Niching in Derandomized Evolution Strategies and its Applications in Quantum Control

A Journey from Organic Diversity to Conceptual Quantum Designs

Ofer Michael Shir

Niching in Derandomized Evolution Strategies and its Applications in Quantum Control

A Journey from Organic Diversity to Conceptual Quantum Designs

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 25 juni 2008
klokke 16:15 uur

door

Ofer Michael Shir

geboren te Jeruzalem, Israël in 1978

Promotiecommissie:

Prof. dr. Thomas Bäck	Promotor
Prof. dr. Marc Vrakking (Amolf-Amsterdam)	Promotor
Dr. Michael Emmerich	Co-promotor
Prof. dr. Darrell Whitley (Colorado State University)	Referent
Prof. dr. Farhad Arbab	
Prof. dr. Joost Kok	

This work is part of the research programme of the 'Stichting voor Fundamenteel Onderzoek der Materie (FOM)', which is financially supported by the 'Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)'.

Niching in Derandomized Evolution Strategies and its Applications in Quantum Control.
Ofer Michael Shir.
Thesis Universiteit Leiden.
ISBN: 978-90-6464-256-2
Printed in the Netherlands.

To my parents, Mira and Yos'ke

Contents

There are more things in heaven and earth, Horatio,
than are dreamt of in your philosophy.
Prince Hamlet, **Hamlet**; William Shakespeare

Introduction

Optimal behavior of natural systems is frequently encountered at all levels of everyday life, and thus has become a major source of inspiration for various fields. The discipline of Natural Computing aims at developing computational techniques that mimic collective phenomena in nature that often exhibit excellent behavior in information processing. Among a long list of natural computing branches, we are particularly interested in the fascinating field of *Organic Evolution*, and its computational derivative, the so-called *Evolutionary Algorithms* (EAs) field. By encoding an optimization problem into an artificial biological environment, EAs mimic certain elements in the Darwinian dynamics and aim at obtaining highly-fit solutions in terms of the problem. A population of trial solutions undergo artificial variations and survive this simulation upon the criteria posed by the selection mechanism. Analogously, it is suggested that this population would evolve into highly-fit solutions of the optimization problem.

The original goal of this work was to extend specific variants of EAs, called Evolution Strategies (ES), to subpopulations of trial solutions which evolve in parallel to various solutions of the problem. This idea stems from the evolutionary concept of *organic speciation*. Essentially, the *natural computing* way of thinking is required here to further deepen into Evolutionary Biology Theory, and attain creative solutions for the artificial population in light of the desired speciation effect. The so-called *niching* techniques are the extension of EAs to speciation forming multiple subpopulations. They have been investigated since the early days of EAs, mainly within the popular variants of Genetic Algorithms (GAs). In addition to the theoretical challenge to design such techniques, which is well supported by the biologically inspired motivation, there is a *real-world incentive* for this effort. The discipline of *decision making*, which makes direct benefit out of the advent of the global optimization field, poses the demand for the multiplicity of different optimal solutions. Ideally, those multiple solutions, as obtained by the optimization routine, would have high diversity among each other, and represent different *conceptual designs*.

Aiming at largely devoting this research to niching in ES, we were also originally interested in applying our proposed algorithms to *experimental optimization*. More specifically, we were aiming at applications in the emerg-

ing field of Quantum Control (QC). The latter offers an enormous variety of high-dimensional continuous optimization problems, both at the *theoretical* as well as the *experimental* levels. In that respect, it is potentially a heavenly testbed for Evolutionary optimization, and particularly for niching methods. This is due to some remarkable properties of QC landscapes, which typically possess an infinite number of optimal solutions, as proved by QC Theory. We thus find the combination of research on *niching* and the application to QC landscapes very attractive. After being exposed to this overwhelming treasure of QC landscape richness, we decided to devote an independent part of this dissertation to Quantum Control.

Symbolically, this interdisciplinary study forms a *closed natural computing circle*, where biologically-oriented investigation of organic evolution and speciation helps to develop methods for solving applications in Physics in general, and in Quantum Control in particular. By our reckoning, this symbolism is even further strengthened upon considering the stochastic nature of Evolutionary Algorithms; This process can be thus considered as throwing dice in order to solve Quantum Mechanics, sometimes referred to as the *science of dice*.

Thus, biologically inspired by organic evolution in general, and organic speciation in particular, armed with the real-world incentive to obtain multiple optimal solutions for better decision making, we hereby begin our journey from diversity in nature to conceptual designs in Quantum Control.

This dissertation therefore consists of two parts: Part I introduces a niching framework to a set of state-of-the-art ES algorithms, namely Derandomized Evolution Strategies (DES), and focuses on testing the proposed algorithms on artificial landscapes. Part II reviews the main aspects of Quantum Control in the general context of global function optimization. It then presents the experimental observation of Derandomized ES as well as the proposed niching algorithms when applied to several QC systems, both at the laboratory and at the numerical simulations levels. As far as we know, this is the first time that Quantum Control search landscapes are comprehensively introduced to the community of Computer Science.

Part I begins with presenting the algorithmic kernels of this study, Derandomized Evolution Strategies. This is done in Chapter 1 by providing the reader with the essential terminology of global optimization, reviewing the fundamentals of the ES field, and eventually introducing explicitly, in detail, the derandomized algorithms.

Upon developing a niching framework for Evolution Strategies, some preliminary topics had to be addressed. We properly introduce the real-world incentive for niching, namely the selection of conceptual designs by the decision maker. Furthermore, we review elementary concepts of the Organic

Speciation Theory, discuss the crucial aspect of *population diversity* within ES, and finally present a short overview of previously introduced niching techniques. Chapter 2 aims at addressing those topics, and therefore it constitutes an important preliminary study for the derivation of our niching framework. Due to the highly interdisciplinary nature of the niching research, this chapter presents a particularly high diversity of topics, which are linked by *niching*.

In Chapter 3 we present our proposed framework of niching within Derandomized ES. We describe it in detail, and thereafter test it on a suite of multimodal artificial landscapes. We analyze the numerical observation, and discuss the algorithmic performance.

Chapter 4 extends the framework of Chapter 3 to self-adaptive niche-shape approaches, for solving the so-called *niche radius problem*. This is an important topic in the field of niching, as it attempts to treat the challenge of defining a generic basin of attraction without *a-priori* knowledge on the landscape.

Another extension of our proposed niching framework, this time to the field of Multi-Objective Optimization, is introduced in Chapter 5. As the two fields of niching and multi-criterion optimization, corresponding to multimodal and multiobjective problems, respectively, have many aspects in common, we show the feasibility of utilizing our niching framework in a multi-objective approach. This concludes Part I of the thesis.

The goal that Part II aims to achieve is two-fold: Firstly, properly introducing the main optimization aspects of the Quantum Control field, and secondly, presenting our work on the optimization of a specific Quantum Control problem, namely Dynamic Molecular Alignment. We thus begin Chapter 6 with a detailed review of Quantum Control Theory and Experiments. The review outlines fundamental concepts of Quantum Control Theory, and mainly focuses on theorems concerning the critical points of the landscapes, as well as on landscape richness and multiplicity of optimal solutions. It then presents Quantum Control Experiments, and discusses our experimental setup for Part II.

Chapter 7 describes our investigation of two optimization problems corresponding to Quantum Control systems of Second Harmonic Generation. We conduct experiments on these optimization problems, by means of numerical simulations as well as laboratory experiments, by employing specific Derandomized ES variants. It is the only chapter where we report on real-world laboratory experiments, while the following chapters focus on numerical simulations exclusively.

Chapter 8 is devoted to the introduction of the *rotational framework*, the fundamental framework upon which the Dynamic Molecular Alignment problem is based. In that respect, this chapter can be considered as a *gateway*

to our work on the alignment problem investigated in Chapter 9. Following a detailed Quantum Mechanical description of the framework, Chapter 8 poses the rotational population transfer optimization problem. It then presents our numerical observation of the Derandomized ES employment to the problem, and finalizes the chapter with applying our proposed niching algorithms.

Chapter 9 reports in detail on our work on the Dynamic Molecular Alignment, which constitutes the main application in our research on Quantum Control landscapes. It describes the alignment problem, and then presents various optimization approaches that we employed in addition to the straightforward application of Derandomized ES. These approaches include a special parameterization method developed for this purpose, optimality investigation of a simplified variant, optimization subject to a dynamically varying environment, multi-objective consideration of the problem, and, finally, the application of niching.

We thereafter complete this journey by summarizing our main results and by presenting promising directions for future research.

A Technical Note Due to technical printing considerations, several plots from various chapters are concentrated in Appendix A. In these particular cases, a plot is referred to in the text as Figure **A.x**.

Part I

Niching in Derandomized Evolution Strategies

If it could be demonstrated that any complex organ existed, which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down.
Charles Darwin

Chapter 1

Evolution Strategies

1.1 Background

The paradigm of *Evolutionary Computation* (EC), which is gleaned from the model of *organic evolution*, studies populations of candidate solutions undergoing variations and selection, and aims at benefiting from the collective phenomena of their generational behavior. The term Evolutionary Algorithms (EAs) essentially refers to the collection of such generic methods, inspired by the theory of *natural evolution*, that encode complex problems into an artificial biological environment, define its genetic operators, and simulate its propagation in time. Motivated by the basic principles of the Darwinian theory, it is suggested that such simulation would yield an optimal solution for the given problem.

Evolutionary Algorithms [1] have three main streams, rooted either in the United States or in Germany, during the 1960s: *Evolutionary Programming* (EP), founded by L. Fogel in San-Diego [2], *Genetic Algorithms* (GAs) founded by J. Holland in Ann Arbor [3, 4], and *Evolution Strategies* (ES), founded by P. Bienert, H.P. Schwefel and I. Rechenberg, three students to that time at the Technical University of Berlin (see, e.g., [5, 6, 7]).

Evolution Strategies for *global parameter optimization*, the general framework of this study, is reviewed in this chapter. We start with laying out the basic foundations and definitions.

1.1.1 The Framework: Global Optimization

Let us introduce the elementary terminology of a continuous real-valued *parameter optimization problem* [8]. The following definition excludes discrete and mixed-integer problems. Given an objective function, also called the target function,

$$f \; : \; \mathcal{S} \subseteq \mathbb{R}^n \rightarrow \mathbb{R}, \quad \mathcal{S} \neq \emptyset$$

where $\mathcal{S}$ is the set of *feasible solutions*

$$\mathcal{S} = \{\vec{x} \in \mathbb{R}^n \mid g_j(\vec{x}) \geq 0 \; \forall j \in \{1, ..., q\}\}, \quad g_j(\vec{x}) \; : \; \mathbb{R}^n \to \mathbb{R}$$

subject to q inequality constraints $g_j(\vec{x})$, the goal is to find a vector $\vec{x}^* \in \mathcal{S}$ which satisfies

$$\forall \vec{x} \in \mathcal{S} \; : \; f(\vec{x}) \geq f(\vec{x}^*) \equiv f^* \tag{1.1}$$

Then, f^* is defined as the *global minimum* and $\vec{x}^*$ is the *global minimum location*.

Due to

$$min\{f(\vec{x})\} = -max\{-f(\vec{x})\},$$

it is straightforward to convert every *minimization* problem into a *maximization* problem. Thus, without loss of generality, we shall assume a minimization problem, unless specified otherwise.

A *local minimum* $\hat{f} = f(\hat{\vec{x}})$ is defined in the following manner:

$$\exists \epsilon > 0 \quad \forall \vec{x} \in \mathcal{S} \; : \; \left\| \vec{x} - \hat{\vec{x}} \right\| < \epsilon \Rightarrow \hat{f} \leq f(\vec{x})$$

Unimodality vs. Multimodality A landscape is said to be *unimodal* if it has only a single minimum, and *multimodal* otherwise. It is called multi-global if there are several minima with equal function values as the global minimum.

Global Minimum in Practice: Characterization While there exists a general criterion for the **automatic identification** of a local minimum, such as the *zero gradient criterion*, in practice there is no equivalent general criterion for the global minimum [8]. The attempt to characterize it is essentially equivalent to posing the multimodal optimization problem and differentiating *de facto* between global and local minima. We outline here a theoretical attempt to accomplish this characterization, by means of the important concept of *level sets* [9, 10]. Given a *level set*,

$$L_f(\alpha) = \{\vec{x} | \, \vec{x} \in \mathcal{S}, \; f(\vec{x}) \leq \alpha\}, \tag{1.2}$$

it is subject to *level set mapping*, which defines its effective domain:

$$G_f = \{\alpha | \, \alpha \in \mathbb{R}, \; L_f(\alpha) \neq \emptyset\}. \tag{1.3}$$

Assuming that G_f is *compact* and *closed*, $L_f(\alpha)$ is said to be **lower semi-continuous** (*lsc*) at the point $\bar{\alpha} \in G_f$ if $\vec{x} \in L_f(\bar{\alpha})$, $\{\alpha^i\} \subset G_f$, $\{\alpha^i\} \to \bar{\alpha}$ imply the existence of $K \in \mathbb{N}$ and a sequence $\{\vec{x}^i\}$ such that $\{\vec{x}^i\} \to \vec{x}$ and $\vec{x}^i \in L_f(\alpha^i)$ for $i \geq K$.

Given this, the following is a sufficient condition for characterizing a global minimum:

Theorem 1.1.1. *Let f be a real-valued function on $\mathcal{S} \subset \mathbb{R}^n$. If every $\vec{x} \in \mathcal{S}$ satisfying $f(\vec{x}) = \bar{\alpha}$ is either a global minimum of $f(\cdot)$ on $\mathcal{S}$ or it is not a local minimum of $f(\cdot)$, then $L_f(\alpha)$ is* lsc *at $\bar{\alpha}$.*

Törn and Zilinskas concluded that the extension to multimodal domains makes the optimization problem unsolvable in the general case, i.e., there is no efficient solution technique for obtaining the global minimum value (see [8] pp. 6).

The Hessian and the Condition Number Given a real-valued *twice differentiable* n-dimensional function f, the *Hessian* matrix of $f(\vec{x})$ is defined as the matrix

$$\mathbf{H}(f(\vec{x})) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix} \tag{1.4}$$

If the second derivatives of f are all continuous, a condition which we shall assume here, the order of differentiation does not matter, and thus the *Hessian* matrix is symmetric. It is then worthwhile to introduce the *condition number* of the Hessian, a scalar which characterizes its degree of complexity, and typically determines the difficulty of a problem to be solved by optimization methods. Let $\left\{\Lambda_i^{\mathbf{H}}\right\}_{i=1}^n$ denote the *eigenvalues* of the Hessian $\mathbf{H}$, and let $\Lambda_{min}^{\mathbf{H}}$ and $\Lambda_{max}^{\mathbf{H}}$ denote its *minimal* and *maximal* eigenvalues, respectively. The *condition number of the Hessian matrix* is defined by:

$$\text{cond}(\mathbf{H}) = \frac{\Lambda_{max}^{\mathbf{H}}}{\Lambda_{min}^{\mathbf{H}}} \geq 1 \tag{1.5}$$

Ill-conditioned problems are often classified as such due to large condition numbers (e.g., 10^{14}) of the Hessian on their landscapes.

Separability Another defining property of problem difficulty is the *separability* of the objective function (see, e.g., [11]). A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *separable* if it can be optimized by solving n 1-dimensional problems separately:

$$\arg\min_{\vec{x}} f(\vec{x}) = \left(\arg\min_{x_1} f(x_1, \ldots), \ldots, \arg\min_{x_n} f(\ldots, x_n) \right)$$

1.1.2 Evolutionary Algorithms

Whereas ES and EP are similar algorithms and share many basic characteristics [12], the principal difference between them and GAs is the *encoding of*

Algorithm 1 An Evolutionary Algorithm

1: $t \leftarrow 0$
2: $P_t \leftarrow \text{Init}()$ $\{P_t \in \mathcal{S}^\mu$: Set of solutions$\}$
3: Evaluate(P_t)
4: **while** $t < t_{\max}$ **do**
5: $\quad G_t \leftarrow \text{Generate}(P_t)$ {Generate λ variations}
6: $\quad \text{Evaluate}(G_t)$
7: $\quad P_{t+1} \leftarrow \text{Select}(G_t \cup P_t)$ {Rank and select μ best}
8: $\quad t \leftarrow t + 1$
9: **end while**

the genetic information. Traditional GAs encode the genome with discrete values (as in nature), whereas ES as well as EP do that with continuous real-values. Moreover, ES and EP focused more on development of mutation operators, while in classical GA research the recombination operator received most attention. Today, GA, ES, and EP subsume under the term Evolutionary Algorithms (EAs).

Here, we offer an introductory generic description of an EA. The latter considers a *population* (i.e., set) of *individuals* (i.e., trial solutions), and models its collective learning process. Each individual in the population is initialized according to an algorithm-dependent procedure, and may carry not only a specific search point in the landscape, but also some environmental information concerning the search. A combination of stochastic as well as deterministic processes such as *mutation*, *recombination*, and *selection*, dictate the propagation in time towards successively *better* individuals, corresponding to better regimes of the landscape. The quality of an individual, or alternatively the merit of a trial solution, are determined by a so-called *fitness function*, which is typically the objective function or its rescaling. Thus, certain individuals are favored over others during the selection phase, which is based upon the fitness evaluation of the population. The selected individuals become the candidate solutions of the next generation, while the others die out.

More explicitly, an EA starts with initializing the *generation* counter t. After generating the initial population with μ individuals in $\mathcal{S}$, a set G_t of λ new solutions is generated by means of *mutation* and possibly *recombination*. The new candidate solutions are evaluated and ranked in terms of their quality (*fitness* value). The μ best solutions in $G_t \cup P_t$ are selected to form the new *parent* population P_{t+1}.

A generalized EA pseudocode is outlined in Algorithm 1.

1.2 The Standard Evolution Strategy

Evolution Strategies were originally developed at the Technical University of Berlin as a procedure for automated experimental design optimization, rather than a global optimizer for continuous landscapes. Following a sequence of successful applications (e.g., shape optimization of a bended pipe, drag minimization of a joint plate, and hardware design of a two-phase flashing nozzle), a diploma thesis [13] and a dissertation [14] laid out the solid foundations for ES as an optimization methodology. There has been extensive work on ES analysis and algorithmic design since then [7, 15, 16].

This section, which is mostly based on [1] and [7], will describe the standard ES in detail. Section 1.2.1 will introduce notation and basic terminology. Section 1.2.2 will present the $(1+1)$ algorithm, which was originally analyzed for theoretical purposes, but continued to play an important role in several aspects of Evolution Strategy design. The self-adaptation principle will be described in Section 1.2.3, while Section 1.2.4 will outline the ES algorithm.

1.2.1 Notation and Terminology

The typical application domain of Evolution Strategies is the minimization of non-linear objective functions of signature $f : \mathcal{S} \subseteq \mathbb{R}^n \to \mathbb{R}$. Given a search problem of dimension n, let $\vec{x} := (x_1, x_2, ..., x_n)^T \in \mathbb{R}^n$ denote the set of *decision parameters* or *object variables* to be optimized: It is defined as an individual associated with a trial solution. In optimization problems, which are of our main interest, it is then straightforward to define the *fitness* of that individual: It is the objective function(s) value(s) of $\vec{x}$, i.e., $f(\vec{x})$.

Evolution Strategies consider a population of candidate solutions of the given problem. This population undergoes stochastic as well as deterministic variations, with the so-called *mutation* operator, and possibly with the *recombination* operator. The mutation operator is typically equivalent to sampling a random variation from a normal distribution. Due to the continuous nature of the parameter space, the biological term *mutation rate* can be associated here with the actual size of the mutation step in the decision space, also referred to as the *mutation strength*.

Explicitly, an individual is represented by a tuple of continuous real-values, sometimes referred to as a *chromosome*, which comprises the decision parameters to be optimized, $\vec{x}$, their fitness value, $f(\vec{x})$, as well as a set of *endogenous* (i.e., evolvable) strategy parameters, $\vec{s} \in \mathbb{R}^m$.

The k^{th} individual of the population is thus denoted by:

$$\vec{a}_k = (\vec{x}_k, \vec{s}_k, f(\vec{x}_k))$$

The dimension m of the strategy parameter space is subject to the desired parameter control approach, to be discussed shortly. The endogenous pa-

rameters are a unique concept for ES, in particular in the context of the mutation operator, and they play a crucial role in the so-called *self-adaptation principle* (see Section 1.2.3).

Strategy-specific parameters, such as the population characteristic parameters μ, λ, and the so-called *mixing number* ν, are called *exogenous* strategy parameters, as they are kept constant during the simulated evolution. The mixing number determines the number of individuals involved in the application of the recombination operator.

1.2.2 Motivation: The $(1+1)$ Evolution Strategy

Rechenberg [6] considered a simple $(1+1)$ Evolution Strategy, with a fixed mutation strength σ, in order to investigate analytically two basic objective functions, namely the *corridor model* and the *sphere model*. From the historical perspective, that study laid out the foundations for the theory of Evolution Strategies.

Rechenberg derived explicitly the expressions for the *convergence rate* of his $(1+1)$ ES for the two models. By definition, neither self-adaptation nor recombination were employed in this strategy. Given the probability of the mutation operator to cover a distance k' towards the optimum, $p(k')$, the convergence rate φ is defined as the expectation of the distance k' covered by the mutation:

$$\varphi = \int_0^\infty p(k') \cdot k' \; dk' \tag{1.6}$$

The expression for the optimal step-size for the two models was first derived. It was observed to depend on the so-called *success probability* p_s,

$$p_s = \mathcal{P}\left\{f(Mutate\left\{\vec{x}\right\}) \leq f(\vec{x})\right\}. \tag{1.7}$$

By setting

$$\left.\frac{d\varphi}{d\sigma}\right|_{\sigma^*} = 0, \tag{1.8}$$

the optimal step-sizes for the two models were calculated, yielding also the optimal success probabilities. The obtained values were both close to $1/5$, regardless of the search space dimensionality. This led to the formulation of the well-known $1/5th$-success rule:

> *The ratio of successful mutations to all mutations should be* $1/5$. *If it is greater than* $1/5$, *increase the standard deviation, if it is smaller, decrease the standard deviation.*

For more details see [1]. The implementation of the $1/5th$-success rule within the (1+1)-ES is given as Algorithm 2. As practical hints, p_s can be calculated over intervals of $10 \cdot n$ trials, and the adaptation constant should be set between the boundaries $0.817 \leq c \ll 1$.

Algorithm 2 The $(1+1)$ Evolution Strategy

1: $t \leftarrow 0$
2: $P_t \leftarrow \text{Init}()$ $\{P_t \in \mathcal{S}$: Set of solutions$\}$
3: Evaluate(P_t)
4: **while** $t < t_{\max}$ **do**
5: $\quad \vec{x}(t) := Mutate\{\vec{x}(t-1)\}$ with step-size σ
6: $\quad$ Evaluate$(P'(t) := \{\vec{x}(t)\})$: $\{f(\vec{x}(t))\}$
7: $\quad Select\{P'(t) \cup P(t)\}$
8: $\quad t \leftarrow t+1$
9: $\quad$ **if** $t \mod n = 0$ **then**
10:
$$\sigma = \begin{cases} \sigma(t-n)/c & \text{if } p_s > 1/5 \\ \sigma(t-n) \cdot c & \text{if } p_s < 1/5 \\ \sigma(t-n) & \text{if } p_s = 1/5 \end{cases}$$
11: $\quad$ **else**
12: $\quad\quad \sigma(t) = \sigma(t-1)$
13: $\quad$ **end if**
14: **end while**

It should be noted that $1/5th$-success rule has been kept alive, and continued to play an important role in several aspects, including the construction of the elitist strategy of the Covariance Matrix Adaptation ES algorithm ([17] and also see Section 1.4).

