017.

[32] K. Deb, L. Thiele, M. Laumanns, and E. Zitzler. "Scalable Multi-Objective Optimization Test Problems". In: *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*. Vol. 1. Honolulu, HI, USA, USA: IEEE, May 2002, 825–830 vol.1. DOI: 10.1109/CEC.2002.1007032.

[33] Kalyanmoy Deb and Himanshu Jain. "An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints". In: *IEEE Transactions on Evolutionary Computation* 18.4 (Aug. 2014), pp. 577–601. ISSN: 1941-0026. DOI: 10.1109/TEVC.2013.2281535.

[34] Shaul Druckmann, Yoav Banitt, Albert Gidon, Felix Schürmann, Henry Markram, and Idan Segev. "A Novel Multiple Objective Optimization Framework for Constraining Conductance-Based Neuron Models by Experimental Data". In: *Frontiers in Neuroscience* 1 (2007). ISSN: 1662-453X.

[35] Fabian Duddeck. "Multidisciplinary Optimization of Car Bodies". In: *Structural and Multidisciplinary Optimization* 35.4 (Apr. 2008), pp. 375–389. ISSN: 1615-1488. DOI: 10.1007/s00158-007-0130-6.

[36] Paul Dufossé and Cheikh Touré. "Benchmarking MO-CMA-ES and COMO-CMA-ES on the Bi-Objective Bbob-Biobj Testbed". In: *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. Prague Czech Republic: ACM, July 2019, pp. 1920–1927. ISBN: 978-1-4503-6748-6. DOI: 10.1145/3319619.3326892.

[37] Francis Ysidro Edgeworth. *Mathematical Psychics. An Essay on the Application of Mathematics to the Moral Sciences.* Vol. 10. CK Paul, 1881.

[38] Matthias Ehrgott. *Multicriteria Optimization*. Second. Berlin Heidelberg: Springer-Verlag, 2005. ISBN: 978-3-540-21398-7. DOI: 10.1007/3-540-27659-9.

[39] Michael T. M. Emmerich and André H. Deutz. "A Tutorial on Multiobjective Optimization: Fundamentals and Evolutionary Methods". In: *Natural Computing* 17.3 (Sept. 2018), pp. 585–609. ISSN: 1567-7818, 1572-9796. DOI: 10.1007/s11047-018-9685-y.

[40] Jesús Guillermo Falcón-Cardona, Saúl Zapotecas-Martínez, and Abel García-Nájera. "Pareto Compliance from a Practical Point of View". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. GECCO '21. New York, NY, USA: Association for Computing Machinery, June 2021, pp. 395–402. ISBN: 978-1-4503-8350-9. DOI: 10.1145/3449639.3459276.

[41] Renato Fiorenza. *Hölder and locally Hölder Continuous Functions, and Open Sets of Class $\mathcal{C}^k$, $\mathcal{C}^{k,\lambda}$*. Frontiers in Mathematics. Birkhäuser Basel, 2016. ISBN: 978-3-319-47939-2. DOI: 10.1007/978-3-319-47940-8.

[42] J. Fliege, L. M. Graña Drummond, and B. F. Svaiter. "Newton's Method for Multiobjective Optimization". In: *SIAM Journal on Optimization* 20.2 (Jan. 2009), pp. 602–626. ISSN: 1052-6234, 1095-7189. DOI: 10.1137/08071692X.

[43] Ellen H. Fukuda and L. M. Graña Drummond. "On the Convergence of the Projected Gradient Method for Vector Optimization". In: *Optimization* 60.8-9 (Aug. 2011), pp. 1009–1021. ISSN: 0233-1934. DOI: 10.1080/02331934.2010.522710.

[44] Judith Grabiner. ""It's All for the Best": Optimization in the History of Science". In: *Journal of Humanistic Mathematics* 11.1 (Jan. 2021), pp. 54–80. ISSN: 21598118. DOI: 10.5642/jhummath.202101.05.

[45] Andreia P. Guerreiro, Carlos M. Fonseca, and Luís Paquete. "The Hypervolume Indicator: Problems and Algorithms". In: *arXiv:2005.00515 [cs]* (May 2020). arXiv: 2005.00515 [cs].

[46] N. Hansen and A. Ostermeier. "Adapting Arbitrary Normal Mutation Distributions in Evolution Strategies: The Covariance Matrix Adaptation". In: *Proceedings of IEEE International Conference on Evolutionary Computation*. May 1996, pp. 312–317. DOI: 10.1109/ICEC.1996.542381.