1.2.3 The Self-Adaptation Principle

Section 1.2.2 provided us with the motivation to adapt the endogenous strategy parameters during the course of evolution, e.g., tuning the mutative step-size according to the $1/5th$-success rule. The basic idea of the self-adaptation principle is to consider the strategy parameters as endogenous parameters, that undergo an evolutionary process themselves. The idea of coupling endogenous strategy parameters to the object variables can be found in organisms, where self-repair mechanisms exist, such as *repair enzymes* and *mutator genes* [18]. This allows an individual to adapt to the changing environment of its trajectory in the landscape, while keeping the potentially harmful effect of mutation within reasonable boundaries. Hence, when mutative self-adaptation is applied, there is no deterministic control in the hands of the user with respect to the mutation strategy.

The crucial claim regarding ES is that self-adaptation of strategy parameters works [19]. It succeeds in doing so by applying the mutation, recombination and selection operators in the strategy, and without the use of any exogenous control. The link between strategy and decision parameters is exploited, even if it is only indirect. Experiments upon which this claim was

based had found several boosting conditions for self-adaptation to work, such as recombination on strategy parameters, selection pressure within certain bounds, and others.

1.2.4 The Canonical $(\mu/\nu \dagger \lambda)$-ES Algorithm

We describe here the specific operators for the standard Evolution Strategy, sometimes referred to as the Schwefel approach, and provide the reader with the implementation details.

Mutation

The mutation operator is the dominant variation operator within ES, and thus we choose to elaborate in this section on its characteristics. As a retrospective analysis, we choose to begin with the outline of some general rules for the design of mutation operators, as suggested by Beyer [15]:

1. **Reachability**. Given the current generation of individuals, any other search point in the landscape should be reached within a finite number of mutation operations.
2. **Unbiasedness**. Variation operators in general, and the mutation operator in particular, should not introduce any bias, and satisfy the *maximum entropy principle*. In the case of continuous unconstrained landscapes, this would suggest the use of the *normal distribution*.
3. **Scalability**. The mutation strength should be adaptive with respect to the landscape.

The ES mutation operator considers stochastic continuous variations, which are based on the *multivariate normal distribution*. Given a normally-distributed random vector, denoted by $\vec{z} = (z_1, z_2, \ldots, z_n)^T$, the mutation operator is then defined as follows:

$$\vec{x}^{NEW} = \vec{x}^{OLD} + \vec{z} \tag{1.9}$$

A *multivariate normal distribution* is uniquely defined by a *covariance matrix*, $\mathbf{C} \in \mathbb{R}^{n \times n}$, which is a symmetric positive semi-definite matrix, as well as by a *mean vector* $\vec{m} \in \mathbb{R}^n$. Its *probability density function* (PDF) is given by:

$$\Phi_{\mathcal{N}}^{pdf}(\vec{z}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{C}}} \cdot \exp\left(-\frac{1}{2}(\vec{z} - \vec{m})^T \cdot \mathbf{C}^{-1} \cdot (\vec{z} - \vec{m})\right) \tag{1.10}$$

A random vector $\vec{z}$ drawn from a multivariate normal distribution, is denoted by

$$\vec{z} \sim \mathcal{N}(\vec{m}, \mathbf{C}).$$

The ES mutation operator always considers a distribution with *zero mean*, i.e., $\vec{m} = \vec{0}$, and thus the covariance matrix $\mathbf{C}$ is the defining component of this operator. It is characterized by its $(n \cdot (n-1))/2$ covariance elements,

$$c_{ij} = cov(x_i, x_j) = cov(x_j, x_i) = c_{ji},$$

as well as by its n variances,

$$c_{ii} \equiv \sigma_i^2 = var(x_i).$$

Overall, we have,

$$\mathbf{C} = \begin{pmatrix} var(x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_n) \\ cov(x_2, x_1) & var(x_2) & \cdots & cov(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(x_n, x_1) & cov(x_n, x_2) & \cdots & var(x_n) \end{pmatrix}$$

Essentially, the $(n \cdot (n+1))/2$ independent elements of the covariance matrix are the endogenous strategy parameters that evolve along with the individual:

$$\vec{s} \leftarrow \mathbf{C},$$

i.e., the strategy parameter vector $\vec{s}$ represents the covariance matrix $\mathbf{C}$ in this case.

For the definition of the *update rule* for the strategy parameters, it is convenient to represent the off-diagonal elements of $\mathbf{C}$ by means of the rotational angles between the principal axes of the decision parameters. Let α_{ij} denote these angles,

$$c_{ij} = cov(x_i, x_j) = \frac{1}{2}(var(x_i) - var(x_j)) \cdot \tan(2\alpha_{ij}) \tag{1.11}$$

According to the *self-adaptation* principle, the covariance matrix elements also evolve every generation. The adaptation of the covariance matrix elements is dictated by non-linear update rules: The diagonal terms, $c_{ii} = \sigma_i^2$, are updated according to the *log-normal* distribution:

$$\sigma_i^{NEW} = \sigma_i^{OLD} \cdot \exp\left(\tau' \cdot \mathcal{N}(0,1) + \tau \cdot \mathcal{N}_i(0,1)\right) \tag{1.12}$$

and the off-diagonal terms are updated through the rotational angles:

$$\alpha_{ij}^{NEW} = \alpha_{ij}^{OLD} + \beta \cdot \mathcal{N}_\ell(0,1) \tag{1.13}$$

where $\mathcal{N}(0,1)$, $\mathcal{N}_i(0,1)$, and $\mathcal{N}_\ell(0,1)$ $(\ell = 1, \ldots, (n \cdot (n-1))/2)$ denote independent random variables, and where $\tau \sim 1/\sqrt{2\sqrt{n}}$, $\tau' \sim 1/\sqrt{2n}$, and $\beta = \frac{5}{180}\pi$ are constants. After those two update steps, the covariance matrix can be updated (off-diagonal terms are calculated by means of Eq. 1.11).

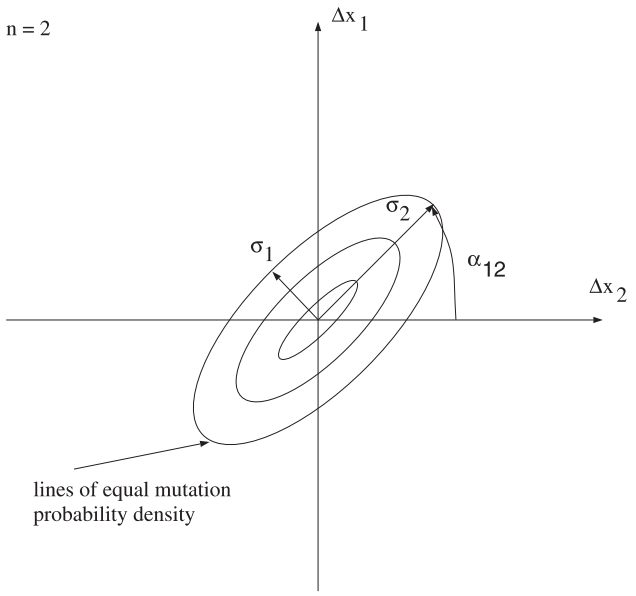

Figure 1.1: Mutation ellipsoids for $n = 2$, drawn from a general non-singular covariance matrix, with $c_{1,2} \sim \tan(2\alpha_{1,2})$. Figure courtesy of Thomas Bäck.

Geometrical Interpretation The equal probability density contour lines of a multivariate normal distribution are ellipsoids, centered about the mean. The *principal axes* of the ellipsoids are defined by the *eigenvectors* of the covariance matrix $\mathbf{C}$. The lengths of the principal axes are proportionate to the corresponding *eigenvalues*. Figure 1.1 provides an illustration for mutation ellipsoids in the case of $n = 2$.

Correlated Mutations: Strategy Considerations Given a decision parameter space of dimension n, a general mutation-control mechanism considers the *covariance matrix* $\mathbf{C}$, but may apply various different strategies, for computational considerations. There are three common approaches:

1. A covariance matrix proportionate to the identity matrix, i.e., having a single free strategy parameter σ, often referred to as the *global step-size*:
$$\mathbf{C}_1 = \sigma^2 \cdot \mathbf{I} \tag{1.14}$$

2. A diagonalized covariance matrix, i.e., having a vector of n free strategy parameters, $\left(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2\right)^T$, typically referred to as the *individual step-sizes*:
$$\mathbf{C}_2 = diag\left(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2\right) \tag{1.15}$$

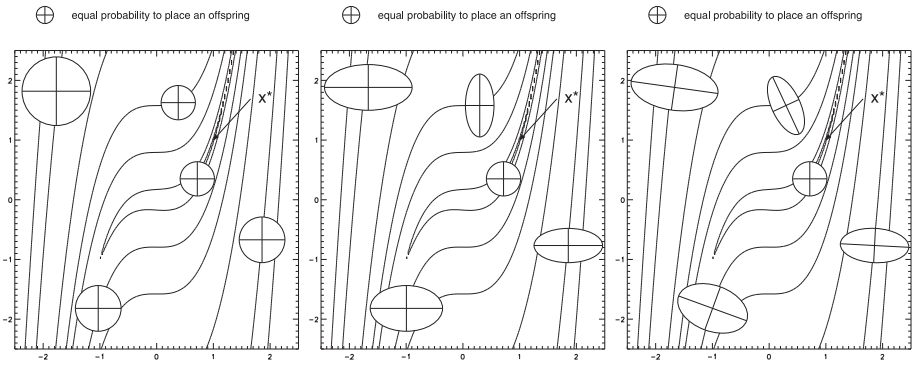

Figure 1.2: Equidensity probability contours for the three different approaches with respect to a $2D$ landscape. Left: A single *global step-size* (circles). Middle: n independent parameters (axis-parallel ellipsoids). Right: $(n \cdot (n+1))/2$ independent parameters (arbitrarily oriented ellipsoids). Figures courtesy of Thomas Bäck [20].

3. A general non-singular covariance matrix, with arbitrary $(n \cdot (n+1))/2$ free strategy parameters:
$$\mathbf{C}_3 = (c_{ij}) \tag{1.16}$$

Thus, the three approaches propose orders of $\mathcal{O}(1)$, $\mathcal{O}(n)$, or $\mathcal{O}(n^2)$ strategy parameters to be learned, respectively, at the cost of different invariance properties. Obviously, a single *global step-size* approach is very limited in its ability to generate successful moves on a generic landscape. The generalization into *individual step-sizes* assigns different variances to each coordinate axis, achieving an invariance with respect to *translation*, but still having dependency on the coordinate system (no invariance with respect to *rotation*). Finally, the most general approach with an arbitrary normal mutation distribution introduces complete invariance with respect to *translation* and *rotation*. Figure 1.2 offers an illustration for the three different approaches, on a given $2D$ landscape.

Recombination

Inspired by the organic mechanism of a *meiotic cell division*, where the genetic material is reordered by means of *crossover* between the chromosomes, the ES recombination operator considers sharing the information from up to ν parent individuals [21]. When $\nu > 2$, it is usually referred to as *multi-recombination*. Unlike other Evolutionary Algorithms (e.g., GAs), the ES recombination operator obtains only a single offspring.

Due to the continuous nature of the parameters at hand, decision as well as strategy parameters, there are two fundamental ways to recombine

parents:

- *Discrete recombination*: one of the alleles is randomly chosen among ν parents. Given a parental matrix of the old generation, $\mathbf{A}^O = \left(\vec{a}_1^O, \vec{a}_2^O, ..., \vec{a}_\nu^O\right)$, the new *recombinant* $\vec{a}^N$ is constructed by:

$$\left(\vec{a}^N\right)_i := \left(\mathbf{A}^O_{m_i}\right)_i, \quad m_i := rand\{1, .., \nu\}$$

- *Intermediate recombination*: the values of ν parents are averaged, typically with uniform weights. Essentially, this is equivalent to calculating the *centroid* of the ν parent vectors:

$$\left(\vec{a}^N\right)_i := \frac{1}{\nu}\sum_{j=1}^{\nu}\left(\vec{a}_j^O\right)_i \tag{1.17}$$

The recombination operator in the standard ES could be applied as follows:

1. For each *object variable* choose ν parents, and apply *discrete recombination* on the corresponding variables.

2. For each *strategy parameter* choose ν parents, and apply *intermediate recombination* on the corresponding variables.

It should be noted that there are no generally known best settings of the recombination operator, and the above are typical implementations of it.

Within the GA research, the *building block hypothesis* (BBH) (see, e.g., [22]) offered an explanation for the working mechanism of the crossover: The combination of good, but yet different, building blocks, i.e., specific portions of the genetic encoding from different parents, is supposed to be the key role for propagating high fitness. The debate over this hypothesis has been kept alive. In ES populations, the diversity decreases rapidly. Therefore, BBH is unlikely to fit in a similar way it does in GA populations.

On the other hand, ES research has given rise to the *genetic repair* hypothesis [23], stating that the *common* good properties of the different parents, rather than their different features, are the key role in the working mechanism of recombination. Also, recombination would typically decrease the harmful effect of mutation and would allow for high step-sizes while achieving the same convergence rates.

Selection

Natural selection is the driving force of organic evolution: Clearing-out an old generation, and allowing its individuals with the fitness advantage to increase their representation in the genetic pool of future generations. As dramatic as it might sound, *death* is an essential part in this process.

Algorithm 3 The $(\mu/\nu \stackrel{+}{,} \lambda)$ Evolution Strategy

1: $t \leftarrow 0$
2: $P_t \leftarrow \text{Init}()$ $\{P_t \in \mathcal{S}^\mu$: Set of solutions$\}$
3: Evaluate(P_t)
4: **while** $t < t_{\max}$ **do**
5: Select ν mating parents from P_t {Marriage}
6: $\vec{a}'_k(t) := Recombine\{P(t)\}\ \forall k \in \{1,\ldots,\lambda\}$ {Recombination}
7: $\vec{a}''_k(t) := Mutate\{\vec{a}'_k(t)\}\ \forall k \in \{1,\ldots,\lambda\}$ {Mutation}
8: Evaluate$(P'(t) := \{\vec{a}''_1(t),\ldots,\vec{a}''_\lambda(t)\})$ $(\{f(\vec{x}''_1(t)),\ldots,f(\vec{x}''_\lambda(t))\})$
9: **if** (μ,λ)-ES **then**
10: $Select\{P'(t)\}$
11: **else if** $(\mu+\lambda)$-ES **then**
12: $Select\{P'(t) \cup P(t)\}$
13: **end if**
14: $t \leftarrow t+1$
15: **end while**

Evolution Strategies adopt this principle, and employ *deterministic* operators in order to select the best μ individuals with the highest fitness, e.g., minimal objective function values, to be transferred into the next generation. Two selection operators are introduced in the standard ES using an elegant notation due to Schwefel. The notation characterizes the selection mechanism, as well as the number of parents and offspring involved:

- $(\mu+\lambda)$-selection: the next generation of parents will be the best μ individuals selected out of the **union** of current parents and λ offspring.
- (μ,λ)-selection: the next generation of parents will be the best μ individuals selected out of the current λ offspring.

In the case of *comma* selection, it is rather intuitive that setting $\mu < \lambda$ would be a necessary condition for an efficient convergence. In *plus* selection, however, any $\mu > 0$ can be chosen in principle. In the latter, the so-called *elitist* selection occurs, when the survival of the best individual found so far is guaranteed, leading to a possible scenario of a parent surviving for the entire process.

We are now in a position to introduce a pseudocode of the Standard Evolution Strategy (Algorithm 3).

A Note on Population Sizes One of the important topics in ES research is the study of optimal population sizes. By definition, the magnitude of λ determines the number of function evaluations per generation, which should preferably be kept small.

Typical population sizes in ES keep a ratio of $\frac{1}{7}$ between the parent and the offspring populations; a popular choice is $\mu = 15$ and $\lambda = 100$ (see, e.g., [1] and [20]).

Based on experimental observations, when individual step-sizes are chosen as strategy parameters (Eq. 1.15), λ has to scale linearly with n. In the case of arbitrary normal mutations (Eq. 1.16), Rudolph [24] showed that successful adaptation to the landscape (i.e., learning successfully the Hessian matrix) can be achieved with an upper bound of $\mu + \lambda = (n^2 + 3n + 4)/2$, but it is certainly not likely to be achieved with the typical population sizes of $\{\mu = 15,\ \lambda = 100\}$.

1.3 Derandomized Evolution Strategies (DES)

Mutative step-size control (MSC) tends to work well in the Standard-ES for the adaptation of a single global step-size (Eq. 1.14), but tends to fail when it comes to the individual step-sizes or arbitrary normal mutations (Eq. 1.15 or Eq. 1.16). Schwefel claimed that the adaptation of the strategy parameters in those cases is impossible within small populations [19], and suggested larger populations as a solution to the problem.

Due to the crucial role that the mutation operator plays within Evolution Strategies, its mutative step-size control was investigated intensively. In particular, the disruptive effects to which the MSC is subject, were studied at several levels [25, 16], and are reviewed here:

- **Indirect selection**. By definition, the goal of the mutation operator is to apply a stochastic variation to an object variable vector, which will increase its selection probability. The selection of the *strategy parameters* setting is indirect, i.e., the vector of a successful mutation is not used to adapt the step-size parameters, but rather the parameters of the distribution that led to this mutation vector.

- **Realization of parameter variation**. Due to the sampling from a random distribution, the *realization* of the parameter variation does not necessarily reflect the nature of the strategy parameters. Thus, the difference *de facto* between good and bad strategy settings of strategy parameters is only reflected in the difference between their probabilities to be selected - which can be rather small. Essentially, this means that the selection process of the strategy parameters is *strongly disturbed*.

- The *strategy parameter change rate* is defined as the difference between strategy parameters of two successive generations. Hansen and Ostermeier [16] argue that the change rate is an important factor, as it gives an indication concerning the adaptation speed, and thus it has a direct influence on the performance of the algorithm. The principal claim is that **this change rate basically vanishes** in the standard-ES.

The *change rate* depends on the *mutation strength* to which the strategy parameters are subject. While aiming at attaining the maximal change rate, the latter is underposed to an upper bound, due to the finite selection information that can be transferred between generations. Change rates that exceed the upper bound would lead to a stochastic behavior. Moreover, the mutation strength that obtains optimal change rate is typically smaller than the one that obtains good diversity among the mutants - a desired outcome of the mutation operator, often referred to as *selection difference*. Thus, the conflict between the objective of *optimal change rate* versus the objective of *optimal selection difference* cannot be resolved at the mutation strength level [25]. A possible solution to this conflict would be to unlink the change rate from the mutation strength.

The so-called *derandomized mutative step-size control* aims to treat those disruptive effects, regardless of the problem dimensionality, population size, etc.

1.3.1 $(1, \lambda)$ Derandomized ES Variants

The concept of derandomized Evolution Strategies has been originally introduced by scholars at the Technical University of Berlin in the beginning of the 1990's. It was followed by the release of a new generation of successful ES variants by Hansen, Ostermeier, and Gawelczyk [26, 27, 28, 29].

The first versions of *derandomized ES algorithms* introduced a controlled global step-size in order to monitor the individual step-sizes by decreasing the stochastic effects of the probabilistic sampling. The selection disturbance was completely removed with later versions by omitting the adaptation of strategy parameters by means of probabilistic sampling. This was combined with individual information from the last generation (the successful mutations, i.e., of selected offspring), and then adjusted to *correlated mutations*. Later on, the concept of *adaptation by accumulated information* was introduced, aiming to use wisely the past information for the purpose of step-size adaptation: Instead of using the information from the last generation only, it was successfully generalized to a weighted average of the previous generations.

Note that the different derandomized-ES variants strictly follow a $(1, \lambda)$ strategy, postponing the treatment of recombination or plus-strategies for later stages[1]. In this way, the question how to update the strategy parameters when an offspring does not improve its ancestor is not relevant here.

Moreover, the different variants hold different numbers of strategy parameters to be adapted, and this is a factor in the learning speed of the

[1]When asked about comma versus plus strategies, Hansen states that "with a good enough algorithm at hand, employing the plus strategy is unnecessary, as your algorithm should be able to revisit the best attainable solution".

optimization routine. The different algorithms hold a number of strategy parameters scaling either linearly ($\mathcal{O}(n)$ parameters responsible for individual step-sizes) or quadratically ($\mathcal{O}(n^2)$ parameters responsible for arbitrary normal mutations) with the dimensionality n of the search space.

1.3.2 First Level of Derandomization

The so-called *first level of derandomization* achieved the following desired effects:

- A degree of freedom with respect to the mutation strength of the strategy parameters.
- Scalability of the *ratio* between the change rate and the mutation strength.
- Independence of population size with respect to the adaptation mechanism.

We choose to review the implementation of the first level of derandomization through *three* particular derandomized ES variants:

DR1

The first derandomized attempt [26] coupled the successful mutations to the selection of decision parameters, and learned the mutation step-size as well as the scaling vector based upon the successful variation. The mutation step is formulated for the k^{th} individual, $k = 1, \ldots, \lambda$:

$$\vec{x}^{(g+1)} = \vec{x}^{(g)} + \xi_k \delta^{(g)} \vec{\xi}^{k}_{scal} \vec{\delta}^{(g)}_{scal} \vec{z}_k \qquad \vec{z}_k \in \{-1, +1\}^n \tag{1.18}$$

Note that $\vec{z}_k$ is a random vector of ± 1, rather than a normally distributed random vector, while $\vec{\xi}^{k}_{scal} \sim \vec{\mathcal{N}}(0,1)^+$, i.e., distributed over the positive part of the normal distribution. The evaluation and selection are followed by the adaptation of the strategy parameters (subscripts *sel* refer to the selected individual):

$$\delta^{(g+1)} = \delta^{(g)} \cdot (\xi_{sel})^{\beta} \tag{1.19}$$

$$\vec{\delta}^{(g+1)}_{scal} = \vec{\delta}^{(g)}_{scal} \cdot \left(\vec{\xi}^{sel}_{scal} + b \right)^{\beta_{scal}} \tag{1.20}$$

$\mathcal{P}\left(\xi_k = \frac{7}{5}\right) = \mathcal{P}\left(\xi_k = \frac{5}{7}\right) = \frac{1}{2}$; $\beta = \sqrt{1/n}$, $\beta_{scal} = 1/n$, $b = 0.35$, and $\xi_k \in \left\{\frac{7}{5}, \frac{5}{7}\right\}$ are constants. Note that the multiplication in Eq. 1.20 is between two vectors and carried out as element-by-element multiplication, yielding a vector of the same dimension n.

DR2

The second derandomized ES variant [27] aimed to accumulate information about the correlation or anti-correlation of past mutation vectors in order to adapt the *global step-size* as well as the *individual step-sizes* - by introducing a quasi-memory vector. This accumulated information allowed omitting the stochastic element in the adaptation of the strategy parameters - updating them only by means of successful variations, rather than with random steps.