[47] Nikolaus Hansen. "The CMA Evolution Strategy: A Comparing Review". In: *Towards a New Evolutionary Computation: Advances in the Estimation of Distribution Algorithms*. Ed. by Jose A. Lozano, Pedro Larrañaga, Iñaki Inza, and Endika Bengoetxea. Studies in Fuzziness and Soft Computing. Berlin, Heidelberg: Springer, 2006, pp. 75–102. ISBN: 978-3-540-32494-2. DOI: 10.1007/3-540-32494-1_4.

[48] Nikolaus Hansen. *The CMA Evolution Strategy: A Tutorial*. Apr. 2016. arXiv: 1604.00772 [cs, stat].

[49] Nikolaus Hansen, Anne Auger, Raymond Ros, Olaf Mersmann, Tea Tušar, and Dimo Brockhoff. "COCO: A Platform for Comparing Continuous Optimizers in a Black-Box Setting". In: *Optimization Methods and Software* 36.1 (Jan. 2021), pp. 114–144. ISSN: 1055-6788, 1029-4937. DOI: 10.1080/10556788.2020.1808977. arXiv: 1603.08785 [cs, math, stat].

[50] Nikolaus Hansen and Stefan Kern. "Evaluating the CMA Evolution Strategy on Multimodal Test Functions". In: *Parallel Problem Solving from Nature - PPSN VIII*. Ed. by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Xin Yao, Edmund K. Burke, José A. Lozano, Jim Smith, Juan Julián Merelo-Guervós, John A. Bullinaria, Jonathan E. Rowe, Peter Tiňo, Ata Kabán, and Hans-Paul Schwefel. Vol. 3242. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 282–291. ISBN: 978-3-540-23092-2 978-3-540-30217-9. DOI: 10.1007/978-3-540-30217-9_29.

[51] Christian Igel, Nikolaus Hansen, and Stefan Roth. "Covariance Matrix Adaptation for Multi-objective Optimization". In: *Evolutionary Computation* 15.1 (Mar. 2007), pp. 1–28. ISSN: 1063-6560. DOI: 10.1162/evco.2007.15.1.1.

[52] Yaochu Jin. *Multi-Objective Machine Learning*. Springer Science & Business Media, Feb. 2006. ISBN: 978-3-540-30676-4.

[53] Dennis Kafura. *Algorithms - Variables and State: CS 1014 (Fall 2018)*. https://canvas.vt.edu/cours 3-dot-3-algorithms-variables-and-state.

[54] Haim Kaplan, Natan Rubin, Micha Sharir, and Elad Verbin. "Efficient Colored Orthogonal Range Counting". In: *SIAM Journal on Computing* 38.3 (Jan. 2008), pp. 982–1011. ISSN: 0097-5397, 1095-7111. DOI: 10.1137/070684483.

[55] Maureen C. Kennedy, E. David Ford, Peter Singleton, Mark Finney, and James K. Agee. "Informed Multi-Objective Decision-Making in Environmental Management Using Pareto Optimality". In: *Journal of Applied Ecology* 45.1 (2008), pp. 181–192. ISSN: 1365-2664. DOI: 10.1111/j.1365-2664.2007.01367.x.

[56] Mohand Ou Idir Khemmoudj, Marc Porcheron, and Hachemi Bennaceur. "When Constraint Programming and Local Search Solve the Scheduling Problem of Electricité de France Nuclear Power Plant Outages". In: *Principles and Practice of Constraint Programming - CP 2006*. Ed. by Frédéric Benhamou. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, pp. 271–283. ISBN: 978-3-540-46268-2. DOI: 10.1007/11889205_21.

[57] Kathrin Klamroth, Renaud Lacour, and Daniel Vanderpooten. "On the Representation of the Search Region in Multi-Objective Optimization". In: *European Journal of Operational Research* 245.3 (2015), pp. 767–778. ISSN: 03772217. DOI: 10.1016/j.ejor.2015.03.031.

[58] J. Knowles and D. Corne. "On Metrics for Comparing Nondominated Sets". In: *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*. Vol. 1. May 2002, 711–716 vol.1. DOI: 10.1109/CEC.2002.1007013.