The mutation step for the k^{th} individual, $k = 1, \ldots, \lambda$, reads:

$$\vec{x}^{(g+1)} = \vec{x}^{(g)} + \delta^{(g)} \vec{\delta}^{(g)}_{scal} \vec{z}_k \qquad \vec{z}_k \sim \vec{\mathcal{N}}(0,1) \tag{1.21}$$

Introducing a quasi-memory vector $\vec{Z}$:

$$\vec{Z}^{(g)} = c\vec{z}_{sel} + (1-c)\,\vec{Z}^{(g-1)} \tag{1.22}$$

The adaptation of the strategy parameters according to the selected offspring:

$$\delta^{(g+1)} = \delta^{(g)} \cdot \left(exp\left(\frac{\|\vec{Z}^{(g)}\|}{\sqrt{n}\sqrt{\frac{c}{2-c}}} - 1 + \frac{1}{5n} \right) \right)^{\beta} \tag{1.23}$$

$$\vec{\delta}^{(g+1)}_{scal} = \vec{\delta}^{(g)}_{scal} \cdot \left(\frac{\left|\vec{Z}^{(g)}\right|}{\sqrt{\frac{c}{2-c}}} + b \right)^{\beta_{scal}}, \qquad \left|\vec{Z}^{(g)}\right| = \left(|Z_1^{(g)}|, |Z_2^{(g)}|, ..., |Z_n^{(g)}| \right) \tag{1.24}$$

with $\beta = \sqrt{1/n}$, $\beta_{scal} = 1/n$, $b = 0.35$, and the quasi-memory rate $c = \sqrt{1/n}$ as constants. Note that the multiplication in Eq. 1.24 is between two vectors and carried out as element-by-element multiplication, yielding a vector of the same dimension n.

DR3

This third variant [28], usually referred to as the *Generation Set Adaptation* (GSA), considered the derandomization of arbitrary normal mutations for the first time, aiming to achieve invariance with respect to the scaling of variables and the rotation of the coordinate system. This naturally came with the cost of a quasi-memory matrix, $\mathbf{B} \in \mathbb{R}^{m \times n}$, setting the dimension of the strategy parameters space to $n^2 \leq m \leq 2n^2$. The adaptation of the global step-size is *mutative* with stochastic variations, just like in the **DR1**.

The mutation step is formulated for the k^{th} individual, $k = 1, \ldots, \lambda$:

$$\vec{x}^{(g+1)} = \vec{x}^{(g)} + \delta^{(g)} \xi_k \vec{y}_k \tag{1.25}$$

$$\vec{y}_k = c_m \mathbf{B}^{(g)} \cdot \vec{z}_k \qquad \vec{z}_k \sim \vec{\mathcal{N}}(0,1) \tag{1.26}$$

The update of the *memory matrix* is formulated as:

$$\begin{aligned} \mathbf{B}^{(g)} &= \left(\vec{b}_1^{(g)}, \ldots, \vec{b}_m^{(g)}\right) \\ \vec{b}_1^{(g+1)} &= (1-c) \cdot \vec{b}_1^{(g)} + c \cdot \left(c_u \xi_{sel} \vec{y}_{sel}\right), \qquad \vec{b}_{i+1}^{(g+1)} = \vec{b}_i^{(g)} \end{aligned} \tag{1.27}$$

The step-size is updated as follows:

$$\delta^{(g+1)} = \delta^{(g)} \left(\xi_{sel}\right)^{\beta} \tag{1.28}$$

where $\mathcal{P}\left(\xi_k = \frac{3}{2}\right) = \mathcal{P}\left(\xi_k = \frac{2}{3}\right) = \frac{1}{2}$; $\beta = \sqrt{1/n}$, $c_m = (1/\sqrt{m})(1 + 1/m)$, $c = \sqrt{1/n}$, $\xi_k \in \left\{\frac{3}{2}, \frac{2}{3}\right\}$, and $c_u = \sqrt{(2-c)/c}$ are constants.

1.4 The Covariance Matrix Adaptation ES

Following a series of successful derandomized ES variants addressing the first level of derandomization, and a continuous effort at the Technical University of Berlin, the so-called *Covariance Matrix Adaptation* (CMA) Evolution Strategy was released in 1996 [29], as a completely derandomized Evolution Strategy – the *fourth* generation of derandomized ES variants.

Second Level of Derandomization The so-called *second level of derandomization* targeted the following effects:

- The probability to regenerate the same mutation step is increased.
- The *change rate* of the strategy parameters is subject to explicit control.
- Strategy parameters are stationary when subject to random selection.

The second level of derandomization was implemented by means of the CMA.

The CMA combines the robust mechanism of ES with powerful *statistical learning* principles, and thus it is sometimes subject to informal criticism for not being a genuine Evolution Strategy. In short, it aims at satisfying the *maximum likelihood principle* by applying *Principle Components Analysis* (PCA) to the successful mutations, and it uses *cumulative global step-size adaptation*.

1.4.1 Preliminary

One of the goals of the CMA is to achieve a successful statistical learning process of the optimal mutation distribution, which is equivalent to **learning a covariance matrix proportional to the inverse of the Hessian matrix** (see, e.g., [30]), without calculating the actual derivatives:

$$\mathbf{C} \propto \mathbf{H}^{-1}$$

Rather than representing a mutation step with a normal variation with zero mean (Eq. 1.9), it is convenient to refer to the original notation of the normal distribution. Thus, in the notation we use here, the vector $\vec{m}$ represents the mean of the mutation distribution, but is also associated with the favorite solution at present (i.e., $\vec{x}^{OLD}$ of Eq. 1.9), σ denotes the *global step-size*, and the covariance matrix $\mathbf{C}$ determines the shape of the distribution ellipsoid:

$$\vec{x}^{NEW} \sim \mathcal{N}(\vec{m}, \sigma^2\mathbf{C}) = \vec{m} + \sigma \cdot \mathcal{N}(\vec{0}, \mathbf{C}) = \vec{m} + \sigma \cdot \vec{z}$$

Different principles dictate the adaptation of the covariance matrix, $\mathbf{C}$, versus the adaptation of the global step-size σ:

- The mean $\vec{m}$ and the covariance matrix $\mathbf{C}$ of the normal distribution are updated according to the *maximum likelihood principle*, such that good mutations are likely to appear again. $\vec{m}$ is updated such that

  $$\mathcal{P}\left(\vec{x}_{sel} | \mathcal{N}\left(\vec{m}, \sigma^2\mathbf{C}\right)\right) \longrightarrow \max$$

  and $\mathbf{C}$ is updated such that

  $$\mathcal{P}\left(\left.\frac{\vec{x}_{sel} - \vec{m}_{old}}{\sigma}\right| \mathcal{N}\left(\vec{0}, \mathbf{C}\right)\right) \longrightarrow \max$$

  considering the prior $\mathbf{C}$. This is implemented through the so-called *Covariance Matrix Adaptation* (CMA) mechanism.

- σ is updated such that it is conjugate perpendicular to the consecutive steps of $\vec{m}$. This is implemented through the so-called *Cumulative Step-size Adaptation* (CSA) mechanism.

The Evolution Path

The most intuitive way to update the covariance matrix would be to construct an $n \times n$ *matrix analogue* to the **DR2** mechanism (see Eq. 1.22), with the *outer-product* of the selected mutation vector $\vec{z}_{sel}$:

$$\mathbf{C} \longleftarrow (1 - c_{cov})\mathbf{C} + c_{cov}\vec{z}_{sel}\vec{z}_{sel}^T$$

However, the sign information of $\vec{z}_{sel}$ is lost due to $\vec{z}_{sel}\vec{z}_{sel}^T = -\vec{z}_{sel}\left(-\vec{z}_{sel}\right)^T$. The solution lies within the definition of the so-called *evolution path*, which accumulates the history information using an *exponentially weighted moving average*:

$$\vec{p}_c \propto \sum_{i=0}^{g} (1 - c_c)^{g-i} \vec{z}_{sel}^{(i)}$$

And now the covariance matrix adaptation step reads:

$$\mathbf{C} \longleftarrow (1 - c_{cov})\mathbf{C} + c_{cov}\vec{p}_c\vec{p}_c^T$$

The Path Length Control

The covariance matrix update is not likely to increase the variance in all directions simultaneously, and thus a global step-size control is much needed. The basic idea of the so-called *path length control* is to measure the length of the evolution path, which is also the consecutive steps of $\vec{m}$, and adapt the step-size according to the following argument: If the *evolution path* is longer than expected, the steps are likely parallel, and thus the step-size should be increased; Alternatively, if it is shorter than expected, the steps are probably anti-parallel, and the step-size should be decreased accordingly. The expected length is defined in a straightforward manner as the expected length of a normally distributed random vector.

The actual measurement is done by means of the "conjugate" evolution path:

$$\vec{p}_\sigma \propto \sum_{i=0}^{g} (1-c_\sigma)^{g-i} \mathbf{C}^{(i)\ -\frac{1}{2}}\ \vec{z}_{sel}^{(i)}$$

where the factorization of $\mathbf{C}$ is required in order to align all directions within the *rotated frame*. Then, the update of the step-size depends on the comparison between $\|\vec{p}_\sigma\|$ and the expected length of a normally distributed random vector, $E\left[\|\mathcal{N}\left(0,\mathbf{I}\right)\|\right]$:

$$\sigma \longleftarrow \sigma \cdot \exp\left(\frac{\|\vec{p}_\sigma\|}{E\left[\|\mathcal{N}\left(0,\mathbf{I}\right)\|\right]} - 1\right)$$

1.4.2 The $(1,\lambda)$ Rank-One CMA

We are now in a position to introduce the explicit formulation of the *rank-one update with cumulation* Covariance Matrix Adaptation Evolution Strategy, following the notation introduced in Section 1.4.1. Additionally, consider the *diagonalization* of the covariance matrix, denoted by

$$\mathbf{C}^{(g)} = \mathbf{B}^{(g)}\mathbf{D}^{(g)}\left(\mathbf{B}^{(g)}\mathbf{D}^{(g)}\right)^T \tag{1.29}$$

where $\mathbf{B}^{(g)}$ is an orthonormal rotation matrix which defines the coordinate system, and $\mathbf{D}^{(g)} = diag\left(\sqrt{\Lambda_1}, \sqrt{\Lambda_2}, ..., \sqrt{\Lambda_n}\right)$ holds the square-roots of the eigenvalues.

The mutation step for the k^{th} individual, $k = 1, \ldots, \lambda$, is then defined as:

$$\vec{x}_k^{(g+1)} = \vec{x}^{(g)} + \sigma^{(g)}\mathbf{B}^{(g)}\mathbf{D}^{(g)}\vec{z}_k^{(g+1)} \tag{1.30}$$

with $\vec{z}_k \sim \vec{\mathcal{N}}\left(\vec{0},\mathbf{I}\right)$.

The *evolution path*, initialized $\vec{p}_c^{(0)} = \vec{0}$, is explicitly updated as follows:

$$\vec{p}_c^{(g+1)} = (1-c_c)\cdot\vec{p}_c^{(g)} + \sqrt{c_c(2-c_c)}\cdot\mathbf{B}^{(g)}\mathbf{D}^{(g)}\vec{z}_{sel}^{(g+1)} \tag{1.31}$$

and then the covariance matrix, initialized as identity $\mathbf{C}^{(0)} = \mathbf{I}$, is adapted accordingly:

$$\mathbf{C}^{(g+1)} = (1 - c_{cov}) \cdot \mathbf{C}^{(g)} + c_{cov} \cdot \vec{p}_c^{(g+1)} \left(\vec{p}_c^{(g+1)}\right)^T \tag{1.32}$$

The calculation of the "conjugate" evolution path, initialized $\vec{p}_\sigma^{(0)} = \vec{0}$, reads:

$$\vec{p}_\sigma^{(g+1)} = (1 - c_\sigma) \cdot \vec{p}_\sigma^{(g)} + \sqrt{c_\sigma(2 - c_\sigma)} \cdot \mathbf{B}^{(g)} \vec{z}_{sel}^{(g+1)} \tag{1.33}$$

and then followed by the update of the global step-size:

$$\sigma^{(g+1)} = \sigma^{(g)} \cdot exp\left(\frac{c_\sigma}{d_\sigma} \cdot \left(\frac{\left\|\vec{p}_\sigma^{(g+1)}\right\|}{E\left[\|\mathcal{N}(0, \mathbf{I})\|\right]} - 1\right)\right) \tag{1.34}$$

The various learning coefficients are typically set as $c_c = 4/(n+4)$, $c_{cov} = 2/(n+1.4)^2$, $c_\sigma = 3/(n+4)$, and $d_\sigma = 1 + c_\sigma$. The expectation of the length of a normally distributed random vector is given by:

$$E\left[\|\mathcal{N}(0, \mathbf{I})\|\right] = \sqrt{2} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \tag{1.35}$$

where the Gamma function is defined by:

$$\Gamma(n) = \int_0^\infty x^{n-1} \exp(-x) dx \tag{1.36}$$

but may also be approximated by $E\left[\|\mathcal{N}(0, \mathbf{I})\|\right] \approx \sqrt{n}\left(1 - \frac{1}{4n} + \frac{1}{21n^2}\right)$.

Implementation Additional implementation remarks are outlines here:

- Arnold offered[2] a dramatic simplification to the global step-size update (Eq. 1.34) with replacing $\left(\frac{\left\|\vec{p}_\sigma^{(g+1)}\right\|}{E[\|\mathcal{N}(0,\mathbf{I})\|]} - 1\right)$ by $\left(\frac{\left\|\vec{p}_\sigma^{(g+1)}\right\|^2 - n}{2n}\right)$. This was reported to perform equally well [16].
- The update of the *evolution path* (Eq. 1.31) is usually implemented with a conditional threshold as follows:

$$\vec{p}_c^{(g+1)} = (1 - c_c) \cdot \vec{p}_c^{(g)} + H_\sigma^{(g+1)} \sqrt{c_c(2 - c_c)} \cdot \mathbf{B}^{(g)} \mathbf{D}^{(g)} \vec{z}_{sel}^{(g+1)} \tag{1.37}$$

$$H_\sigma^{(g+1)} = \begin{cases} 1 & \text{if } \frac{\left\|\vec{p}_\sigma^{(g+1)}\right\|}{\sqrt{1-(1-c_\sigma)^2}} < H_{thresh} \\ 0 & \text{otherwise} \end{cases} \tag{1.38}$$

where $H_{thresh} = \left(1.5 + \frac{1}{n-0.5}\right) \cdot E\left[\|\mathcal{N}(0, \mathbf{I})\|\right]$.

[2] Hansen et al. cite this source of information as *personal communications*.

1.4.3 The (μ_W, λ) Rank-μ CMA

The Rank-μ Covariance Matrix Adaptation [31] is an extension of the original update rule for larger population sizes. The idea is to use $\mu > 1$ vectors in order to update the covariance matrix $\mathbf{C}$ in each generation, based on *weighted intermediate recombination*.

Let $\vec{x}_{i:\lambda}$ denote the i^{th} ranked solution point, such that

$$f\left(\vec{x}_{1:\lambda}\right) \leq f\left(\vec{x}_{2:\lambda}\right) \leq \cdots \leq f\left(\vec{x}_{\lambda:\lambda}\right)$$

The updated *mean* is now defined as follows:

$$\vec{m} \leftarrow \sum_{i=1}^{\mu} w_i \vec{x}_{i:\lambda} = \vec{m} + \sigma \sum_{i=1}^{\mu} w_i \vec{z}_{i:\lambda} \equiv \langle \vec{x} \rangle_W$$

with a set of weights:

$$w_1 \geq w_2 \geq \cdots \geq w_\mu > 0, \qquad \sum_{i=1}^{\mu} w_i = 1$$

Essentially, this is a generalization of the *intermediate recombination* concept (Eq. 1.17), suggested by Rechenberg[3].

By setting $\forall i : w_i = \frac{1}{\mu}$, the original recombination is restored, which is then noted by (μ_I, λ) (note, however, that the $(\mu/\mu_I, \lambda)$ notation is also used [32]).

The covariance matrix update can now be formalized by means of rank-μ update, using an *outer-product of the weighted mutation vectors*:

$$\mathbf{C} \longleftarrow (1 - c_{cov})\mathbf{C} + c_{cov} \sum_{i=1}^{\mu} w_i \vec{z}_{i:\lambda} \vec{z}_{i:\lambda}^T$$

It can be even furthermore combined with the rank-one update:

$$\mathbf{C} \longleftarrow (1 - c_{cov})\mathbf{C} + \frac{c_{cov}}{\mu_{cov}} \vec{p}_c \vec{p}_c^T + c_{cov} \left(1 - \frac{1}{\mu_{cov}}\right) \sum_{i=1}^{\mu} w_i \vec{z}_{i:\lambda} \vec{z}_{i:\lambda}^T$$

We shall now present the (μ_W, λ) rank-μ CMA characteristic equations:

$$\vec{x}_k^{(g+1)} = \langle \vec{x} \rangle_W^{(g)} + \sigma^{(g)} \mathbf{B}^{(g)} \mathbf{D}^{(g)} \vec{z}_k^{(g+1)}, \qquad k = 1, \ldots, \lambda \tag{1.39}$$

$$\vec{p}_c^{(g+1)} = (1 - c_c) \cdot \vec{p}_c^{(g)} + \sqrt{c_c(2 - c_c)} \cdot c_W \mathbf{B}^{(g)} \mathbf{D}^{(g)} \langle \vec{z} \rangle_W^{(g+1)} \tag{1.40}$$

$$\mathbf{C}^{(g+1)} = (1 - c_{cov}) \cdot \mathbf{C}^{(g)} + \frac{c_{cov}}{\mu_{cov}} \cdot \vec{p}_c^{(g+1)} \left(\vec{p}_c^{(g+1)}\right)^T + c_{cov} \left(1 - \frac{1}{\mu_{cov}}\right) \sum_{i=1}^{\mu} w_i \vec{x}_{i:\lambda} \vec{x}_{i:\lambda}^T \tag{1.41}$$

[3] Reported as *personal communications* between Hansen, Ostermeier and Rechenberg.

$$\vec{p}_\sigma^{(g+1)} = (1 - c_\sigma) \cdot \vec{p}_\sigma^{(g)} + \sqrt{c_\sigma(2 - c_\sigma)} \cdot \mathbf{B}^{(g)} c_W \langle \vec{z} \rangle_W^{(g+1)} \tag{1.42}$$

$$\sigma^{(g+1)} = \sigma^{(g)} \cdot exp\left(\frac{c_\sigma}{d_\sigma} \cdot \left(\frac{\left\| \vec{p}_\sigma^{(g+1)} \right\|}{E\left[\| \mathcal{N}(0, \mathbf{I}) \| \right]} - 1 \right) \right) \tag{1.43}$$

The weights are typically set to:

$$w_{i=1\ldots\mu} = \frac{\ln(\mu + 1) - \ln(i)}{\sum_{j=1}^{\mu} \ln(\mu + 1) - \ln(j)} \tag{1.44}$$

The constant c_W is defined such that $c_W \langle \vec{z} \rangle_W^{(g+1)}$ and $\vec{z}_k^{(g+1)}$ are identically distributed with the same variance under random selection:

$$c_W = \frac{\sum_{i=1}^{\mu} w_i}{\sqrt{\sum_{i=1}^{\mu} w_i^2}} \tag{1.45}$$

The special *rank-μ* constant, μ_{cov}, is the *variance effective selection mass*:

$$\mu_{cov} = \frac{1}{\sum_{i=1}^{\mu} w_i^2} \tag{1.46}$$

which becomes $\mu_{cov} = \mu$ in the special case of (μ_I, λ).

The rest of the constants are set as in the $(1, \lambda)$ *rank-one CMA*.

Population Size Given a search space of dimension n, the default CMA population sizes introduced a *revolutionary order of magnitude* into the ES field, $\mathcal{O}(\log(n))$, especially when we take into account the goal to learn the full covariance matrix of the decision parameters space.

The explicit suggested values are as follows:

$$\lambda = 4 + \lfloor 3 \cdot \ln(n) \rfloor \qquad \mu = \lfloor \lambda / 2 \rfloor \tag{1.47}$$

1.4.4 The $(1 + \lambda)$ CMA

This elitist version [17] of the CMA-ES algorithm, which had been originally derived for the sake of a *multi-objective* CMA algorithm [33], combined the classical concept of the $(1 + 1)$ ES strategy, and in particular the *success probability* and *success rule* components (see Eq. 1.7 as well as Section 1.2.2), with the Covariance Matrix Adaptation concept. The so-called *success rule based step size control* replaces the *path length control* of the CMA-comma strategy. The same notation as in Section 1.4.2 is used here:

$$\vec{x}_k^{(g+1)} = \vec{x}^{(g)} + \sigma^{(g)} \mathbf{B}^{(g)} \mathbf{D}^{(g)} \vec{z}_k^{(g+1)}, \qquad k = 1, \ldots, \lambda \tag{1.48}$$

After the evaluation of the new generation, the success rate is updated $p_{succ} = \lambda_{succ}^{(g+1)} / \lambda$, where:

$$\bar{p}_{succ} = (1 - c_p) \cdot \bar{p}_{succ} + c_p \cdot p_{succ} \tag{1.49}$$

$$\sigma^{(g+1)} = \sigma^{(g)} \cdot \exp\left(\frac{1}{d} \cdot \left(\bar{p}_{succ} - \frac{p_{succ}^{target}}{1 - p_{succ}^{target}}\left(1 - \bar{p}_{succ}\right)\right)\right) \tag{1.50}$$

The covariance matrix is updated only if the selected offspring is better than the parent. Then,

$$\vec{p}_c = \begin{cases} \left(1 - c_c\right)\vec{p}_c + \sqrt{c_c\left(2 - c_c\right)} \cdot \frac{\vec{x}_{sel}^{(g+1)} - \vec{x}^{(g)}}{\sigma_{parent}^{(g)}} & \text{if } \bar{p}_{succ} < p_\Theta \\ \left(1 - c_c\right)\vec{p}_c & \text{otherwise} \end{cases} \tag{1.51}$$

$$\mathbf{C}^{(g+1)} = \begin{cases} \left(1 - c_{cov}\right) \cdot \mathbf{C}^{(g)} + c_{cov} \cdot \vec{p}_c \vec{p}_c^{T} & \text{if } \bar{p}_{succ} < p_\Theta \\ \left(1 - c_{cov}\right) \cdot \mathbf{C}^{(g)} + c_{cov} \cdot \left(\vec{p}_c \vec{p}_c^{T} + c_c\left(2 - c_c\right)\mathbf{C}^{(g)}\right) & \text{otherwise} \end{cases} \tag{1.52}$$

The default parameters are set as follows: $d = 1 + \frac{n}{2}$, $p_{succ}^{target} = \frac{2}{11}$, $c_p = \frac{1}{12}$, $c_c = \frac{2}{n+2}$, $c_{cov} = \frac{2}{n^2+6}$, and $p_\Theta = 0.44$.

A Note on Usage As mentioned earlier, this plus-strategy version was constructed for multi-objective optimization. Unofficially, it is not recommended to use it otherwise. In this work, we will restrict the use of the CMA+ to the niching framework exclusively, and thus will not consider it upon the employment of the DES variants to single-criterion Quantum Control optimization tasks in Chapter 7.

1.4.5 Constraints Handling

The broad topic of *constraints handling* [34] is certainly not of a major concern in this study, but it does have an indirect impact on the niching techniques to be introduced here, as will become more clear in the following chapters. We thus choose to specify here, in short, the general approach to handle constraints when derandomized-ES are in use, in light of the rule of thumb suggested by Hansen and Ostermeier for the CMA (see [16], pp. 21).

A possible way to handle constraints would be to repeat the generation step (e.g., Eq. 1.30) until λ, or at least μ *feasible solutions* are generated. This should be strictly enforced, before the following update equations are applied. It is claimed that this method should perform in a satisfying manner, if a sufficient number of feasible solutions are initially generated - due to the symmetry of the mutation distribution. However, if the global minimum is located at the edge of the feasible domain, it is suggested that other constraints handling techniques should be used.

1.4.6 Discussion

The Covariance Matrix Adaptation Evolution Strategy is a state-of-the-art optimization routine, which combines *classical deterministic concepts* (e.g.,

Hessian or Covariance matrices learning) and *statistical learning tools* (e.g., Principal Components Analysis) with the powerful stochastic mechanism of Evolution Strategies. In terms of standard performance criteria, it was ranked as the best Evolutionary Algorithm at hand [35].

The CMA-ES has been informally criticized for not being a genuine evolution strategy, since it incorporates those non-evolutionary components. Even as such, and despite its considerable success-rate as a global optimizer, we would like to stress that it certainly has a nature of a local search routine. The fact that it learns a unimodal distribution in the search space - no matter how well it does so - makes it a local search. We believe that this provides us with some motivation to use the CMA-ES, as well as other derandomized-ES routines, as algorithmic kernels for a multi-distribution approach - which would construct a niching algorithm. The idea would be essentially to use multiple CMAs in parallel, aiming to achieve a good coverage of the landscapes with local-searchers. This idea would become more clear in the next chapter, when we introduce the *gateway to niching*.

The genes are the master programmers, and they are programming for their lives. They are judged according to the success of their programs in coping with all the hazards that life throws at their survival machines, and the judge is the ruthless judge of the court of survival.

The Selfish Gene; Richard Dawkins

Chapter 2

Introduction to Niching

2.1 Speciation Theory vs. Conceptual Designs

Evolutionary Algorithms have the tendency to lose diversity within their population of feasible solutions and to converge into a single solution [1, 36, 37], even if the search landscape has multiple globally optimal solutions.

Niching methods, the extension of EAs to finding multiple optima in multi-modal optimization within one population, address this issue by maintaining the diversity of certain properties within the population. Thus, they aim at obtaining parallel convergence into multiple basins of attraction in a multi-modal landscape within a single run.

The study of niching is challenging both from the **theoretical** point of view and from the **practical** point of view. The theoretical challenge is two-fold - maintaining the diversity within a population-based stochastic algorithm from the computational perspective, but also having an insight into *speciation* theory or *population genetics* from the Evolutionary Biology perspective. The practical aspect provides a real-world incentive for this problem - there is an increasing interest of the *applied optimization community* in providing the decision maker with multiple solutions which ideally represent different conceptual designs, for single-criterion or multi-criterion search spaces [38, 39]. The concept of *"going optimal"* is often extended now into the aim for *"going multi-optimal"*, so to speak: **Obtaining optimal results but also providing the decision maker with different choices**. On this particular note, it is worth mentioning the so-called *Second Toyota Paradox* [40]:

> *"Delaying decisions, communicating ambiguously, and pursuing an excessive number of prototypes, can produce better cars faster and cheaper."*

Niching methods have been studied in the past 35 years, mostly in the context of Genetic Algorithms, and the focus has been mainly on the theoretical aspect. As will be discussed here, niching methods have been mostly

a by-product of studying *population diversity*, and were hardly ever at the front of the EC research.

This chapter, the *gateway to niching*, discusses a variety of introductory topics - ranging from biological aspects of diversity and speciation, mathematical definitions of basins of attraction, to GA niching methods - which reflect the strong interdisciplinary nature of this subject.

2.2 From DNA to Organic Diversity

In this section we introduce the *biological* elementary concepts that correspond to the core of niching methods: *population diversity*. This section is mainly based on [41] and personal lecture notes[1].

A Preliminary Note on Terminology A species is defined as the *smallest evolutionary independent unit*. The term *niche*, however, stems from ecology, and it has several different definitions. It is sometimes referred to as the collective environmental components which are favored by a specific species, but could also be considered as the ecosystem itself which hosts individuals of *various species*. Most definitions would typically also consider the *hosting capacity* of the niche, which refers to the limited available resources for sustaining life in its domain.

In the context of function optimization, *niche* is associated with a *peak*, or a basin of attraction, whereas a *species* corresponds to the subpopulation of individuals occupying that *niche*.

2.2.1 Genetic Drift

Organic evolution can be broken down into four defining fundamental mechanisms: *natural selection*, *mutation*, *migration* or *gene flow*, and *genetic drift*. The latter, which essentially refers to *sampling errors in finite populations*, was overlooked by Darwin, who had not been familiar with Mendelian genetics, and thus did not discuss this effect in his "Origin of Species" [42]. In short, *genetic drift* is a stochastic process in which the diversity is lost in finite populations. A distribution of genetic properties is transferred to the next generation in a limited manner, due to the finite number of generated offspring, or equivalently the limited statistical sampling of the distribution. As a result, the distribution is likely to approach an *equilibrium distribution*, e.g., fixation of specific alleles when subject to equal fitness. This is why *genetic drift* is often considered as a *neutral effect*. The smaller the population, the faster and stronger this effect occurs. An analogy is occasionally drawn between genetic drift to *Brownian motion* of particles in mechanics.

[1] Notes were taken in the course "Evolutionary Biology" of Prof. David Stern (EEB309), Princeton University, Fall 2007.

In order to demonstrate the genetic drift effect, we conducted simulations[2] on the following basic model of population genetics: The evolution of random-mating populations with *two alleles*, namely, **A** and **a**, equal fitnesses of the *three genotypes* (i.e., no preferences for **AA**, **Aa**, nor **aa**), no mutations, no migration between the replicate populations, and finite population size N. We simulated ten simultaneously evolving populations, for three test-cases of population sizes: $N_1 = 10$, $N_2 = 100$, and $N_3 = 1000$. Figure 2.1 offers an illustration for the three different simulations. It is easy to observe a clear trend in this simple experiment: Alleles' loss/fixation is very likely to occur in small population sizes, and is not likely to occur in large population sizes.

The *genetic drift* effect had been originally recognized by R.A. Fisher [43] (referred to as *random survival*), and was explicitly mentioned by S. Wright when studying Mendelian populations [44]. It was, however, revisited and given a new interpretation in the *Neutral Theory of Molecular Evolution* of Kimura [45]. The *Neutral Theory* suggested that the *random genetic drift* effect is the main driving force within molecular evolution, rather than the *non-random natural selection* mechanism. *Natural selection* as well as *genetic drift* are considered nowadays, by the contemporary evolutionary biology community, as the combined driving force of organic evolution. Moreover, the importance of the *Neutral Theory* is essentially in its being a **null hypothesis model** for the *Natural Selection Theory* - by definition.

2.2.2 Organic Diversity

Diversity among individuals or populations in nature can be attributed to different evolutionary processes which occur at different levels. We distinguish here between variations that are observed within a single species to a *speciation* process, during which a new species arises, and review shortly both of them.

Variations within a Species Diversity of organisms within a single species stems from variance at the genotypic level, referred to as *genetic diversity*, or from the existence of spectrum of phenotypic realizations to a specific genotype. These effects are quantified and are usually associated with *genotypic variance* and *phenotypic variance*, respectively. Several hypotheses explaining *genetic diversity* have been proposed within the discipline of *population genetics*, including the *neutral evolution theory*. It should be noted that genetic diversity is typically considered to be advantageous for survival, as it may allow better adaptation of the population to environmental changes, such as climate variations, diseases, etc.

Phenotypic variance is measured on a continuous spectrum, also known

[2]Simulations were conducted with the *PopG Genetic Simulation Program*, version 3.1.

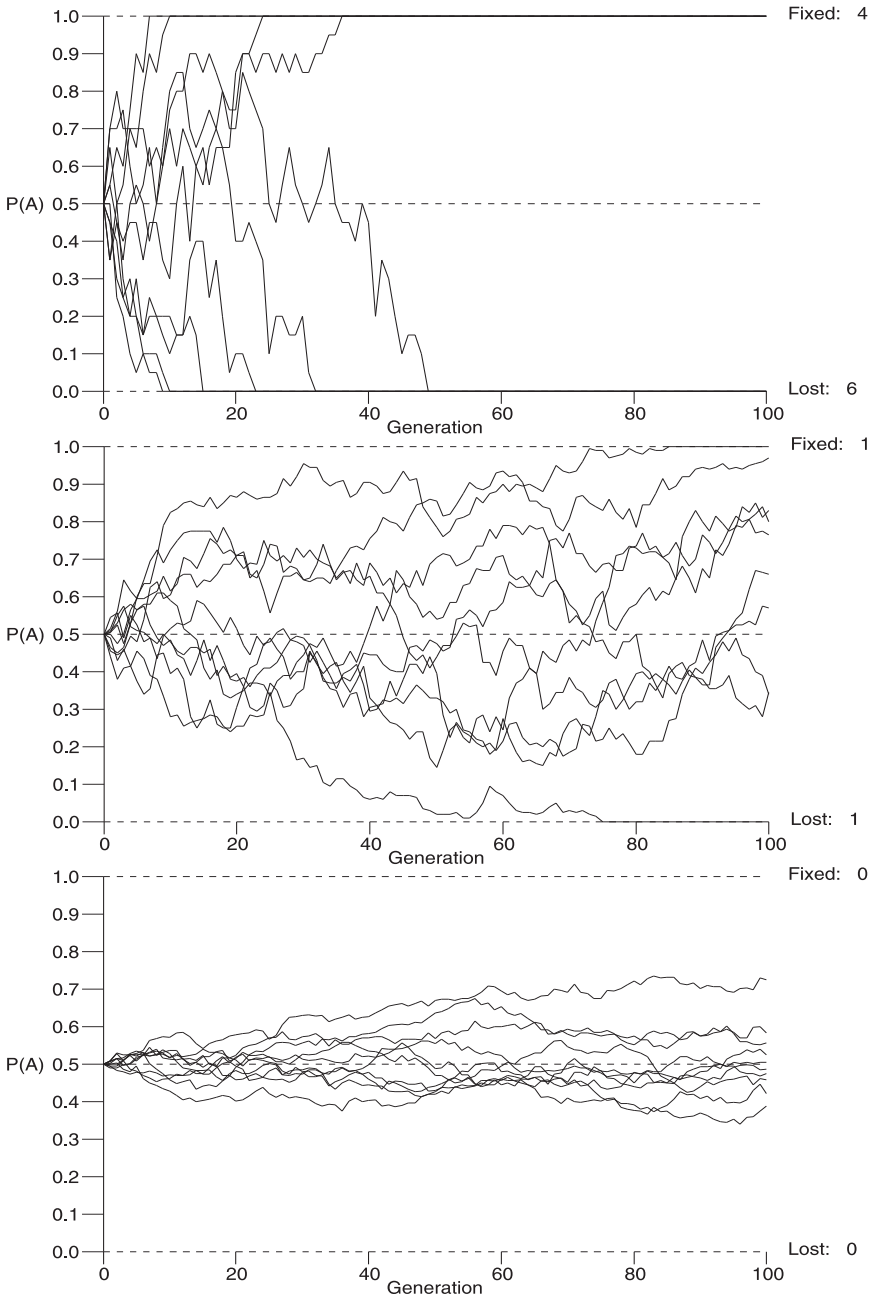

Figure 2.1: Ten simultaneously evolving populations, for three test-cases of population sizes: $N_1 = 10$ [TOP], $N_2 = 100$ [CENTER], and $N_3 = 1000$ [BOTTOM]. The vertical axis corresponds to the allele frequency of **A** in the population, as a function of generations, indicated on the horizontal axis.

as quantitative variation. Roughly speaking, the main sources of quantitative variations [41, 46] are outlined here:

1. Genes have *multiple loci*, and hence are mapped into a large set of phenotypes.

2. *Environmental effects* have direct influence on natural selection; fitness is time-dependent, and thus phenotypic variations in the outcome of selection are expected.

3. *Phenotypic plasticity* is the amount in which the *genotypic expression* vary in different environments[3], and it is a direct source of variation at the phenotypic level.

4. The plastic response of the genotype to the environment, i.e., the joint effect of genetic and environmental elements, also affects the selection of a specific phenotype, and thus can lead to variations. This effect is known as *Genotype-Environment Interaction* ("G-by-E").

Thus, *quantitative variations* are mainly caused by genotypic and phenotypic realizations and their interaction with the environment. The ratio between *genetic variance* to total *phenotypic variance* is defined as *heritability* [44].

Speciation The essence of the speciation process is **lack of gene flow**, where physical isolation often plays the role of the barrier to gene flow. Lack of gene flow is only one of the necessary conditions for speciation. Another necessary condition for speciation to occur is that the reduction of gene flow will be followed by a phase of **genetic divergence**, by means of *mutation*, *selection*, and *drift*. Finally, the completion or elimination of divergence can be assessed via the so-called *secondary contact* phase: interbreeding between the parental populations would possibly fail (offspring is less fit), succeed (offspring is fitter), or have a neutral outcome (offspring has the same fitness). This would correspond respectively to increasing, decreasing or stabilizing the differentiation between the two arising species. Note that the speciation can occur de facto, without the actual secondary contact taking place; the latter is for observational assessment purposes.

In organic evolution, four different levels of *speciation* are considered, corresponding to four levels of physical linkage between the subpopulations:

1. **Allopatric speciation** The split in the population occurs only due to complete geographical separation, e.g., migration or mountain building. It results in two geographically isolated populations.

[3] Bradshaw [47] gave the following qualitative definition to phenotypic plasticity: "The amount by which the expressions of individual characteristics of a genotype are changed by different environments is a measure of the plasticity of these characters".

2. **Peripatric speciation** Species arise in small populations which are not geographically separated but rather isolated in practice; the effect occurs mainly due to the *genetic drift* effect.

3. **Parapatric speciation** The geographical separation is limited, with a physical overlap between the two zones where the populations split from each other.

4. **Sympatric speciation** The two diverging populations coexist in the same zone, and thus the speciation is strictly non-geographical. This is observed in nature in parasite populations, that are located in the same zone, but associated with different plant or animal hosts [48].

These four modes of speciation correspond to four levels of geographically decreasing linkages. Roughly speaking, *statistical association* of genetic components in nature, such as *loci*, typically results from *physical linkage*. In this case, we claim that statistical disassociation, which is the trigger to speciation, originates from gradually decreasing physical linkage.

In summary, speciation typically occurs throughout three steps:

1. Geographic isolation or reduction of gene flow.

2. Genetic divergence (mutation, selection, drift).

3. Secondary contact (observation/assessment).

2.3 "Ecological Optima": Basins of Attraction

We devote this section to the definition of basins of attraction. This section is mainly based on Törn and Zilinskas [8].

The task of defining a *generic basin of attraction* seems to be one of the most difficult problems in the field of *global optimization*, and there have only been few attempts to treat it theoretically[4] [8].

Rigorously, it is possible to define the basin by means of a *local optimizer*. In particular, consider a gradient descent algorithm starting from $\vec{x}_0$, which is characterized by the following dynamics:

$$\frac{d\vec{x}(t)}{dt} = -\nabla f\left(\vec{x}(t)\right) \tag{2.1}$$

with the initial condition $\vec{x}\,(0) = \vec{x}_0$. Now, consider the set of points for which the limit exists:

$$\Upsilon = \left\{ \vec{x} \in \mathbb{R}^n \,\middle|\, \vec{x}(0) = \vec{x} \wedge \vec{x}(t)|_{t\geq 0} \text{ satisfies Eq. 2.1} \wedge \lim_{t\to\infty} \vec{x}(t) \text{ exists} \right\} \tag{2.2}$$

[4] Intuitively, and strictly metaphorically speaking, we may think of a *region of attraction* of $\vec{x}_L$ as the region, where if water is poured, it will reach $\vec{x}_L$. Accordingly, we may then think of the basin of $\vec{x}_L$ as the maximal region that will be covered when the cavity at $\vec{x}_L$ is filled to the lowest part of its rim.

Definition 2.3.1. The *region of attraction* $A(\vec{x}_L)$ of a local minimum $\vec{x}_L$ is

$$A(\vec{x}_L) = \left\{ \vec{x} \in \Upsilon \,\middle|\, \vec{x}(0) = \vec{x} \wedge \vec{x}(t)|_{t \geq 0} \text{ satisfies Eq. 2.1} \wedge \lim_{t \to \infty} \vec{x}(t) = \vec{x}_L \right\}. \tag{2.3}$$

The *basin* of $\vec{x}_L$ is the **maximal level set** that is fully contained in $A(\vec{x}_L)$.

In the case of several disconnected local minima with the same function value, it is possible to define the region of attraction as the union of the non-overlapping connected sets.

2.3.1 Classification of Optima: The Practical Perspective

On the note of the theoretical definition of the basin, it is worth mentioning the practical perspective for the *classification of optima shapes*, also referred to as global topology. This topic is strongly related to the emerging subfield of *robustness study* (see, e.g., [49]), which aims at attaining high-yield optima with large basins (i.e., low partial derivative values in the proximity of the peak). Moreover, yet visited from a different direction, another approach was introduced recently by Lunacek and Whitley for classifying different classes of multimodal landscapes with respect to algorithmic performance [50]. The latter defines the *dispersion metric* of a landscape as the degree to which the local optima are globally clustered near one another. Landscapes with low dispersion have their best local optima clustered together in a single *funnel*[5]. This classification to low dispersion versus high dispersion may be associated with the algorithmic trade-off between exploration of the landscape and exploitation of local structures. In the broad context of this work, it is interesting to note that the CMA was shown in [50] to perform well on low-dispersion landscapes, and was less efficient on high-dispersion landscapes.

2.4 Population Diversity within EAs

The term *population diversity* is commonly used in the context of Evolutionary Algorithms, but it rarely refers to a rigorous definition. Essentially, it is associated both with *genetic diversity* and *speciation* - the two different concepts from organic evolution that were discussed in Section 2.2 - at the same time. This is simply due to the fact that the differences between the two concepts do not have any practical effect on the evolutionary search and the goal of maintaining diversity among the evolving candidate solutions. In the well known trade-off between *exploration* and *exploitation* of the landscape during a search, *maintaining population diversity* is a driving force in the *exploration front*, and thus it is an important component. Among EC

[5] We deliberately avoid the definition of a funnel, as its definition is rather vague. We refer the reader to [51].

researchers, population diversity is first considered as a component due to play a role in a fruitful exploration of the landscape for the sake of obtaining a single solution, while its role in obtaining multiple solutions is typically considered as a secondary one.

Mahfoud's Formalism Mahfoud constructed a formalism for characterizing *population diversity* in the framework of Evolutionary Algorithms (see [37], pp. 50-59). Mahfoud's formal framework was based on the partitioning of the search space into equivalence classes (set to *minima* in the search landscape), a descriptive relation (typically, *genotypic* or *phenotypic* mappings), and the measurement of distance between the current distribution of subpopulations to some given *goal-distribution*.

Let $P = \{p_i\}_{i=1}^{\ell}$ be a discrete distribution describing the current partitioning of the population into subpopulations, i.e., p_i is the portion of the population located at the i^{th} site. Let $Q = \{q_i\}_{i=1}^{\ell}$ be the goal-distribution of the population with respect to the defined sites. We demand that by construction we have $\sum_{i=1}^{\ell} p_i = 1$, as well as $\sum_{i=1}^{\ell} q_i = 1$. The formalism focuses in defining the *directed divergence*, or distance, of distribution P to distribution Q. Several well-known metrics follow this formalism by satisfying its various criteria. We review some of them here.

1. The *entropy* of a system is a quantitative measurement of its *disorder* or *randomness* [52]. Although it had originated in Physics, in the *Second Law of Thermodynamics*, it also became an important criterion in *information systems*, also referred to as *Shannon's Information Entropy*. Accordingly, this general concept has several definitions, where we choose here to introduce a relevant definition to probability distributions.

   Definition 2.4.1. The *entropy* of a discrete probability distribution, $\{p_i\}_{i=1}^{\ell}$, is defined as:

   $$S(P) = \sum_{i=1}^{\ell} p_i \cdot \ln\left(\frac{1}{p_i}\right) = -\sum_{i=1}^{\ell} p_i \cdot \ln\left(p_i\right) \tag{2.4}$$

   The following measure, developed by Kullback and Leibler [53], quantifies the *directed divergence* between the two distributions, P and Q, as long as it is well defined (i.e., $\forall i \;\; p_i > 0, \; q_i > 0$):

   $$D\left(P, Q\right) = \sum_{i=1}^{\ell} p_i \cdot \ln\left(\frac{p_i}{q_i}\right) \tag{2.5}$$

   Given a uniform goal-distribution, the Kullback-Leibler measure is re-

duced to the following:

$$D(P, \mathbf{U}) = \sum_{i=1}^{\ell} p_i \cdot \ln\left(\frac{p_i}{1/\ell}\right) = \ln(\ell) - S(P) \tag{2.6}$$

Mahfoud shows that the Kullback-Leibler measure satisfies the criteria of his formalism, and can be used as a diversity measure.

2. The standard distance metrics are useful measures of directed divergence between the distributions.

Definition 2.4.2. A family of distance metrics is defined as follows:

$$D(P, Q) = \sqrt{\sum_{i=1}^{\ell} |p_i - q_i|^k}, \qquad 0 < k \leq \infty \tag{2.7}$$

Mahfoud shows that the family of distance metrics, with $0 < k \leq \infty$, satisfies the criteria and can be used as diversity measures.

This analytical framework, with its derived measurements of diversity, allowed Mahfoud to compare the role of population diversity among different GA niching techniques, and essentially became a performance criterion in his study.

Diversity Loss Subject to the complex dynamics of the various forces within an evolutionary algorithm, population diversity is typically lost, and the search is likely to converge into a single basin of attraction in the landscape.

Population diversity loss within the population of solutions is the fundamental effect which niching methods aim to treat. In fact, from the historical perspective, the quest for diversity-promoting-techniques was the main goal within the EC community for some time, and niching methods were merely obtained as *by-products*, so to speak, of that effort. As will be argued here, population diversity is an important component in a population-based search, and it even becomes critical in extended techniques, such as *Evolutionary Multi-Objective* approaches (see Chapter 5).

Next, we describe the effect of *diversity loss* within Evolution Strategies. This will be followed by some conclusions drawn by the GA research concerning diversity loss within GAs, as a point of reference to ES.

2.4.1 Diversity Loss in Evolution Strategies

The defining mechanism of ES is strongly dictated by the mutation operator as well as by the deterministic selection operator. As defining operators,

they have a direct influence on the diversity property of the population. The recombination operator, nevertheless, does not play a critical role in the ES mechanism. In practice, especially in the context of derandomized ES, it is not an essential component.

We attribute two main components to the *population diversity loss* within ES: fast *take-over*, which is associated with the *selection* operator, and *genetic drift* (or *neutrality* effect), which is associated both with the *selection* and the *recombination* operators, respectively.

Selective Pressure: Fast Take-Over

Evolution Strategies have a strictly deterministic, rank-based approach, to selection. In the two traditional approaches, (μ, λ) and $(\mu + \lambda)$, the best individuals are selected - implying, rather intuitively, high *selective pressure*. Due to the crucial role of the selection operator within the evolution process, its impact within the ES field has been widely investigated.

Goldberg and Deb introduced the important concept of *takeover time* [54], which gives a quantitative description of selective pressure **with respect to the selection operator exclusively**:

Definition 2.4.3. The *takeover time* τ^* is the minimal number of generations until repeated application of the selection operator yields a uniform population filled with copies of the best individual.