[59] Oswin Krause, Tobias Glasmachers, Nikolaus Hansen, and Christian Igel. "Unbounded Population MO-CMA-ES for the Bi-Objective BBOB Test Suite". In: *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion*. Denver Colorado USA: ACM, July 2016, pp. 1177–1184. ISBN: 978-1-4503-4323-7. DOI: 10.1145/2908961.2931699.

[60] Gottfried Leibniz. "Nova Methodus pro Maximis et Minimis". In: *Acta Eruditorum* (1684).

[61] Ilya Loshchilov and Tobias Glasmachers. *Anytime Bi-Objective Optimization with a Hybrid Multi-Objective CMA-ES (HMO-CMA-ES)*. May 2016. DOI: 10.48550/arXiv.1605.02720. arXiv: 1605.02720 [cs].

[62] Eugénie Marescaux and Anne Auger. "Multiobjective Hypervolume Based ISOOMOO Algorithms Converge with At Least Sublinear Speed to the Entire Pareto Front". 2021.

[63] Eugénie Marescaux and Nikolaus Hansen. "Hypervolume in biobjective optimization cannot converge faster than $\Omega(1/p)$". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. GECCO '21. New York, NY, USA: Association for Computing Machinery, June 2021, pp. 430–438. ISBN: 978-1-4503-8350-9. DOI: 10.1145/3449639.3459371.

[64] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. "An Analysis of Approximations for Maximizing Submodular Set Functions—I". In: *Mathematical Programming* 14.1 (Dec. 1978), pp. 265–294. ISSN: 1436-4646. DOI: 10.1007/BF01588971.

[65] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Corr. 2. print. Springer Series in Operations Research. New York, NY: Springer, 2000. ISBN: 978-0-387-98793-4.

[66] Vilfredo Pareto. *Manual of Political Economy*. A. M. Kelley, 1971. ISBN: 978-0-678-00881-2.

[67] Sophie N. Parragh and Fabien Tricoire. "Branch-and-Bound for Bi-objective Integer Programming". In: *INFORMS Journal on Computing* 31.4 (Oct. 2019), pp. 805–822. ISSN: 1091-9856, 1526-5528. DOI: 10.1287/ijoc.2018.0856.

[68] Qingfu Zhang and Hui Li. "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition". In: *IEEE Transactions on Evolutionary Computation* 11.6 (Dec. 2007), pp. 712–731. ISSN: 1941-0026, 1089-778X. DOI: 10.1109/TEVC.2007.892759.

[69] Alberto Rojo and Anthony Bloch. *The Principle of Least Action: History and Physics*. Cambridge: Cambridge University Press, 2018. ISBN: 978-0-521-86902-7. DOI: 10.1017/9781139021029.

[70] Oliver Schütze and Carlos Hernández. *Archiving Strategies for Evolutionary Multi-objective Optimization Algorithms*. Vol. 938. Studies in Computational Intelligence. Cham: Springer International Publishing, 2021. ISBN: 978-3-030-63772-9 978-3-030-63773-6. DOI: 10.1007/978-3-030-63773-6.

[71] Ke Shang, Hisao Ishibuchi, Linjun He, and Lie Meng Pang. "A Survey on the Hypervolume Indicator in Evolutionary Multi-objective Optimization". In: *IEEE Transactions on Evolutionary Computation* (2020), pp. 1–1. ISSN: 1941-0026. DOI: 10.1109/TEVC.2020.3013290.

[72] Olivier Teytaud and Sylvain Gelly. "General Lower Bounds for Evolutionary Algorithms". In: *Parallel Problem Solving from Nature - PPSN IX*. Ed. by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Thomas Philip Runarsson, Hans-Georg Beyer, Edmund Burke, Juan J. Merelo-Guervós, L. Darrell Whitley, and Xin Yao. Vol. 4193. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 21–31. ISBN: 978-3-540-38990-3 978-3-540-38991-0. DOI: 10.1007/11844297_3.

[73] Cheikh Toure, Anne Auger, Dimo Brockhoff, and Nikolaus Hansen. "On Bi-objective Convex-Quadratic Problems". In: *Evolutionary Multi-Criterion Optimization*. Ed. by Kalyanmoy Deb, Erik Goodman, Carlos A. Coello Coello, Kathrin Klamroth, Kaisa Miettinen, Sanaz Mostaghim, and Patrick Reed. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 3–14. ISBN: 978-3-030-12598-1. DOI: 10.1007/978-3-030-12598-1_1.