The selective pressure has been further investigated by Bäck [36], who analyzed all the ES selection mechanisms also with respect to takeover times. Here, we introduce the results for the takeover times of the main selection mechanisms in the absence of mutation, where we chose to omit the derivations. See [1] for the proofs.

Theorem 2.4.4. *The takeover time of* (μ, λ)*-selection is :*

$$\tau^*_{(\mu,\lambda)} = \frac{\ln(\lambda)}{\ln\left(\frac{\lambda}{\mu}\right)} \tag{2.8}$$

Theorem 2.4.5. *The takeover time of* $(\mu + \lambda)$*-selection is given implicitly by:*

$$\begin{aligned} \lambda &= \frac{\left(\alpha_1^{\tau^*+1} - \alpha_2^{\tau^*+1}\right)}{\sqrt{\frac{\lambda}{\mu} \cdot \left(\frac{\lambda}{\mu} + 4\right)}} \\ \alpha_{1,2} &= \frac{\lambda}{2\mu} \pm \frac{1}{2} \cdot \sqrt{\left(\frac{\lambda}{\mu}\left(\frac{\lambda}{\mu} + 4\right)\right)} \end{aligned} \tag{2.9}$$

Corollary 2.4.6. *It is easy to verify that upon the substitution of the traditional population sizes of the standard-ES, one obtains very short* takeover times *for the given selection mechanisms, which imply high* selective pressure.

The ratio $\frac{\lambda}{\mu}$ clearly plays a dominant role in the derived takeover times of the two selection approaches. Not surprisingly, the term *selective pressure* is occasionally associated with this ratio. It should be noted that the same ratio also governs the convergence velocity of the $(\mu \stackrel{+}{,} \lambda)$-ES for large population sizes, i.e., $\mu \gg 1$ (see [1] pp. 89-90).

ES Genetic Drift

We consider two different ES neutral effects, that could be together ascribed as a general ES genetic drift: *Recombination drift* and *selection drift*. We argue that these two components are directly responsible to the loss of population diversity in ES.

Recombination Drift Beyer explored extensively the so-called *mutation-induced speciation by recombination* (MISR) principle (see, e.g., [55]). According to this important principle, repeated application of the mutation operator, subject to a dominant recombination operator, would lead to a stable distribution of the population, which resembles a species or a cloud of individuals. When fitness-based selection is applied, this cloud is likely to move together towards fitter regions of the landscape. Furthermore, Beyer managed to prove analytically [55] that the MISR principle is indeed universal when finite populations are employed, subject to sampling-based recombination. The latter was achieved by analyzing the ES dynamics without fitness-based selection, deriving the expected population variance, and showing that it is reduced with random sampling in finite populations. This result was also corroborated by numerical simulations. That study provides us with an analytical result that a sampling-based recombination is subject to genetic drift, and leads to loss of population diversity.

Selection Drift At the same time, a recent study on the extinction of subpopulations on a simple *bimodal equi-fitness* model investigated the drift effect of the selection operator [56]. It considered the application of *selection* on finite populations, when the fitness values of the different attractors were equal (i.e., eliminating the possibility of a *take-over effect*), and argued that a neutral effect (*drift*) would occur, pushing the population into a single attractor. The latter study indeed demonstrated this effect of *selection drift* in ES, which resulted in a convergence to an equilibrium distribution around a single attractor. It was also shown that the time of extinction increases proportionally with μ. The analysis was conducted by means of Markov chain models, supported by statistical simulations.

Corollary 2.4.7. *Evolution Strategies that employ finite populations are typically underposed to several effects that are responsible for the loss of population diversity. It has been shown that the standard selection mechanisms may lead to a fast take-over effect. In addition, we argued that both the* recombination *and the* selection *operators experience their own* drift *effects that lead to population diversity loss. We conclude that an Evolution Strategy with a* small population *is likely to encounter a rapid effect of diversity loss.*

2.4.2 Point of Reference: Diversity Loss within GAs

Mahfoud devoted a large part of his thesis to studying population diversity within GAs [37]. He concluded that three main components can be attributed to the effect of population diversity loss within GAs:

- **Selection Pressure** The traditional GA applies a probabilistic selection mechanism, namely the *Roulette-Wheel Selection* (RWS). This mechanism belongs to a broad set of selection mechanisms which follow the fitness-proportionate selection principle. Selection pressure is thus associated with the 1^{st} *moment of the selection operator*. It has been demonstrated by Mahfoud [37] that the selection pressure, or equivalently the non-zero expectation of the selection operator, prevents the algorithm from converging in parallel into more than a single attractor.

- **Selection Noise** Selection noise is associated with the 2^{nd} *moment of the selection operator*, or its *variance*. Mahfoud [37] demonstrated that the high variance of the RWS, as well as of other selection mechanisms, is responsible for the fast convergence of a population into a single attractor, even when there exists a set of equally fit attractors. We consider this effect as a *genetic drift* in its broad definition - sampling error of a distribution - although it was not explicitly referred to as such by Mahfoud.

- **Operator Disruption** Evolutionary operators in general, and the *mutation* and *recombination* operators in particular, boost the evolution process toward exploration of the search space. In that sense, they have a constructive effect on the process, since they allow locating new and better solutions. However, their action also has a destructive effect. This is due to the fact that by applying them good solutions that have been located previously might be lost. In that sense, they eliminate competition between highly fit individuals, and "assist" some of them to take-over. The mutation operator usually has a small effect, since it acts in small steps - low mutation probability in the traditional GA, which means infrequent occurrence of bit flips. Thus, the mutation operator can be considered to have a negligible disruption. The recombination operator, on the other hand, has a more considerable

effect. In the GA field, where the *crossover* operator is in use (single-point, two-point or n-point crossovers), it has been shown to have a disruptive nature by breaking desired patterns within the population (the well known *Schema Theorem* discusses the schema disruption by the crossover operator and states that schemata with high defining length will most likely be disrupted by the crossover operator; see, e.g., [22]).

It should be noted that an equivalent ES disruptive-recombination effect was analyzed in [57], and was shown to boost the extinction of subpopulations located around a basin of attraction. Furthermore, it was observed that by omitting the recombination operator the stability of the subpopulations was indeed strengthened.

2.4.3 Neutrality in ES Variations: Mutation Drift

The mutation operator, the defining operator of Evolution Strategies, applies normally-distributed variations of finite sample sizes, and thus is expected to experience sampling errors as the sample sizes decrease. These sampling errors lead to an undirected movement of the population center of mass, with speed which depends on the population size. We shall call this effect *mutation drift*.

Simulations In order to demonstrate and analyze this *mutation drift* effect, we conducted simulations on the following basic ES model: The parallel evolution of several populations in an n-dimensional space, based on sequential normally-distributed variations (with a fixed identity matrix as the covariance of the distribution), without selection nor recombination. The ES variation can be then considered as a *continuous random walk* of μ individuals in an n-dimensional space. Essentially, this corresponds to **mutation-only ES** of multiple populations.

We simulated 10 simultaneously evolving populations, for three test-cases of population sizes: $\mu_1 = 10$, $\mu_2 = 100$, and $\mu_3 = 1000$, subject to three space dimensions: $n_1 = 1$, $n_2 = 10$, and $n_3 = 1000$. For each simulation, we measured the distance of the *population mean*, or *center of mass*, to the starting point, as a function of generational steps. More precisely, we measured the location of the *population mean* for n_1, and the *Euclidean distance from the origin* for $\{n_2, n_3\}$. Figure 2.2 presents the outcome of these calculations. It is easy to observe in those simulations a similar trend to the equivalent simulations of Section 2.2.1: The center of mass strongly drifts away from the origin when the population is small, and shows the contrary behavior when the population is large. We therefore conclude that mutation drift is very likely to occur in small population sizes, and is not likely to occur in large population sizes.

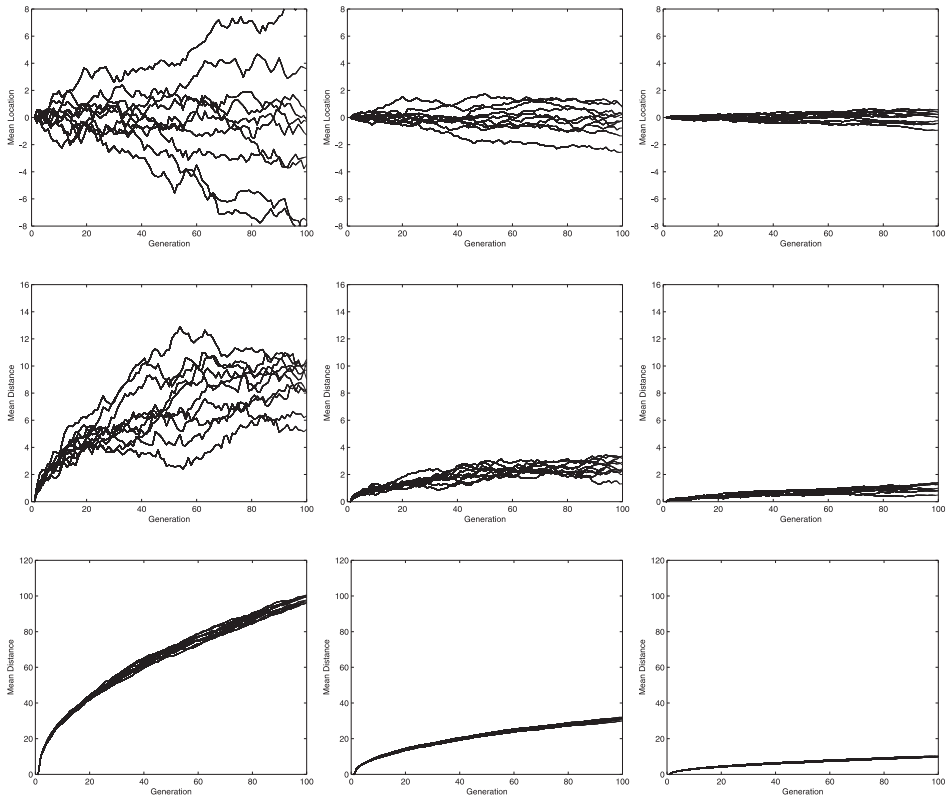

Figure 2.2: Illustration of the *mutation drift* effect in ES, for 10 simultaneously evolving populations, as a function of population size [$\mu_1 = 10$ (left), $\mu_2 = 100$ (center), and $\mu_3 = 1000$ (right)] and landscape dimensionality [$n_1 = 1$ (top), $n_2 = 10$ (center), and $n_3 = 1000$ (bottom)]. The vertical axes correspond to the **location of the center of mass of the population** (for $n_1 = 1$, top row) or **distance from the origin to the center of mass of the population** (for $n_2 = 10$ or $n_3 = 1000$, in the center or bottom rows, respectively). The horizontal axis corresponds to the generational step of the calculation.

We thus demonstrated here that the *center of mass* of a small ES population is subject to a so-called *mutation drift*. This is an equivalent effect to the genetic drift of alleles, as described in Section 2.2.1. We claim that it allows for easy translation of small populations from one location to another, having the potential to boost fast and efficient *speciation*. Therefore, we argue that drift in this context can be a blessing for the fast formation of species in niching.

Since small populations are typically employed by Evolution Strategies, and especially by the derandomized variants, we consider this effect of *mutation drift* as a positive potential component for niching with ES. This result

provides us with further motivation to introduce DES with small populations into the niching framework.

2.5 Classical Niching Techniques

Despite the fact that the motivation for multimodal optimization is beyond doubt, and the biological inspiration is real, there is no unique definition of the mission statement for *niching techniques*. There have been several attempts to provide a proper definition and functional specification for niching; we review some of them here:

1. Mahfoud [37] chose to put emphasis on locating as well as maintaining good optima, and formulated the following:

   > The litmus test for a niching method, therefore, will be whether it possesses the capability to find multiple, final solutions within a reasonable amount of time, and to maintain them for an extended period of time.

2. Beyer et al. [58] put forward also the actual maintenance of population diversity:

   > *Niching*: process of separation of individuals according to their states in the search space or maintenance of diversity by appropriate techniques, e.g. local population models, fitness sharing, or distributed EA.

3. Preuss [59] considered the two definitions mentioned above, and proposed a third:

   > Niching in EAs is a two-step procedure that **(a)** concurrently or subsequently distributes individuals onto distinct basins of attraction and **(b)** facilitates approximation of the corresponding (local) optimizers.

GA Niching Methods Niching methods within Genetic Algorithms have been studied during the past few decades, initially triggered by the necessity to promote *population diversity* within EAs. The research has yielded a variety of different methods, which are the vast majority of existing work on niching in general. The remainder of this section will focus on GA niching techniques, by providing a short overview of the main known methods, with emphasis on the important concepts of *Sharing* and *Crowding*. This survey is mainly based on [37] and [60].

2.5.1 Fitness Sharing

The *sharing* concept was one of the pioneering niching approaches. It was first introduced by Holland in 1975 [4], and later implemented as a niching technique by Goldberg and Richardson [61]. This strong approach of **considering the fitness as a shared resource** has essentially become an important concept in the broad field of Evolutionary Algorithms, and laid the foundations for various successful niching techniques for multimodal function optimization, mainly within GAs. A short description of the *fitness sharing* mechanism follows.

The basic idea of *fitness sharing* is to consider the fitness of the landscape as a resource to be shared among the individuals, in order to decrease redundancy in the population. Given the similarity metric of the population, which can be *genotypic* or *phenotypic*, the *sharing function* is defined as follows:

$$sh(d_{i,j}) = \begin{cases} 1 - \left(\frac{d_{i,j}}{\rho}\right)^{\alpha_{sh}} & \text{if } d_{i,j} < \rho \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

where $d_{i,j}$ is the distance between individuals i and j, ρ (traditionally noted as σ_{sh}) is the fixed radius of every niche, and $\alpha_{sh} \geq 1$ is a control parameter, typically set to 1. Using the *sharing function*, the *niche count* is given by

$$m_i = \sum_{j=1}^{N} sh(d_{i,j}) \tag{2.11}$$

Let an individual raw fitness be denoted by f_i, then the *shared fitness* is defined by:

$$f_i^{sh} = \frac{f_i}{m_i} \tag{2.12}$$

assuming that the fitness is *strictly positive* and subject to *maximization*. The evaluation of the shared fitness is followed by the selection phase, which is typically based on the *roulette wheel selection* (RWS) operator [22]; The latter takes into consideration the shared fitness. Thus, the *sharing* mechanism practically punishes individuals that have similar members within the population via their fitness, and by that it aims at reducing redundancy in the gene pool, especially around the peaks of the fitness landscape.

One important auxiliary component of this approach is the *niche radius*, ρ. Essentially, this approach makes a strong assumption concerning the fitness landscape, stating that the optima are far enough from one another with respect to the *niche radius*, which is estimated for the given problem and remains fixed during the course of evolution. This poses the so-called *niche radius problem*, to be discussed later, especially in Chapters 3 and 4.

It is important to note that the formulas for determining the value of ρ, which will be given in Chapter 3, are dependent on q, the number of peaks of the target function. Hence, a second assumption is that q can be estimated.

In practice, an accurate estimation of the expected number of peaks q in a given domain may turn out to be extremely difficult. Moreover, peaks may vary in shape, and this would make the task of determining ρ rather complicated. This provides us with the motivation to treat the issue of niche shapes in Chapter 4.

In the literature, several GA niching *sharing*-based techniques, which implement and extend the basic concept of sharing, can be found [37, 61, 62, 63, 64, 65, 66]. Furthermore, the concept of sharing was successfully extended to other "yields of interest", such as *concept sharing* [38].

2.5.2 Dynamic Fitness Sharing

In order to improve the *sharing* mechanism, a dynamic approach was proposed. The *dynamic niche sharing* method [64], which extended the *fitness sharing* technique, aimed at dynamically recognizing the q peaks of the forming niches, and based on that information classified the individuals as either members of one of the niches, or as members of the "non-peaks domain".

Explicitly, let us introduce the *dynamic niche count*:

$$m_i^{dyn} = \begin{cases} n_j & \text{if individual } i \text{ is within dynamic niche } j \\ m_i & \text{otherwise (non-peak individual)} \end{cases} \tag{2.13}$$

where n_j is the size of the j^{th} dynamic niche (i.e., the number of individuals which were classified to niche j), and m_i is the standard *niche count*, as defined in Eq. 2.11.

The shared fitness is then defined as follows:

$$f_i^{dyn} = \frac{f_i}{m_i^{dyn}} \tag{2.14}$$

The identification of the dynamic niches can be carried out by means of a *greedy* approach, as proposed in [64] as the Dynamic Peak Identification (DPI) algorithm (see Algorithm 4). As in the original *fitness sharing* technique, the *shared fitness evaluation* is followed by the selection phase, typically implemented with the RWS operator. Thus, this technique does not fixate the peak individuals, but rather provides them with an advantage in the selection phase, which is probability-based within GAs.

2.5.3 Clearing

Another variation to the *fitness sharing* technique, called *clearing*, was introduced by Petrowski [65] at the same time as the *dynamic fitness sharing* [64]. The essence of this mechanism is the *'winner takes it all'* principle, and its idea is to designate a specific number of individuals per niche, referred to as *winners*, which could enjoy the resources of that niche. This is equivalent to the introduction of a "death penalty" to the *losers* of the niche, the

Algorithm 4 Dynamic Peak Identification (DPI)

input: population Pop, number of niches q, niche radius ρ

```
1: Sort Pop in decreasing fitness order
2: i := 1
3: NumPeaks := 0
4: DPS := ∅ {Set of peak elements in population}
5: while NumPeaks ≠ q and i ≤ popSize do
6:   if Pop[i] is not within sphere of radius ρ around peak in DPS then
7:     DPS := DPS ∪ {Pop[i]}
8:     NumPeaks := NumPeaks + 1
9:   end if
10:  i := i + 1
11: end while
```

output: DPS

individuals of each niche which lose the generational competition to the actual peak-individuals. Following a *radius-based* procedure of identifying the winners and losers of each niche in each generation, the winners are assigned with their raw-fitness values, whereas all the other individuals are assigned with *zero* fitness. This is called the *clearing phase*. The selection phase, typically based on the RWS operator, considers *de facto* only the winners of the different niches. The allowed number of winners per niche, also referred to as the *niche capacity*, is a control parameter that reflects the degree of elitism. In any case, as in previous techniques, the peaks are never fixated, and are subject to the probabilistic selection of the GA.

This methods was shown to outperform the *fitness sharing* technique on a specific set of low-dimensional test problems [65].

2.5.4 Crowding

Crowding was one of the pioneering methods in this field, as introduced by de Jong in 1975 [67]. The *crowding* approach aimed at reducing changes in the population distribution between generations, in order to prevent *premature convergence*; it does so by applying *restricted replacement*. Next, we will describe the method in more detail.

Given the traditional GA, a proportion G of the population is selected in each generation via fitness-proportionate selection to undergo variations (i.e., *crossover* and *mutation*) - out of which a part is chosen to die and to be replaced by the new offspring. Each offspring finds the individuals it replaces by taking a random sample of CF (referred to as **crowding factor**) individuals from the population, and replacing the **most similar individual** from the sample. An appropriate *similarity metric* should be chosen.

The crucial point of this niching mechanism is the calculation of the

Algorithm 5 Deterministic Crowding: Replacement Selection

1: Select two parents, p_1 and p_2, randomly, without replacement
2: Generate two variations, c_1 and c_2
3: **if** $d(p_1, c_1) + d(p_2, c_2) \leq d(p_1, c_2) + d(p_2, c_1)$ **then**
4: **if** $f(c_1) > f(p_1)$ **then** replace p_1 with c_1
5: **if** $f(c_2) > f(p_2)$ **then** replace p_2 with c_2
6: **else**
7: **if** $f(c_2) > f(p_1)$ **then** replace p_1 with c_2
8: **if** $f(c_1) > f(p_2)$ **then** replace p_2 with c_1
9: **end if**

so-called *crowding distance* **between parents and offspring**, in order to control the *change rate* between generations. A different use of the *crowding distance*, applied among individuals of the same generation and assigned with reversed ranking, will be revisited in the context of Evolutionary Multi-Objective Optimization in Chapter 5; In the context of niching see also Deb's "Omni-Optimizer" ([68] and Section 5.2.1).

Mahfoud, who analyzed the *crowding* niching technique [37], concluded that it was subject to disruptive effects, mainly *drift*, which prevented it from maintaining more than two peaks. He then proposed a mechanism called *deterministic crowding*, as an improvement to the original *crowding* niching technique. The proposed procedure applies variation operators to pairs of individuals in order to generate their offspring, who are then all evaluated with respect to the crowding distance, and undergo *replacement selection* (see Algorithm 5, which assumes *maximization*).

2.5.5 Clustering

The application of *clustering* for niching is very intuitive from the computational perspective, as well as straightforward in its implementation. Yin et al. [62] proposed a clustering framework for niching with GAs, which we describe here briefly. A clustering algorithm, such as the *K-Means* algorithm [69], first partitions the population into niches, and then considers the *centroids*, or center points of mass, of the newly partitioned subpopulations.

Let d_{ic} denote the distance between individual i and its *centroid*, and let f_i denote the raw fitness of individual i. Assuming that there are n_c individuals in the niche of individuals i, its fitness is then defined as:

$$f_i^{Clustering} = \frac{f_i}{n_c \cdot (1 - (d_{ic}/2d_{max})^{\alpha})}, \tag{2.15}$$

where d_{max} is the maximum distance allowed between an individual and its niche centroid, and α is a defining parameter. It should be noted that the clustering algorithm uses an additional parameter, d_{min}, for determining the

minimal distance allowed between centroids, playing an equivalent role to the *niche radius* ρ of the *sharing*-based mechanisms.

This method is often subject to criticism for its strong dependency on a relatively large number of parameters. However, this *clustering* technique has become a popular kernel for niching with EAs, and its application was reported in various studies (see, e.g., [56, 70, 71, 72, 73, 74, 75]).

2.5.6 The Sequential Niche Technique

The straightforward approach of *iteration* can be used to locate sequentially multiple peaks in the landscape, by means of an *iterative local search* [76]. This procedure is blind to any information gathered in previous searches, and sequentially restarts stochastic searches, hoping to hit a different peak every run. Obviously, it is likely to encounter *redundancy*, and the number of expected iterations is then increased by a factor. A **redundancy factor** can be estimated if the peaks are of equal height (equi-fitness landscape), i.e., the probability to converge into any of the q peaks is equal to $1/q$:

$$R = \sum_{i=1}^{q} \frac{1}{i}$$

For $q > 3$, this can be approximated by:

$$R \approx \gamma + \ln(q), \tag{2.16}$$

where $\gamma \approx 0.577$ is the Euler-Mascheroni constant. This *redundancy factor* remains reasonably low for any practical value of q, but is expected to considerably increase if all optima are not equally likely to be found.

On a related note, we would like to mention a multi-restart with increasing population size approach that was developed with the CMA algorithm [77]. The latter aims at attaining the global minimum, while possibly visiting local minima along the process and restarting the algorithm with a larger population size and a modified initial step-size. It is not defined as a niching technique and does not target optima other than the global minimum, but it can capture sub-optimal minima during its search.

Beasley et al. extended the naive *iteration* approach, and developed the so-called *Sequential Niche* technique [78]. This method, in contrast to the other niching methods presented earlier, does not modify the genetic operators nor any characteristics of the traditional GA, but rather creates a general search framework suitable for locating multiple solutions. By means of this method the search process turns into a sequence of independent runs of the traditional GA, where the basic idea is to suppress the fitness function at the observed optimum that was obtained in each run, in order to prevent the search from revisiting that optimum.