[74] Cheikh Touré, Nikolaus Hansen, Anne Auger, and Dimo Brockhoff. "Uncrowded Hypervolume Improvement: COMO-CMA-ES and the Sofomore Framework". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. GECCO '19. New York, NY, USA: Association for Computing Machinery, July 2019, pp. 638–646. ISBN: 978-1-4503-6111-8. DOI: 10.1145/3321707.3321852.

[75] David A Van Veldhuizen and Gary B Lamont. "Evolutionary Computation and Convergence to a Pareto Front". In: *Proceedings of the Third Annual Conference on Genetic Programming*. San Francisco: Stanford University Bookstore, 1998, pp. 221–228.

[76] Thomas Voß, Nikolaus Hansen, and Christian Igel. "Improved Step Size Adaptation for the MO-CMA-ES". In: *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation - GECCO '10*. Portland, Oregon, USA: ACM Press, 2010, p. 487. ISBN: 978-1-4503-0072-8. DOI: 10.1145/1830483.1830573.

[77] Iryna Yevseyeva, Andreia P. Guerreiro, Michael T. M. Emmerich, and Carlos M. Fonseca. "A Portfolio Optimization Approach to Selection in Multiobjective Evolutionary Algorithms". In: *Parallel Problem Solving from Nature – PPSN XIII*. Ed. by Thomas Bartz-Beielstein, Jürgen Branke, Bogdan Filipič, and Jim Smith.

Vol. 8672. Cham: Springer International Publishing, 2014, pp. 672–681. ISBN: 978-3-319-10761-5 978-3-319-10762-2. DOI: 10.1007/978-3-319-10762-2_66.

[78] E. Zitzler, L. Thiele, M. Laumanns, C.M. Fonseca, and V.G. da Fonseca. "Performance Assessment of Multiobjective Optimizers: An Analysis and Review". In: *IEEE Transactions on Evolutionary Computation* 7.2 (Apr. 2003), pp. 117–132. ISSN: 1089-778X, 1089-778X, 1941-0026. DOI: 10.1109/TEVC.2003.810758.

[79] Eckart Zitzler, Dimo Brockhoff, and Lothar Thiele. "The Hypervolume Indicator Revisited: On the Design of Pareto-compliant Indicators Via Weighted Integration". In: *Evolutionary Multi-Criterion Optimization*. Ed. by Shigeru Obayashi, Kalyanmoy Deb, Carlo Poloni, Tomoyuki Hiroyasu, and Tadahiko Murata. Vol. 4403. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 862–876. ISBN: 978-3-540-70927-5 978-3-540-70928-2. DOI: 10.1007/978-3-540-70928-2_64.

[80] Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele. "Comparison of Multiobjective Evolutionary Algorithms: Empirical Results". In: *Evolutionary Computation* 8.2 (June 2000), pp. 173–195. ISSN: 1063-6560, 1530-9304. DOI: 10.1162/106365600568202.

[81] Eckart Zitzler, Joshua Knowles, and Lothar Thiele. "Quality Assessment of Pareto Set Approximations". In: *Multiobjective Optimization: Interactive and Evolutionary Approaches*. Ed. by Jürgen Branke, Kalyanmoy Deb, Kaisa Miettinen, and Roman Słowiński. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, pp. 373–404. ISBN: 978-3-540-88908-3. DOI: 10.1007/978-3-540-88908-3_14.

[82] Eckart Zitzler and Simon Künzli. "Indicator-Based Selection in Multiobjective Search". In: *Parallel Problem Solving from Nature - PPSN VIII*. Ed. by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Xin Yao, Edmund K. Burke, José A. Lozano, Jim Smith, Juan Julián Merelo-Guervós, John A. Bullinaria, Jonathan E. Rowe, Peter Tiňo, Ata Kabán, and Hans-Paul Schwefel. Vol. 3242. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 832–842. ISBN: 978-3-540-23092-2 978-3-540-30217-9. DOI: 10.1007/978-3-540-30217-9_84.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Analyse de convergence et nouveaux algorithmes en optimisation bi-objectifs