In further detail, the traditional GA is run multiple times sequentially: given the best solution of each run, it is first stored as a possible final solution, and secondly the fitness function is artificially suppressed in all the points within the neighborhood of that optimum up to a desired radius. This modification is done immediately after each run. Its purpose is to discourage the following runs from revisiting these optima, and by that to encourage the exploration of other areas of the search landscape - aiming at obtaining all its optima. It should be noted that each function modification might yield artificial discontinuities in the fitness landscape. This method focuses only on locating multiple optima of the given search problem, without considering the concepts of parallel evolution and subpopulations formation. In that sense, it has been claimed that it could not be considered as a niching method, but rather as a modified iterated search.

2.5.7 The Islands Model

This is probably the most intuitive niching approach from the biological perspective, directly inspired by organic evolution. Also referred to as the *Regional Population Model*, this approach (see, e.g., [79, 80, 81]) simulates the evolution of subpopulations on remote computational units (independent processors), aiming at achieving a speciation effect by **monitoring the gene flow**. The population is divided into multiple subpopulations, which evolve independently for a fixed number of generations, called *isolation period*. This is followed by a phase of controlled gene flow, or *migration*, when a portion of each subpopulation migrates to other nodes.

The genetic diversity and the amount of information exchange between subpopulations are determined by the following parameters - the number of exchanged individuals, the *migration rate*, the selection method of the individuals for migration (uniformly at random, or elitist fitness-based approach), and the scheme of migration, e.g., complete net topology, ring topology, or neighborhood topology.

2.5.8 Other GA-Based Methods

Tagging (see, e.g., [82, 83]) is a mechanism that aims at improving the distance-based methods of *fitness sharing* and *crowding*, by labeling individuals with tag-bits. Rather than carrying out distance calculations, the tag-bits are employed for identifying the subpopulations, enforcing *mating restrictions*, and then implementing the *fitness sharing* mechanism. An individual is classified to a subpopulation by its genetic inheritance, so to speak, which is subject to generational variations, rather than by its actual spatial state. This concept simplifies the classification process, and obviously reduces the computational costs per generation, but it also introduces a new bio-inspired approach into niching: individuals belong to a species because

their parents did, and not because they are currently adjacent to a "peak individual", for instance. This technique was shown in [82] to be a rather efficient implementation of the *sharing* technique.

A complex subpopulation differentiation model, the so-called **Multinational Evolutionary Algorithm**, was presented in [84]. This original technique considers a world of *"nations"*, *"governments"*, and *"politicians"*, with dynamics dictated by migration of individuals, merging of subpopulations, and selection. Additionally, it introduces a topology-based auxiliary mechanism of *sampling*, which detects whether feasible solutions share the same basin of attraction. Due to the *curse of dimensionality*, this sampling-based mechanism is expected to lose its efficiency in high-dimensional landscapes.

Stoean et al. [85] constructed the so-called **Elitist Generational Genetic Chromodynamics Algorithm**. The idea behind this radius-based technique was the definition of a *mating region*, a *replacement region*, and a *merging region* — with appropriate mating-, replacement-, and merging-radii — which dictates the dynamic of the genetic operations.

Chapter 4 will elaborate furthermore on specific GA-based niching techniques in the context of the so-called *niche radius problem*.

2.5.9 Miscellaneous: Mating Schemes

It has been observed that once the niche formation process starts, i.e., when the population converges into the multiple basins in the landscape, crossbreeding between different niches is likely to fail in producing good offspring. In biological terms, this is the elimination of the divergence, by means of *hybridization*, in the **secondary contact phase**, as discussed in Section 2.2.2.

Deb and Goldberg [54] proposed a so-called *mating restriction scheme*, which poses a limitation on the choice of partners in the reproduction phase and prevents recombination between competing niches. They used a distance measure, subject to a distance threshold which was set to the niche radius, and showed that it could be used to improve the *fitness sharing* algorithm.

Mahfoud [37] proved that the mating restriction scheme of Deb and Goldberg was not sufficient *per se* in maintaining the population diversity in GA niching. A different approach of Smith and Bonacina [86], however, considered an Evolutionary Computation Multi-Agent System, as opposed to the traditional *centralized* EA, and did manage to show that the same mating restriction scheme in an agent-based framework was capable in maintaining diversity and converging with stability to the desired peaks.

From the biological perspective, the mating restriction scheme is obviously equivalent to keeping the geographical isolation, or the barrier to gene flow, in order to allow the completion of the speciation phase. As discussed earlier, the geographical element in organic evolution is the crucial component which creates the conditions for speciation, and it is not surprising that

artificial niching techniques choose to enforce it, by means of mechanisms such as the niche radius or the mating restriction scheme.

2.6 Niching in Evolution Strategies

Researchers in the field of Evolution Strategies initially showed no particular interest in the field of niching, leaving it essentially for Genetic Algorithms. An exception would be the employment of island models. Roughly speaking, classical niching mechanisms such as *fitness sharing*, which redefine the selection mechanism, are likely to interfere with the core of Evolution Strategies – the *self-adaptation mechanism* – and thus doomed to experience problems in a straightforward implementation. Manipulations of fitness values are usually not suitable for Evolution Strategies, as in the case of constraints handling: death-penalty is typically the preferred approach for constraints violation in ES, rather than a continuous penalty as used in other EAs, in order to avoid the introduction of disruptive effects to the self-adaptation mechanism (see, e.g., [34, 87]). Therefore, niching with Evolution Strategies would have to be addressed from a different direction. Moreover, the different nature of the ES dynamics, throughout the *deterministic selection* and the *mutation operator*, suggests as well that a different treatment is required here.

There are several, relatively new, niching methods that have been proposed within ES, mostly clustering-based [56, 73, 74]. A different approach, which preceded this thesis, was presented in [88, 89, 90].

2.7 Discussion and Mission Statement

Niching techniques, following somehow various *mission statements*, introduce a large variety of approaches, some of which are more biologically inspired, whereas others are multimodal-optimization oriented. In both cases, those techniques were usually tested on **low-dimensional artificial landscapes**, and the application of these methods to real-world landscapes was hardly ever reported. We claim that niching methods should be implemented also for attaining multiple solutions in high-dimensional real-world problems, serving the decision makers by providing them with the choice of optimal solutions, and representing well Evolutionary Algorithms in multimodal domains. By our humble reckoning, the *multimodal front* of real-world applications, i.e. multimodal real-world problems which demand multiple optimal solutions, should also enjoy the powerful capabilities of Evolution Strategies, as other fronts do, e.g., multi-objective domains and constrained domains.

On a different note, Preuss, in an important paper [59], raised the question: *"Under what conditions can niching techniques be faster than iterated local search algorithms?"*. Considering a simplified model, and assuming the

existence of an efficient basin identification method, he managed to show that it pays off to employ Evolutionary Algorithms niching techniques on landscapes whose basins of attraction vary significantly in size. However, the original question in its general form remained open.

Mahfoud [91] drew a comparison of *parallel* versus *sequential* niching methods, while considering *fitness sharing*, *deterministic crowding*, *sequential niching*, and *parallel hillclimbing*. Generally speaking, he concluded that parallel niching GAs outperform parallel hillclimbers on a hard set of problems, and that *sequential niching* is always outperformed by the parallel approaches.

Obviously, there is *no free lunch*, and there is no best technique, especially in niching. In this context, *local search* capabilities should not be underestimated, and *population diversity preservers* should not be overestimated. We claim that like any other complex component in organic as well as artificial systems - the success of niching is about the subtle interplay between the different, sometime conflicting, driving effects.

We thus choose to adopt Preuss' mission statement, and *define the challenge in niching as follows*:

> **Attaining the optimal interplay between partitioning the search space into niches occupied by stable subpopulations, by means of population diversity preservation, and exploiting the search in each niche by means of a highly efficient optimizer with local-search capabilities**.

All animals are equal,
but some animals are more equal than others.
Animal Farm; George Orwell

Chapter 3

Niching with Derandomized Evolution Strategies

3.1 General

Following our *mission statement*, as presented in Section 2.7, we would like to construct a generic niching framework which offers the combination of population diversity preservation and local-search capabilities. We consider Derandomized Evolution Strategies as the best choice for that purpose, as EA variants with local search characteristics (see our discussion in Section 1.4.6). Furthermore, DES typically employ small populations, which was shown to be a potential advantage for a niching technique, as it can boost the speciation effect (Section 2.4.3). Thus, we are now challenged to complete the framework by introducing a mechanism for partitioning the search space into "ecological optima", and stimulating population diversity preservation.

We restrict this chapter to the scope of niching with a fixed niche radius, assuming that the landscapes under investigation would not dramatically suffer from the so-called niche-radius assumptions. Chapter 4 will extend this framework to self-adaptive approaches, which will aim at treating these assumptions.

This chapter presents our proposed algorithm, introduces our test bed of artificial landscapes as well as the performance criteria, and finally discusses the numerical results of our calculations.

3.2 The Proposed Algorithm

The advent of derandomized Evolution Strategies allows successful global optimization with minimal requirements concerning exogenous parameters, mostly without recombination, and with a low number of function evaluations. In particular, consider the $(1 \stackrel{+}{,} \lambda)$ derandomized ES variants presented in Chapter 1. In the context of niching, this generation of modern ES vari-

ants allows the construction of fairly simple and elegant niching algorithms. Next, we outline our proposed method.

Our niching technique is based upon interacting search processes, which simultaneously perform a derandomized $(1, \lambda)$ or $(1 + \lambda)$ search in different locations of the landscape. In case of multimodal landscapes these search processes are meant to explore different attractor basins of local optima.

An important point in our approach is to strictly enforce the fixed allocation of the population resources, i.e. number of offspring, per niche. The idea is thus to prevent a *take-over scenario*, in which a subpopulation located at a fitter optimum generates more offspring in comparison to competing subpopulations. The biological idea behind this fixed allocation of resources lies in the concept of limited *hosting capacities* of given ecological niches, as introduced in Chapter 2.

The *speciation interaction* occurs every generation when all the offspring are considered together to become niches' representatives for the next iteration, or simply the next search points, based on the rank of their fitness and their location with respect to higher-ranked individuals. We focus in a simple framework without recombination ($\mu = 1$), whereas niching with recombination will be considered in the specific context of Chapter 5.

3.2.1 Niching with $(1 \dagger \lambda)$ DES Kernels

Given q, the estimated/expected number of peaks, $q + p$ "D-sets" are initialized, where a D-set is defined as the collection of all the dynamically adapted strategy as well as decision parameters of the derandomized algorithm, which uniquely define the search at a given point of time. These parameters are the current search point, the mutation vector / covariance matrix, the global step-size, as well as other auxiliary parameters. At every point in time the algorithm stores exactly $q + p$ D-sets, which are associated with $q + p$ search points: q for the peaks and p for the "non-peaks domain". The $(q+1)^{th}...(q+p)^{th}$ D-sets are individuals which are randomly re-generated every *epoch*, i.e. a cycle of κ generations, as potential candidates for niche formation. This is basically a *quasi-restart* mechanism, which allows new niches to form dynamically. We stress that the total number of function evaluations allocated for a run should depend on the number of desired peaks, q, and not on p. Setting the value of p essentially reflects the following dilemma: Applying a wide restart approach for further exploring the search space, versus exploiting computational resources for the existing niches. In any case, due to the *curse of dimensionality*, p loses its significance as the dimension of the problem increases.

Until the stopping criterion is met, the following procedure takes place. Each search point samples λ offspring, based on its evolving D-set. After the fitness evaluation of the new $\lambda \cdot (q+p)$ individuals, the classification into niches of the entire population is obtained in a *greedy* manner, by means of

Algorithm 6 $(1 \dagger \lambda)$-DES Niching with a Fixed Niche Radius

1: **for** $i = 1 \dots (q + p)$ search points **do**
2: Generate λ samples based on the D-set of i
3: **end for**
4: Evaluate fitness of the population
5: Compute the Dynamic Peak Set (DPS) with the DPI Algorithm
6: **for all** elements of DPS **do**
7: Set peak as a search point
8: Inherit the D-set and update it respectively
9: **end for**
10: **if** N_{DPS}=size of $DPS < q$ **then**
11: Generate $q - N_{DPS}$ new search points, reset D-sets
12: **end if**
13: **if** $gen \mod \kappa \equiv 0$ **then**
14: Resample the $(q+1)^{th} \dots (q+p)^{th}$ search points
15: **end if**

the DPI routine [64] (Algorithm 4). The latter is based on the fixed niche radius ρ. The peaks then become the new search points, while their D-sets are inherited from their parents and updated respectively.

We would like to point out the dynamic nature of the subpopulations dynamics. Due to the *greedy* classification to niches, which is carried out every generation, some niches can merge in principle, while all the individuals, except for the *peak individual*, die out in practice. Following our principle of fixed resources per niche, only the peak individual will be sampled λ times in the following generations. In socio-biological terms, the peak individual could be associated with an **alpha-male**, which wins the local competition and gets all the sexual resources of its ecological niche.

A pseudo-code for the *niching routine* is presented as Algorithm 6.

Sizing the Population We follow the recommended population size for $(1, \lambda)$ derandomized ES (see, e.g., [25]), and set $\lambda = 10$. On this note, we would like to mention a theoretical work on sizing the population in a derandomized $(1, \lambda)$ ES with respect to the local progress [92]. The latter work obtained theoretical results showing that the local serial progress is maximized when the expected progress of the second best individual vanishes. These results allowed for the construction of a population size adaptation scheme, which sets the value of λ as a function of the fitness difference of the second fittest offspring and its parent. This adaptation scheme was shown to perform well on a set of simple theoretical landscapes [92].

3.3 Niche Radius Calculation

The original formula for the niche radius ρ, for *phenotypic sharing* in GAs, was derived by Deb and Goldberg [54]. Analogously, by considering the ES decision space as the GA decoded parameter space, the same formula can be applied to optimization tasks defined over continuous domains, by employing the Euclidean metric. Given q, the number of peaks in the solution space, every niche is considered to be surrounded by an n-dimensional hypersphere with radius ρ, which occupies $\frac{1}{q}$ of the entire volume of the space. The volume of the hypersphere which contains the entire space is

$$V = cr^n, \tag{3.1}$$

where c is a constant, given explicitly by

$$c = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}, \tag{3.2}$$

with $\Gamma(n)$ as defined in Eq. 1.36. Given lower and upper bound values, $\{x_{k,min},\ x_{k,max}\}$, of each coordinate in the decision parameters space, r is defined as follows:

$$r = \frac{1}{2}\sqrt{\sum_{k=1}^{n}(x_{k,max} - x_{k,min})^2} \tag{3.3}$$

Upon dividing the volume into q parts, we may write

$$c\rho^n = \frac{1}{q}cr^n, \tag{3.4}$$

which yields

$$\rho = \frac{r}{\sqrt[n]{q}} \tag{3.5}$$

Hence, by applying this niche radius approach, two assumptions are made:

1. The expected/desired number of peaks, q, is given or can be estimated.
2. All peaks are at least in distance 2ρ from each other, where ρ is the fixed radius of every niche.

3.4 Experimental Procedure

In order to test our proposed algorithmic niching framework, we would like to apply them to a test suite of artificial landscapes. Their application to Quantum Control landscapes will be reported in Part II.

We describe here our experimental procedure. We begin by discussing the construction of our test suite, and then present the numerical observation of our calculations.

3.4.1 Multi-Modal Test Functions

The choice of a numerical testbed for evaluating the performance of search or optimization methods is certainly one of the core issues among the scholars in the community of algorithms and Operations Research.

In a benchmark article, Whitley et al. [11] criticized the commonly tested artificial landscapes in the Evolutionary Algorithms community, and offered general guidelines for constructing test problems. We state these guidelines here:

1. *Test suites should contain problems that are resistant to hill-climbers.* Hill-climbing strategies, including *line search*, are typically faster than EAs, when they are successful. Hence, it is justified to test EAs on landscapes which cannot be easily hill-climbed.

2. *Test suites should contain problems that are non-linear, non-separable, and non-symmetric.*

3. *Test suites should contain scalable functions.* The dimensionality of the search space is an important issue, and thus should be tested accordingly.

4. *Test suites should contain problems with scalable evaluation cost.* The cost of some evaluation functions grows as a function of the search space dimensionality. This typically characterizes real-world problems, and should be considered.

5. *Test problems should have a canonical form.* This demand is relevant to encoding-based algorithms, such as GAs.

The following remarkable effort was made almost a decade after that document, when a large group of scholars in the EC community joined their efforts and compiled an agreed test suite of artificial landscapes [93], to be tested in an open performance competition reported at IEEE CEC 2005 [35]. The latter also included multimodal functions.

The issue of developing a multimodal test suite received even less attention, likely due to historical reasons. Since multimodal domains were mainly treated by GA-based niching methods, their corresponding test suites were limited to low-dimensional continuous landscapes, typically with two decision parameters to be optimized ($n = 2$) (see, e.g., [61, 37]).

In essence, our study is the first to introduce EA niching methods into high-dimensional continuous landscapes.

When compiling our test suite, we aimed at following Whitley's guidelines, including some traditional GA-niching test functions as well as functions from [93]. The reader should keep in mind that our niching methods will be applied on real-world landscapes in Chapters 8 and 9.

Our test suite contains the following artificial multimodal continuous functions (see Table 3.1 for their mathematical description):

- $\mathcal{M}$ is a basic hyper-grid multimodal function with uniformly distributed minima of equal function value of -1. It is meant to test the stability of a particularly large number of niches: in the interval $[0,1]^n$ it has 5^n minima. We used $\alpha = 6$.

- The well known Ackley function has one global minimum, regardless of its dimension n, which is surrounded isotropically by $2n$ local minima in the first hypersphere, followed by an exponentially increasing number of minima in successive hyperspheres. Ackley's function has been widely investigated in the context of Evolutionary Algorithms (see, e.g., [1]). We used $c_1 = 20$, $c_2 = 0.2$, and $c_3 = 2\pi$.

- $\mathcal{L}$ - also known as $F2$, as originally introduced in [61] - is a sinusoid trapped in an exponential envelope. The parameter k determines the sharpness of the peaks in the function landscape; we set it to $k = 6$. $\mathcal{L}$ has one global minimum, regardless of n and k. It has been a popular test function for GA niching methods. We used $l_1 = 5.1$, $l_2 = 0.5$, $l_3 = 4 \cdot \ln(2)$, $l_4 = 0.0667$ and $l_5 = 0.64$.

- The Rastrigin function [8] has one global minimum, surrounded by a large number of local minima arranged in a lattice configuration.
  We also consider its shifted-rotated variant [93], with a linear transformation matrix of condition number 2 as the rotation operator (see below a note on implementation).

- The Griewank function [8] has its global minimum ($f^* = 0$) at the origin, with several thousand local minima in the area of interest. There are 4 sub-optimal minima: $\tilde{f} \approx 0.0074$ with $\tilde{\vec{x}} \approx \left(\pm\pi, \pm\pi\sqrt{2}, 0, 0, 0, ...0\right)$.
  We also consider its shifted-rotated variant [93], with a linear transformation matrix of condition number 3 as the rotation operator (see a note on implementation below).

- The function after Fletcher and Powell [1] is a non-separable *non-linear parameter estimation problem*, which has a non-uniform distribution of 2^n minima. It has non-isotropic attractor basins. See a note on implementation below.

A Note on Implementation Most of the data for the functions, and in particular the translation and rotation operators, was retrieved from [93][1]. The Fletcher-Powell data (the matrices $\mathbf{A}$, $\mathbf{B}$ and the vector $\vec{\alpha}$) was retrieved from [1].

[1] Data is available for download at `http://www.ntu.edu.sg/home/epnsugan/index_files/`.

Table 3.1 summarizes the unconstrained multimodal test functions as well as their initialization intervals.

3.4.2 Performance Criteria

The traditional GA niching methods research had been strongly interested in the distribution of the final population compared to a goal-distribution, as formalized by Mahfoud (see Section 2.4). While Mahfoud's formalism introduced a generic theoretical tool, being derived from information theory, most of the studies considered *de facto* specific performance calculations. For example, a very popular niching performance measurement, which satisfies Mahfoud formalism's criteria, is the *Chi-square-like performance statistic* (see, e.g., [54]). The latter estimates the deviation of the actual distribution of individuals N_i from an ideal distribution (characterized by mean μ_i and variance σ_i^2) in all the $i = 1 \dots q+1$ subspaces (q peak subspaces and the non-peak subspace):

$$\chi^2 = \sqrt{\sum_{i=1}^{q+1} \left(\frac{N_i - \mu_i}{\sigma_i} \right)^2}, \tag{3.6}$$

where the ideal-distribution characteristic values are derived per function.

Our research focuses on the ability to identify global as well as local optima, and to converge in these directions through time, with no particular interest in the distribution of the population. Thus, as has been done in earlier studies of GA niching [64], we adopt the performance metric called the *maximum peak ratio statistic*. This metric measures the quality as well as the number of optima given as a final result by the evolutionary algorithm. Explicitly, assuming a *minimization problem*, given the fitness values of the subpopulations in the final population $\left\{ \tilde{f}_i \right\}_{i=1}^{q}$, and the fitness values of the real optima of the objective function $\left\{ \hat{\mathcal{F}}_i \right\}_{i=1}^{q}$, the *maximum peak ratio* is defined as follows:

$$MPR = \frac{\sum_{i=1}^{q} \hat{\mathcal{F}}_i}{\sum_{i=1}^{q} \tilde{f}_i}, \tag{3.7}$$

where all values are assumed to be *strictly positive*. If this is not the case in the original parameterization of the landscape, the latter should be scaled accordingly with an additive constant for the sake of this calculation. Also, given a maximization problem, the MPR is defined as the sum of the obtained optima divided by the sum of the real optima. A drawback of this performance metric is that the real optima need to be known *a-priori*. However, for many artificial test problems these can be derived analytically, or tight numerical approximations to them are available.

Table 3.1: Test functions to be *minimized*, initialization domains and number of desired peaks. For some of the non-separable functions, we apply translation and rotation: $\vec{y} = \mathcal{O}(\vec{x} - \vec{r})$, where $\mathcal{O}$ is an orthogonal rotation matrix, and $\vec{r}$ is a shifting vector. See the note on implementation.

Separable:

Name	Function	Init	Niches
$\mathcal{M}$	$\mathcal{M}(\vec{x}) = -\frac{1}{n}\sum_{i=1}^{n}\sin^{\alpha}(5\pi x_i)$	$[0,1]^n$	100
$\mathcal{A}$ [Ackley]	$\mathcal{A}(\vec{x}) = -c_1 \cdot \exp\left(-c_2\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}\right)$ $-\exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos(c_3 x_i)\right) + c_1 + e$	$[-10,10]^n$	$2n+1$
$\mathcal{L}$	$\mathcal{L}(\vec{x}) = -\prod_{i=1}^{n}\sin^{k}(l_1\pi x_i + l_2)\cdot\exp\left(-l_3\left(\frac{x_i - l_4}{l_5}\right)^2\right)$	$[0,1]^n$	$n+1$
$\mathcal{R}$ [Rastrigin]	$\mathcal{R}(\vec{x}) = 10n + \sum_{i=1}^{n}\left(x_i^2 - 10\cos(2\pi x_i)\right)$	$[-1,5]^n$	$n+1$
$\mathcal{G}$ [Griewank]	$\mathcal{G}(\vec{x}) = 1 + \sum_{i=1}^{n}\frac{x_i^2}{4000} - \prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right)$	$[-10,10]^n$	5

Non-separable:

Name	Function	Init	Niches
$\mathcal{F}$ [Fletcher-Powell]	$\mathcal{F}(\vec{x}) = \sum_{i=1}^{n}(A_i - B_i)^2$ $A_i = \sum_{j=1}^{n}(a_{ij}\cdot\sin(\alpha_j) + b_{ij}\cdot\cos(\alpha_j))$ $B_i = \sum_{j=1}^{n}(a_{ij}\cdot\sin(x_j) + b_{ij}\cdot\cos(x_j))$ $a_{ij},\, b_{ij} \in [-100,100]\,;\quad \vec{\alpha} \in [-\pi,\pi]^n$	$[-\pi,\pi]^n$	10
$\mathcal{R}_{SR}$ [S.R. Rastrigin]	$\mathcal{R}_{SR}(\vec{x}) = 10n + \sum_{i=1}^{n}\left(y_i^2 - 10\cos(2\pi y_i)\right)$	$[-5,5]^n$	$n+1$
$\mathcal{G}_{SR}$ [S.R. Griewank]	$\mathcal{G}_{SR}(\vec{x}) = 1 + \sum_{i=1}^{n}\frac{y_i^2}{4000} - \prod_{i=1}^{n}\cos\left(\frac{y_i}{\sqrt{i}}\right)$	$[0,600]^n$	5

3.4.3 New Perspective: MPR vs. Time

Although the MPR metric was originally derived to be analyzed by means of the its saturation value, a new perspective was introduced by us in [90]. Our study investigated the MPR as a function of time, focusing on the early stages of the run. It was shown experimentally that the time-dependent MPR data fits a theoretical function: *the logistic curve*.

The Logistic Equation A simple modeling of the *organic population growth* is often described by the following differential equation:

$$\frac{dy}{dt} = cy\left(1 - \frac{y}{a}\right), \tag{3.8}$$

with the solution

$$y(t) = \frac{a}{1 + \exp\{c\,(t - T)\}}, \tag{3.9}$$

where a is the saturation value of the curve, T is its time shift, and c (in this context always negative) determines the shape of the exponential rise.

This equation, known as the *logistic equation*, describes many processes in nature. All those processes share the same pattern of behavior - growth with *acceleration*, followed by *deceleration* and then a *saturation* phase.

In the context of evolutionary niching methods, we argued [90] that the logistic parameters should be interpreted in the following way - T as the *learning period* of the algorithm, and the absolute value of c as its *niching formation acceleration*.

3.4.4 MPR Analysis: Previous Observation

This MPR time-dependent analysis was applied in [90] to two ES-based niching techniques: niching with the standard-ES according to the Schwefel-approach [94], and niching with the CMA-ES. In short, the standard-ES based method applies the same niching framework as the one described in this thesis except for one conceptual difference: it employs a (μ, λ) strategy in each niche, subject to *restricted mating*. Otherwise, it employs the standard ES operators.

We outline some of the conclusions of that study here:

1. The **niching formation acceleration**, expressed as the absolute value of c, had larger values for the CMA-ES mechanism for all the test-cases. That implied stronger niching acceleration and faster convergence.

2. A trend concerning the absolute value of c as a function of the dimensionality was observed: the higher the dimensionality, the lower the absolute value of c, i.e., the slower the niching process.

3. The **learning period**, expressed as the value of T in the curve fitting, got negative as well as positive values. Negative values mean that the niches formation process, expressed as the exponential rise of the MPR, started immediately from generation zero.

4. The averaged **saturation value** a, i.e., the MPR value, was larger in all of the test-cases for the CMA-ES mechanism. In that respect, the CMA kernel outperformed the standard-ES on the given landscapes.

The study concluded with the claim that there was a clear *trade-off*: Either a long learning period followed by a high niching acceleration (CMA-ES), or a short learning period followed by a low niching acceleration (Standard-ES).

3.5 Numerical Observation

We describe here our numerical observation with respect to the experimental results of our 5 niching variants on the proposed test suite.

3.5.1 Modus Operandi

The 5 niching algorithms are tested on the specified functions for various dimensions. Each test case includes 100 runs per algorithm. All runs are performed with a core mechanism of a $(1 \dot{+} 10)$-strategy per niche and initial points are sampled uniformly within the initialization intervals. Initial global step-sizes are set to $\frac{1}{4}$ of the intervals. The parameter q is set based on *a-priori* knowledge when available, or arbitrarily otherwise.
Function evaluations: the idea is to allocate a fixed number of evaluations per peak $\left(n \cdot 10^4\right)$, and thus each run is stopped after $q \cdot n \cdot 10^4$ function evaluations.

As mentioned earlier, setting the parameter p reflects the trade-off between further sampling the search-space, on the expense of exploiting the granted function evaluations at the existing attraction sites. Here, we set $p = 1$, which means emphasis on the latter.

A curve fitting routine is applied to each run in order to retrieve the characteristic parameters of its logistic curve. This routine uses the least-squared-error method, and runs an optimization procedure to minimize it.

3.5.2 Numerical Results

The numerical results are presented at several levels:

Niching Acceleration

Table 3.2 presents the mean and the standard deviations for the *absolute value* of the parameter c over 100 runs, as obtained by the curve fitting

routine. There is a clear trend in the given numerical results - in the vast majority of the test cases, the DR2 algorithm has the highest absolute values of c, whereas the CMA+ has the lowest absolute values. This trend corresponds to having the highest niching acceleration and the lowest niching acceleration, respectively. Moreover, the 4 comma strategies have absolute c values in the same order of magnitude, whereas the CMA+ typically has a lower absolute value in comparison to them.

MPR Saturation

This scalar value represents, to some degree, the quality of the obtained minima, and thus the final result of the niching process. Table 3.3 presents the mean and the standard deviation of the saturation MPR values for the different test cases. As can be seen in this table, the CMA-$\left(\frac{+}{,}\right)$ kernels achieve the highest MPR values, and thus they outperform together the other methods with respect to the niching process. However, for the given test cases, there is no clear winner for the MPR value.

Global Minimum

Table 3.4 contains the percentage of runs in which the global minimum was located. $\mathcal{M}$ is discarded from the table, as its global minimum was always found, by all algorithms, for every dimension n under investigation. Generally speaking, the CMA-$\left(\frac{+}{,}\right)$ routines, and in particular the CMA+ strategy, were superior with respect to the other derandomized variants.

One can also observe a strong correlation between Tables 3.3 and 3.4: Routines that obtain high MPR saturation values, i.e., locate the top-quality peaks, typically perform well globally and locate the global minimum in a high percentage of the runs.

The $c-T$ Tradeoff Hypothesis

We would like to numerically assess the hypothesis claiming the existence of a tradeoff between the learning period T and the niching acceleration c, as speculated in [90], with respect to the 5 algorithms under investigation.

We consider two test functions of the suite, one per class: the *separable* $\mathcal{M}$ and the *non-separable* $\mathcal{G}_{RS}$ (the *Shifted Rotated Griewank*). For each we run the algorithms for an increasing dimensionality of $n = 3, 4, \ldots, 30$, and obtain the MPR parameters for 100 runs - in order to plot c as a function of T.

Figures 3.1 and 3.2 present the $c-T$ curves for $\mathcal{M}$ and $\mathcal{G}_{RS}$, respectively. The curves reflect a clear trade-off between c and T over the dimensions for the algorithms for both cases (an exception: DR3 over $\mathcal{M}$). We consider this a numerical corroboration of the hypothesis: The longer the learning period, the lower the niching acceleration.

Table 3.2: The **absolute value** of the parameter c, obtained from curve fitting: Mean and standard deviation over 100 runs.

Test-Case	**DR1**	**DR2**	**DR3**	**CMA**	**CMA+**
$\mathcal{M} : n = 3$	0.107 ± 0.006	0.138 ± 0.009	0.106 ± 0.010	0.069 ± 0.005	0.054 ± 0.003
$\mathcal{M} : n = 10$	0.059 ± 0.002	0.072 ± 0.002	0.071 ± 0.003	0.040 ± 0.001	0.015 ± 0.001
$\mathcal{M} : n = 40$	0.027 ± 0.001	0.033 ± 0.001	0.024 ± 0.001	0.013 ± 0.001	0.003 ± 0.001
$\mathcal{A} : n = 3$	0.153 ± 0.038	0.226 ± 0.058	0.167 ± 0.006	0.135 ± 0.033	0.048 ± 0.006
$\mathcal{A} : n = 10$	0.063 ± 0.009	0.079 ± 0.013	0.071 ± 0.011	0.055 ± 0.011	0.017 ± 0.001
$\mathcal{L} : n = 3$	0.164 ± 0.070	0.194 ± 0.124	0.151 ± 0.064	0.148 ± 0.047	0.063 ± 0.030
$\mathcal{L} : n = 10$	0.150 ± 0.015	0.186 ± 0.024	0.143 ± 0.057	0.147 ± 0.016	0.040 ± 0.003
$\mathcal{R} : n = 3$	0.022 ± 0.032	0.035 ± 0.042	0.009 ± 0.012	0.030 ± 0.024	0.010 ± 0.011
$\mathcal{R} : n = 10$	0.046 ± 0.007	0.049 ± 0.010	0.039 ± 0.017	0.022 ± 0.007	0.016 ± 0.002
$\mathcal{G} : n = 3$	0.012 ± 0.014	0.025 ± 0.017	0.012 ± 0.003	0.023 ± 0.040	0.006 ± 0.012
$\mathcal{G} : n = 10$	0.031 ± 0.027	0.102 ± 0.020	0.031 ± 0.030	0.023 ± 0.003	0.019 ± 0.015
$\mathcal{F} : n = 3$	0.022 ± 0.023	0.042 ± 0.017	0.024 ± 0.024	0.023 ± 0.025	0.015 ± 0.012
$\mathcal{F} : n = 10$	0.054 ± 0.093	0.087 ± 0.105	0.078 ± 0.123	0.044 ± 0.083	0.022 ± 0.021
$\mathcal{R}_{RS} : n = 3$	0.157 ± 0.036	0.254 ± 0.053	0.178 ± 0.047	0.200 ± 0.041	0.055 ± 0.008
$\mathcal{R}_{RS} : n = 10$	0.072 ± 0.026	0.095 ± 0.019	0.083 ± 0.025	0.072 ± 0.027	0.020 ± 0.002
$\mathcal{G}_{RS} : n = 3$	0.108 ± 0.067	0.126 ± 0.074	0.118 ± 0.064	0.113 ± 0.069	0.050 ± 0.007
$\mathcal{G}_{RS} : n = 10$	0.056 ± 0.015	0.072 ± 0.015	0.085 ± 0.020	0.090 ± 0.012	0.020 ± 0.004

Table 3.3: The saturation MPR value: Mean and standard deviation over 100 runs.

Test-Case	**DR1**	**DR2**	**DR3**	**CMA**	**CMA+**
$\mathcal{M} : n = 3$	1 ± 0	1 ± 0	1 ± 0	1 ± 0	1 ± 0
$\mathcal{M} : n = 10$	1 ± 0	1 ± 0	1 ± 0	1 ± 0	1 ± 0
$\mathcal{M} : n = 40$	0.997 ± 0.002	1 ± 0	0.988 ± 0.003	1 ± 0	1 ± 0
$\mathcal{A} : n = 3$	0.971 ± 0.029	0.966 ± 0.028	0.960 ± 0.030	0.977 ± 0.024	0.992 ± 0.017
$\mathcal{A} : n = 10$	0.901 ± 0.024	0.905 ± 0.025	0.901 ± 0.025	0.920 ± 0.023	0.942 ± 0.023
$\mathcal{L} : n = 3$	0.963 ± 0.028	0.945 ± 0.038	0.953 ± 0.029	0.962 ± 0.027	0.996 ± 0.006
$\mathcal{L} : n = 10$	0.505 ± 0.163	0.379 ± 0.153	0.167 ± 0.129	0.596 ± 0.148	0.562 ± 0.109
$\mathcal{R} : n = 3$	0.263 ± 0.314	0.245 ± 0.036	0.233 ± 0.042	0.143 ± 0.046	0.481 ± 0.124
$\mathcal{R} : n = 10$	0.052 ± 0.007	0.063 ± 0.007	0.055 ± 0.005	0.057 ± 0.009	0.053 ± 0.005
$\mathcal{G} : n = 3$	0.115 ± 0.168	0.526 ± 0.470	0.366 ± 0.050	0.223 ± 0.288	0.761 ± 0.098
$\mathcal{G} : n = 10$	0.024 ± 0.042	0.026 ± 0.047	0.066 ± 0.018	0.015 ± 0.017	0.079 ± 0.029
$\mathcal{F} : n = 3$	0.002 ± 0.002	0.002 ± 0.002	0.002 ± 0.002	0.003 ± 0.004	0.002 ± 0.001
$\mathcal{F} : n = 10$	0.001 ± 0.001	0.001 ± 0.001	0.001 ± 0.001	0.001 ± 0.001	0.001 ± 0.001
$\mathcal{R}_{RS} : n = 3$	0.409 ± 0.111	0.463 ± 0.067	0.423 ± 0.117	0.469 ± 0.103	0.563 ± 0.098
$\mathcal{R}_{RS} : n = 10$	0.085 ± 0.015	0.099 ± 0.019	0.078 ± 0.015	0.108 ± 0.017	0.071 ± 0.014
$\mathcal{G}_{RS} : n = 3$	0.072 ± 0.043	0.078 ± 0.044	0.085 ± 0.048	0.082 ± 0.036	0.108 ± 0.041
$\mathcal{G}_{RS} : n = 10$	0.134 ± 0.038	0.144 ± 0.037	0.122 ± 0.035	0.161 ± 0.034	0.045 ± 0.013

Table 3.4: Global minimum reached in 100 runs.

Test-Case	**DR1**	**DR2**	**DR3**	**CMA**	**CMA+**
$\mathcal{A}: \ n = 3$	100%	100%	100%	100%	100%
$\mathcal{A}: \ n = 10$	90%	91%	90%	92%	95%
$\mathcal{L}: \ n = 3$	93%	74%	92%	97%	100%
$\mathcal{L}: \ n = 10$	9%	2%	0%	17%	13%
$\mathcal{R}: \ n = 3$	20%	19%	13%	16%	48%
$\mathcal{R}: \ n = 10$	0%	0%	0%	0%	0%
$\mathcal{G}: \ n = 3$	13%	21%	32%	13%	88%
$\mathcal{G}: \ n = 10$	8%	16%	4%	16%	2%
$\mathcal{F}: \ n = 3$	100%	100%	100%	100%	100%
$\mathcal{F}: \ n = 10$	14%	12%	15%	23%	15%
$\mathcal{R}_{RS}: \ n = 3$	45%	40%	39%	54%	72%
$\mathcal{R}_{RS}: \ n = 10$	0%	0%	0%	0%	0%
$\mathcal{G}_{RS}: \ n = 3$	4%	2%	4%	12%	8%
$\mathcal{G}_{RS}: \ n = 10$	6%	1%	3%	14%	0%

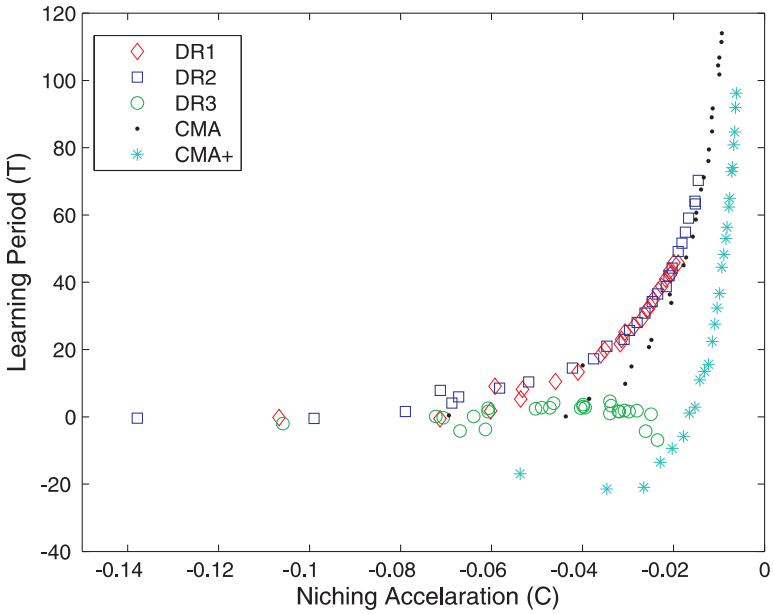

Figure 3.1: The $c - T$ curve for $\mathcal{M}$: A clear trade-off for the different algorithms, except for DR3, which has a flat curve. Each data point is an average of 100 runs, given $n = 3, 4, \ldots, 30$.

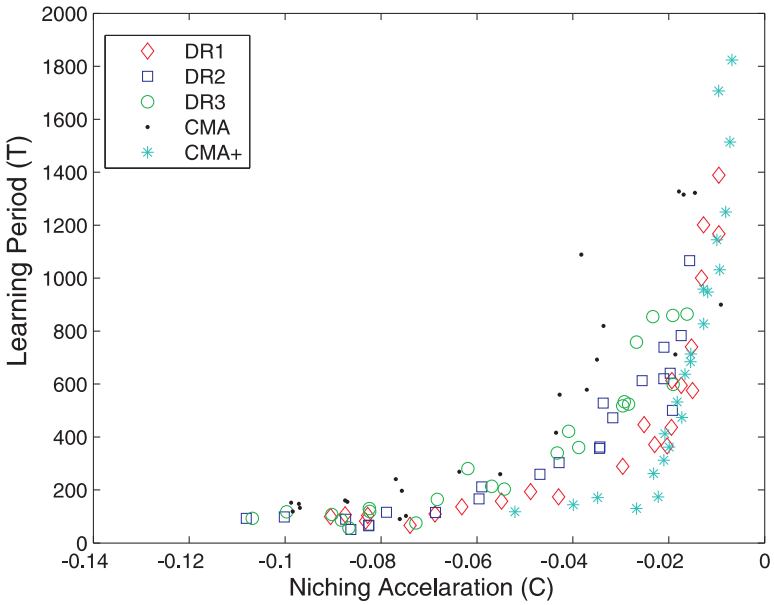

Figure 3.2: The $c - T$ curve for $\mathcal{G}_{RS}$: A clear trade-off for the 5 different algorithms. Each data point is an average of 100 runs, given $n = 3, 4, \ldots, 30$.

3.5.3 Discussion

The elitist CMA strategy was observed to perform very well in the proposed niching framework. A straightforward and rather intuitive explanation for that would be its tendency to maintain convergence in any basin of attraction, versus a higher probability for the comma strategy to escape them. Moreover, we would like to suggest another argument for the advantage of an elitist strategy for niching. The niching problem can be considered as an optimization task with constraints, i.e., the formation of niches that restricts competing niches and their optimization routines from exploring the search space freely. It has been suggested in previous studies (see, e.g., [87]) that ES self-adaptation in constrained problems will tend to fail with a comma-strategy, and thus a plus-strategy is preferable for such problems. We might link this argumentation to the observation of our numerical results here, and suggest that an elitist strategy is preferable for niching.

Adaptability is not imitation.
It means power of resistance and assimilation.
Mahatma Gandhi

Chapter 4

Self-Adaptive Niche-Radii and Niche-Shape Approaches

4.1 General

While the motivation and usefulness of niching cast no doubt, the relaxation of assumptions and limitations concerning the hypothetical landscape is much needed if niching methods are to be valid in a broader range of applications. In short, we choose to treat in this chapter the particular limiting assumption of the fixed niche radius by introducing self-adapting niche-radii and niche-shape mechanisms.

More specifically, niching techniques are often subject to criticism due to the so-called *niche radius problem*. The majority of the niching methods make an *assumption* concerning the fitness landscape, stating that the optima are far enough from one another with respect to the so-called *niche radius*, which is estimated for the given problem and remains fixed during the course of evolution, as outlined in Section 3.3. Obviously, there are landscapes for which this assumption is not applicable, and where this approach is most likely to fail (see Figures 4.1 and 4.2 for illustration). As discussed earlier, the task of defining a generic basin of attraction seems to be one of the most difficult problems in the field of global optimization.

4.1.1 Related Work

There were several GA-oriented studies which addressed this so-called *niche radius problem*, aiming to relax the assumption specified earlier, or even to drop it completely. Jelasity [63] suggested a cooling-based mechanism for the niche-radius, also known as the UEGO, which adapts the global radius as a function of time during the course of evolution. Gan and Warwick [72] introduced the so-called Dynamic Niche Clustering, to overcome the radius problem by using a clustering mechanism. A complex subpopulation

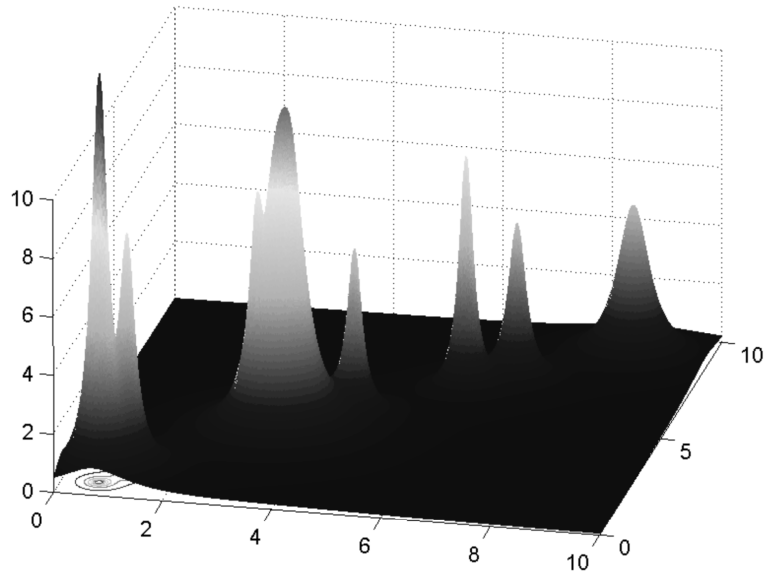

Figure 4.1: The Shekel function (see, e.g., [8]) in a $2D$ decision space: Introducing a dramatically uneven spread of optima; For more details see Table 3.1.

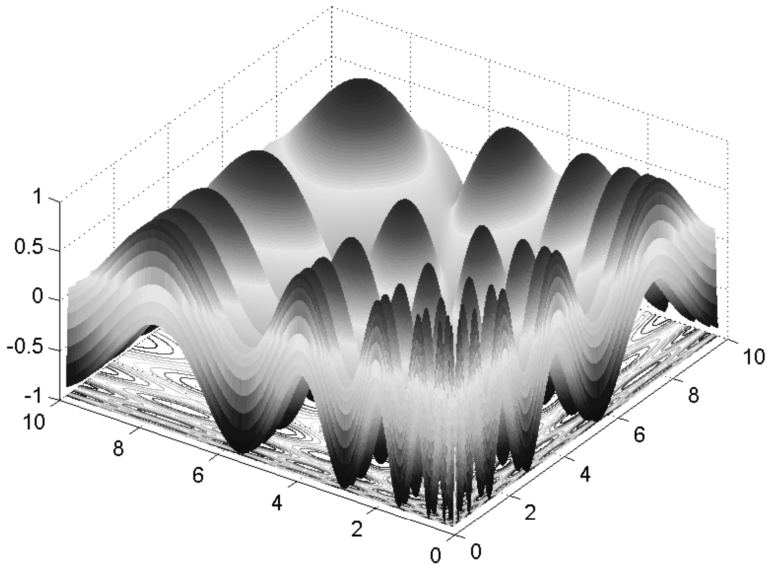

Figure 4.2: The Vincent function in a $2D$ decision space: A sine function with a decreasing frequency.

differentiation model, the so-called Multinational Evolutionary Algorithm, was presented by Ursem [84]. It introduces a topological-based auxiliary mechanism of sampling, which detects whether feasible solutions share the same basin of attraction. A recent study by Stoean et al. [95] considered the hybridization of the latter with a radius-based niching method proposed in [85]. Finally, an iterative statistical-based approach was introduced lately [66] for learning the optimal niche radius, without a-priori knowledge of the landscape. It considers the *fitness sharing* strategy, and optimizes it as a function of the population size and the niche radius, without relaxing the landscape assumption specified earlier – i.e., the niches are eventually obtained using a single fixed niche radius.