Mots clés : optimisation, multi-objectifs, vitesse de convergence, algorithme

Résumé : L'optimisation est le domaine des mathématiques appliquées qui s'intéresse à la minimisation (ou la maximisation) d'une ou plusieurs fonctions objectif. Elle a de nombreuses applications industrielles et scientifiques, de la planification des pannes de courant à la conception de voitures en passant par la description de phénomènes physiques. L'optimisation multi-objectifs s'intéresse à l'approximation de l'ensemble de Pareto, c'est-à-dire l'ensemble des solutions admissibles qui ne peuvent être améliorées suivant tous les objectifs simultanément, et du front de Pareto, son image dans l'espace des objectifs. Dans cette thèse de doctorat, nous cherchons à approfondir les connaissances sur la vitesse de convergence d'algorithmes d'optimisation multi-objectifs vers la totalité du front de Pareto. Nous nous appuyons sur l'hypervolume, un indicateur de qualité d'ensemble largement utilisé, pour quantifier l'écart entre le front de Pareto et une approximation fournie par un algorithme, autrement dit l'erreur. Le coût est mesuré par le nombre de fois où les fonctions objectif ont été évaluées.

Tout d'abord, nous prouvons une borne supérieure théorique sur la vitesse de convergence. Nous démontrons que pour une vaste catégorie de fronts de Pareto, la plus petite erreur associée à une approximation du front de Pareto composée de n solutions est supérieure à $1/(n+1)$ multiplié par une constante. Cela garantit que la vitesse de convergence est au mieux sous-linéaire.

Ensuite, nous définissons une nouvelle classe d'algorithmes, HV-ISOOMOO. Une méta-itération de HV-ISOOMOO correspond à la résolution d'un sous-problème d'optimisation mono-objectif. Les solutions renvoyées par le solveur mono-objectif fournissent une approximation du front de Pareto. Nous démontrons des bornes inférieures sur la vitesse de convergence de cette approximation vers le front de Pareto en supposant que le solveur mono-objectif fournisse toujours un optimum global avec exactitude. Pour les fronts de Pareto convexes, cet algorithme idéal a une vitesse de convergence optimale: $\Theta(1/n)$.

Finalement, nous détaillons une implémentation de HV-ISOOMOO qui utilise le solveur mono-objectif CMA-ES, MO-CMA-ES-2. Cet algorithme se compare favorablement aux meilleurs algorithmes multi-objectifs connus à ce jour. Sur un problème convexe simple, nous analysons empiriquement l'évolution de l'erreur en fonction du nombre de méta-itérations de MO-CMA-ES-2 et du nombre d'itérations de CMA-ES.

Title : Convergence analysis and novel algorithms in bi-objective optimization

Keywords : optimization, multi-objective, convergence speed, algorithm

Abstract : Optimization is the field of applied mathematics concerned with minimizing (or maximizing) one or more objective functions. It has many applications in science and industry, from scheduling power outages through designing cars to describing physical phenomena. Multi-objective optimization aims at finding a good approximation of the Pareto set, that is the set of feasible solutions that cannot be improved in all objectives simultaneously, and the Pareto front, its image in the objective space. In this Ph.D. thesis, we seek to improve the understanding of the convergence speed of multi-objective optimization algorithms towards the entire Pareto front. We rely on the hypervolume, a widely used set quality indicator, to quantify the optimality gap between the Pareto front and its approximation provided by an algorithm. The computational cost is typically measured by the number of evaluations of the objective functions.

First, we derive a theoretical upper bound on the speed of convergence. We prove that for a wide class of Pareto fronts, the smallest optimality gap associated with a set of n points is larger than $1/(n+1)$ multiplied by a constant. This result shows that the speed of convergence is sublinear at best.

Then, we introduce an algorithmic framework, HV-ISOOMOO. A meta-iteration of HV-ISOOMOO corresponds to the solving of a single-objective optimization subproblem. We provide lower bounds on the convergence speed of the Pareto front approximation formed by the solutions returned by the single-objective solver (which we call the *final incumbents Pareto front approximation*) under the assumption that the solver returns a global optimum with infinite precision. For convex Pareto fronts, this ideal algorithm reaches the best achievable convergence rate: $\Theta(1/n)$.

Finally, we provide an implementation of HV-ISOOMOO coupled with CMA-ES, a state-of-the-art single-objective optimization algorithm: MO-CMA-ES-2. The MO-CMA-ES-2 algorithm presents state-of-the-art performance. On a simple convex problem, we empirically analyze the evolution of the optimality gap of its final incumbents Pareto front approximation with respect to the number of meta-iterations of HV-ISOOMOO and iterations of CMA-ES.

Institut Polytechnique de Paris
91120 Palaiseau, France