4.1.2 Our Approach

Our study introduces a new concept into the niche radius problem, inspired by the ES self-adaptation concept - an **adaptive individual niche radius**. The idea is that each individual, i.e., feasible solution in the artificial population, updates every generation a niche radius along with its adaptive strategy parameters. This study is an "adaptive extension" to niching with the CMA-ES.

Two new approaches are presented here. The first exploits the self-adaptation of the step-size in the CMA-ES mechanism, the *cumulative step-size adaptation* (CSA) mechanism, and couples the individual niche-radius to it. Since the step-size does not hold any further spatial information concerning the landscape, the classification into niches uses hyperspheres, based on the *Euclidean distance*. The second approach introduces the *Mahalanobis distance* into the niching mechanism, aiming to allow more accurate spatial classification by using ellipsoids which are based upon the evolving distribution, rather than the uniform hyperspheres of the *Euclidean metric*. This idea can be easily implemented into the CMA-ES niching routines, since the covariance matrix of the distribution — an essential component of the Mahalanobis distance — is already learned by the algorithm. These two new approaches are tested with the CMA-$(\dagger,\dagger)$ routines, and evaluated on a suite of artificial landscapes, including problems with an uneven spread of optima as well as with non-isotropic attractor basins.

4.2 New Proposed Approaches

In this section we present two new approaches for the adaptation of the niches classification mechanism, in the framework of niching with the CMA-ES. Section 4.2.1 presents the self-adaptive niche radius mechanism which is based upon the coupling to the step-size, and Section 4.2.2 introduces niching with the *Mahalanobis distance*, relying on the evolving covariance matrix.

4.2.1 Self-Adaptive Radius: Step-Size Coupling

Aiming to follow the successful mechanism of the step-size adaptation, the idea of this approach is to *couple* the niche radius to the global step-size σ, whereas the *indirect selection* of the niche radius is governed by the objective that every niche should ideally consist of λ individuals. This is implemented by means of a quasi *dynamic fitness sharing* mechanism. A detailed description follows.

The Niching-CMA method is used as outlined earlier (Chapter 3), with the following modifications. q is given as an input to the algorithm, but it is now merely a prediction or a demand for the maximal number of solutions the decision maker would like to obtain. Given the i^{th} individual in the population, a niche radius denoted by ρ_i^0 is initialized by means of a rule ($\rho_i^0 = \sqrt{n} \cdot \sigma_{init}$) in the beginning of the search. Its update step in generation $(g+1)$ is based on the parent's radius and step-size:

$$\rho_i^{(g+1)} = \left(1 - c_i^{(g+1)}\right) \cdot \rho_{parent}^{(g)} + c_i^{(g+1)} \cdot \sqrt{n} \cdot \sigma_{parent}^{(g+1)} \tag{4.1}$$

where $c_i^{(g)} \in [0, 1)$ is the individual learning coefficient. The latter is updated by means of the *step-size difference*, i.e., $\Delta\sigma_i^{(g+1)} = \left|\sigma_{parent}^{(g+1)} - \sigma_{parent}^{(g)}\right|$:

$$c_i^{(g+1)} = \gamma \cdot \left(1 - \exp\left\{-\alpha \cdot \Delta\sigma_i^{(g+1)}\right\}\right) \tag{4.2}$$

See Figure 4.3 for an illustration. As for the constants, γ and α are set differently for the two selection strategies:

$$\gamma = \begin{cases} \frac{1}{5} & \text{for } (1,\lambda)\text{-selection} \\ \frac{4}{5} & \text{for } (1+\lambda)\text{-selection} \end{cases} \quad \alpha = \begin{cases} 10 & \text{for } (1,\lambda)\text{-selection} \\ 100 & \text{for } (1+\lambda)\text{-selection} \end{cases} . \tag{4.3}$$

γ determines the saturation value of the learning coefficient: Strong coupling to the parent's step-size for the plus strategy, versus a weak coupling for the comma strategy. α dictates the strength of the exponential convergence towards the saturation value: Slow convergence for the plus strategy, versus a rapid convergence for the comma strategy. This rule for parametric setting works reliably on a wide range of problems, as we will show later. The rational behind it stems from the different niching convergence behavior of the two strategies, as was already discussed in Section 3.5. Furthermore, we shall discuss the use of new parameters in Section 4.4.

The DPI routine (Algorithm 4 is run using the **individual niche radii**, for the identification of the peaks and the classification of the population.

Furthermore, introduce:

$$g(x,\lambda) = 1 + \Theta(\lambda - x) \cdot \frac{(\lambda - x)^2}{\lambda} + \Theta(x - \lambda) \cdot (\lambda - x)^2 , \tag{4.4}$$

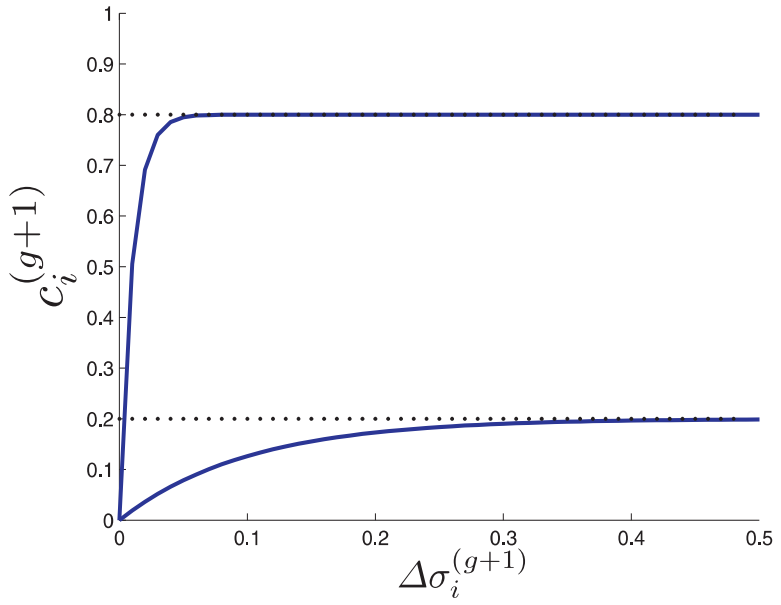

Figure 4.3: The learning coefficient $c_i^{(g+1)}$ (Eq. 4.2) is plotted as a function of the *step-size difference*, $\Delta\sigma_i^{(g+1)}$, for the two strategies, as derived from Eq. 4.3 – (γ, α) are substituted for the two strategies: $\left\{\left(\frac{1}{5}, 10\right), \left(\frac{4}{5}, 100\right)\right\}$.

where $\Theta(y)$ is the *Heaviside step function*. Given a fixed λ, $g(x, \lambda)$ is a parabola with unequal branches, centered at $(x = \lambda,\ g = 1)$ (see Figure 4.4 for illustration). The justification for its geometrical asymmetry will be described shortly. Then, by applying the calculation of the *dynamic niche count* m_i^{dyn} (Eq. 2.13), based on the appropriate radii, we **define** the *niche fitness* of individual i by:

$$f_i^{niche} = \frac{f_i}{g\left(m_i^{dyn}, \lambda\right)} \tag{4.5}$$

We assume, again, that the raw fitness is strictly positive and subject to maximization. Finally, the selection of the next parent in each niche, i.e., the so-called *alpha-male* of the local site, is based on this *niche fitness*.

Eq. 4.5 enforces the requirement for having a fixed resource of λ individuals per niche, since $g(x, \lambda)$ yields values greater than 1 for any niche count different than λ. The asymmetry of $g(x, \lambda)$ is therefore meant to penalize more the niches which exceed λ members, in comparison to those with less than λ members. This equation is a variant of the dynamic shared fitness (Eq. 2.14), and is used now in the context of niche radius adaptation.

The *self-adaptive* niching routine is presented in Algorithm 7.

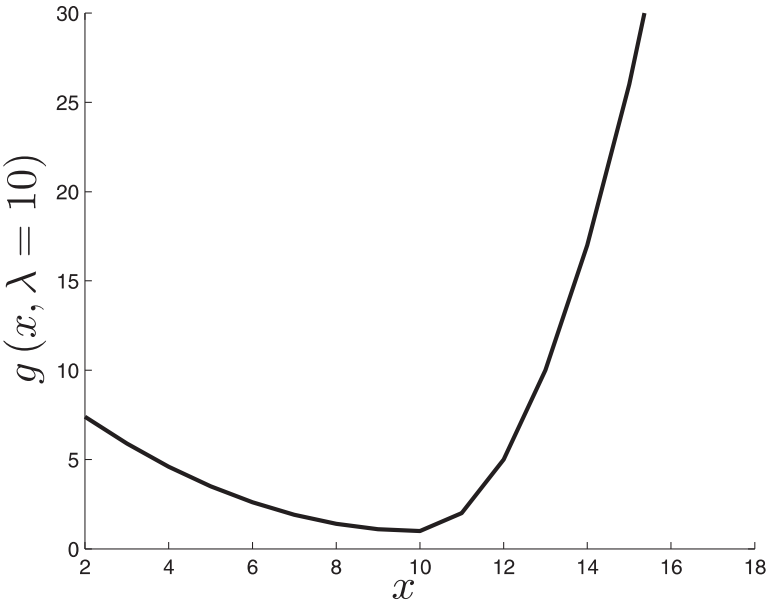

Figure 4.4: An illustration for $g\,(x, \lambda = 10)$ (Eq. 4.4): A parabola with uneven branches; The *niche fitness* (Eq. 4.5) is penalized more for an overpopulated niche ($\lambda > 10$) due to the steep branch, in comparison to an underpopulated niche.

Algorithm 7 Niching-CMA with an Adaptive Niche Radius

1: **for** $i = 1 \ldots (q + p)$ search points **do**
2: Generate λ samples based on the CMA-set of i
3: Update the niche radius ρ_i^{g+1} according to Eq. 4.1
4: **end for**
5: Evaluate fitness of the population
6: Compute the DPS with the DPI Algorithm, based on individual radii
7: Compute the Dynamic Niche Count of every individual
8: **for all** elements of DPS **do**
9: Compute the Niche Fitness (Eq. 4.5)
10: Set individual with best niche fitness as a search point
11: Inherit the CMA-set and update it respectively
12: **end for**
13: **if** N_{DPS} =size of $DPS < q$ **then**
14: Generate $q - N_{DPS}$ new search points, reset CMA-sets
15: **end if**
16: **if** $gen \mod \kappa \equiv 0$ **then**
17: Resample the $(q + 1)^{th} \ldots (q + p)^{th}$ search points
18: **end if**

4.2.2 Mahalanobis Metric: Covariance Exploitation

Existing niching techniques, and in particular those presented in Chapter 3 and Section 4.2.1, use the *Euclidean distance* in the decision space for the classification of feasible solutions to the niches under formation. This approach is likely to encounter problems in high-dimensional landscapes with non-isotropic basins of attraction. Since the CMA-ES algorithm already learns the covariance matrix of the decision space distribution, it is worthwhile to use it for a better spatial classification mechanism within the niching framework. In essence, this can be considered as an upgrade of the niching mechanism, as it captures a more accurate spatial formation of the niches. **Most importantly, this approach is also self-adaptive.**

After giving this motivation, we proceed with discussing the details of this idea.

The Mahalanobis Distance In the following, we consider the *Mahalanobis distance*, for instance in a probability distribution. Given a *mean vector* $\vec{m}$ and a *covariance matrix* Σ, the Mahalanobis distance of a vector $\vec{v}$ from the *mean vector* is defined as:

$$d(\vec{v}, \vec{m}) = \sqrt{(\vec{v} - \vec{m})^T \Sigma^{-1} (\vec{v} - \vec{m})} \tag{4.6}$$

It can be shown that the *iso-distance surfaces* of this metric are ellipsoids which are centered about the *mean* $\vec{m}$. In the special case where $\Sigma \sim \mathbf{I}$ (e.g., features are uncorrelated and all variances equal) the Mahalanobis distance reduces to the normalized *Euclidean* distance, and the iso-distance surfaces become Euclidean hyperspheres. Though the Mahalanobis distance is typically applied in statistics, it can also be applied in different contexts as a metric on vector spaces given a positive-semidefinite and symmetric matrix Σ determining the elliptic iso-distance surfaces.

Mahalanobis CMA-ES Niching

In the context of *niching*, given an individual $\vec{x}$, representing a niche with a covariance matrix $\mathbf{C_x}$, we choose to define, accordingly, the Mahalanobis distance of an individual $\vec{y}$ to the niche by

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \mathbf{C_x}^{-1} (\vec{x} - \vec{y})}\,.$$

Since different individuals have different covariance matrices, this operation is *asymmetric*. Hence, the actual classification into niches depends not only on the identity of the so-called *peak individuals*, which are selected according to their higher fitness, but also on their individual covariance matrices. Due to the fact that the classification itself is carried out individually by means of independently evolving distance measures, an equivalent classification by

means of the Euclidean metric would possibly result in a different outcome when compared to this approach.

Notably, the proposed routine does not have a secondary selection mechanism, which was necessary for the self-adaptive niche radius approach, as introduced in Section 4.2.1. The reason why it is not required here is that the local shape of the attractor basins, as approximated by the CMA, is equivalent to the desired shape for the niches, and thus sufficient for successful classification of individuals to the niche.

Numerical Implementation

As for the technical details, we discuss here the numerical implementation of the Mahalanobis metric, considering the *matrix inversion* which is required. We show here that the *matrix inversion*, in this context, can be replaced by *matrix multiplication* - which leads to a significant performance gain for the dimensions that are typically under study.

In the CMA-ES mechanism, the *eigenvalue-decomposition* of the covariance matrix $\mathbf{C}$, which is calculated every generation, reads

$$\mathbf{C} = \mathbf{BD}\left(\mathbf{BD}\right)^T, \tag{4.7}$$

where $\mathbf{D} = diag\left(\sqrt{\Lambda_1}, \sqrt{\Lambda_2}, ..., \sqrt{\Lambda_n}\right)$, with the *eigenvalues* $\{\Lambda_i\}_{i=1}^n$. In order to obtain $\mathbf{C}^{-1}$, one can derive,

$$\begin{aligned} \mathbf{C}^{-1} &= \left[\mathbf{BD}\left(\mathbf{BD}\right)^T\right]^{-1} = \mathbf{B}^{T\ -1}\mathbf{D}^{T\ -1}\mathbf{D}^{-1}\mathbf{B}^{-1} = \\ &\mathbf{B} \cdot diag\left(\frac{1}{\Lambda_1}, \frac{1}{\Lambda_2}, ..., \frac{1}{\Lambda_n}\right) \cdot \mathbf{B}^T \end{aligned} \tag{4.8}$$

and thus the *matrix inversion* calculation can be replaced, within the CMA-ES routine, by a *matrix multiplication* calculation.

Despite the fact that these two operations are equivalent in terms of numerical complexity (see, e.g., [96]), we observe in practice a difference between the two procedures for obtaining $\mathbf{C}^{-1}$. For dimensions up to $n = 30$, it is observed that the multiplication procedure takes on average half the calculation time in comparison to the inversion procedure[1]. Hence, it pays off to follow the derivation given here.

Due to numerical features of the eigenvalue-decomposition, which were also discussed by Hansen et al. (see [16], pp. 20), but are crucial here for the inversion operation of the covariance matrix, we introduce a lower bound to the eigenvalues: $\Lambda_{min} = 10^{-10}$.

[1] The calculations were done with MATLAB 7.0.

Table 4.1: Additional test functions to be *minimized* and initialization domains.

Name	Function	Init	Niches
$\mathcal{S}$ [Shekel]	$\mathcal{S}(\vec{x}) = -\sum_{i=1}^{10} \frac{1}{k_i(\vec{x}-a_i)(\vec{x}-a_i)^T+c_i}$	$[0,10]^n$	8
$\mathcal{V}$ [Vincent]	$\mathcal{V}(\vec{x}) = -\frac{1}{n}\sum_{i=1}^{n} \sin(10 \cdot \ln(x_i))$	$[0.25,10]^n$	50

Self-Adaptive Mahalanobis Approach

The self-adaptive niche radius mechanism presented in Section 4.2.1, can easily be adjusted to employ the Mahalanobis distance for the classification of the niches. In the context of this study, it would become a hybrid approach in the sense that it applies both a self-adaptive niche radius and a self-adaptive distance metric for the sake of the classification phase. This hybridization will also be considered in the experimental procedure as an independent niching routine.

4.3 Experimental Procedure

We apply the same experimental setup of Chapter 3, with the following modifications:

- We consider additional test-functions with an uneven spread of optima, introducing a challenge in the light of the niche radius problem:

  1. The Vincent function is a sine function with a decreasing frequency. It has 6^n global optima in the interval $[0.25,10]^n$.
  2. The Shekel function, suggested in [8], introduces a landscape with a dramatically uneven spread of optima. It has one global optimum, and 7 ordered local optima. The Shekel data was retrieved from [8].

  Table 4.1 is an extension to Table 3.1, summarizing the additional test-functions.

- In order to keep the behavior as simple as possible, the parameter p is set here to $p = 0$ (no so-called restart mechanism).

- We keep the same experimental framework of function evaluations granted per niche: $n \cdot 10^4$ function evaluations are allocated per niche, and thus a run is terminated after $q \cdot n \cdot 10^4$ function evaluations.

4.3.1 Numerical Observation

We discuss here the performance analysis at three levels:

Global Minimum

Table 4.2 contains the percentage of runs in which the global minimum was located. $\mathcal{M}$ and $\mathcal{V}$ are discarded from the table, as their global minimum was always found, by all algorithms, for every dimension n under investigation. For the *comma* strategy (four left columns), we observe that the Mahalanobis metric usually improves the global optimization – both for the fixed, as well as for the self-adaptive niche radius approaches. On the other hand, this does not seem to be the general trend for the *plus* strategy - on average the employment of the Mahalanobis distance does not improve the global optimization. We may conclude that there is no clear 'winner', and that the routines employing the Mahalanobis distance do not achieve a dramatic improvement in global optimization. This is an expected result, as the employment of this metric assists in the formation of the niches.

MPR Saturation

Tables 4.3 and 4.4 present the mean and the standard deviation of the saturation MPR values for the different test cases.

We observe a trend of better performance for the routines employing the Mahalanobis distance for both strategies. On average, the MPR values are higher, reflecting a better niching process.

Note that the niching routines, except for the fixed niche radius case, fail on the Ackley landscape, i.e., they locate only the global minimum, where all other niches are located in the global basin of attraction. This effect can be explained by the strong basin of attraction of the global minimum, in comparison to the sub-optimal minima.

Moreover, most of the MPR values for the Fletcher-Powell and shifted-rotated Griewank test-cases are much lower than unity, due to the extreme scaling of the landscape: It has false traps with very high function values. Thus, upon being trapped in these local minima, the MPR value is expected to be very low.

Niching Acceleration

The MPR analysis allows us to compare the *niching acceleration* of the different routines. Tables 4.5 and 4.6 present the mean values and the standard deviation of the niching acceleration values for the different test cases, by means of the *absolute value* of the parameter c of Eq. 3.9. The curve-fitting routine did not attain data with acceptable high quality for the Fletcher-Powell test-case, and it suffered from extremely large standard deviations. We thus choose to discard it from this table.

There are some general trends in the attained data. The comma strategy has typically higher niching acceleration values, as expected from previous observations (Chapter 3). Within each strategy, there is a trend of higher

Table 4.2: Global minimum reached in 100 runs (CMA denotes a $(1, \lambda)$-strategy, CMA+ denotes a $(1 + \lambda)$-strategy). The best result, per strategy, is emphasized in bold scripts.

Test-Case	CMA	M-CMA	S-CMA	MS-CMA	CMA+	M-CMA+	S-CMA+	MS-CMA+
$\mathcal{A} : n = 3$	100%	100%	100%	100%	100%	100%	100%	100%
$\mathcal{A} : n = 10$	94%	100%	100%	100%	97%	100%	100%	100%
$\mathcal{L} : n = 3$	64%	**66%**	43%	54%	**94%**	89%	65%	70%
$\mathcal{L} : n = 10$	**16%**	8%	2%	13%	**9%**	5%	1%	0%
$\mathcal{R} : n = 3$	54%	**59%**	13%	40%	**67%**	62%	14%	30%
$\mathcal{R} : n = 10$	0%	0%	0%	0%	0%	0%	0%	0%
$\mathcal{G} : n = 3$	12%	19%	10%	**25%**	19%	19%	16%	**52%**
$\mathcal{G} : n = 10$	20%	**31%**	27%	27%	0%	0%	0%	0%
$\mathcal{S} : n = 5$	91%	97%	82%	**100%**	83%	62%	**98%**	91%
$\mathcal{S} : n = 10$	21%	48%	46%	90%	97%	92%	**100%**	75%
$\mathcal{F} : n = 4$	100%	100%	100%	100%	100%	100%	100%	100%
$\mathcal{F} : n = 10$	25%	40%	36%	**46%**	17%	22%	34%	**37%**
$\mathcal{R}_{SR} : n = 3$	46%	**54%**	14%	24%	50%	**66%**	10%	26%
$\mathcal{R}_{SR} : n = 10$	**8%**	2%	0%	0%	0%	0%	0%	0%
$\mathcal{G}_{SR} : n = 3$	**10%**	0%	0%	0%	**9%**	0%	0%	0%
$\mathcal{G}_{SR} : n = 10$	0%	0%	0%	0%	0%	0%	0%	0%

Table 4.3: MPR saturation values for the $(1, \lambda)$-Strategy: Mean values and standard deviations over 100 runs. Emphasized in bold-script are winner algorithms with respect to the specified landscape, also in reference to the results of Table 4.4. Landscapes with several winners do not apply bold scripts.

Test-Case	**CMA**	**M-CMA**	**S-CMA**	**MS-CMA**
$\mathcal{M} : n = 3$	1 ± 0	1 ± 0	1 ± 0	1 ± 0
$\mathcal{M} : n = 10$	0.994 ± 0.002	0.967 ± 0.003	1 ± 0	1 ± 0
$\mathcal{M} : n = 40$	0.956 ± 0.006	0.953 ± 0.008	0.994 ± 0.001	0.995 ± 0.002
$\mathcal{A} : n = 3$	0.938 ± 0.044	N.A.	0.860 ± 0.143	N.A.
$\mathcal{A} : n = 10$	0.909 ± 0.033	N.A.	N.A.	N.A.
$\mathcal{L} : n = 3$	0.864 ± 0.092	0.870 ± 0.106	0.713 ± 0.083	0.834 ± 0.099
$\mathcal{L} : n = 10$	0.240 ± 0.086	0.389 ± 0.114	0.478 ± 0.080	$\mathbf{0.564 \pm 0.105}$
$\mathcal{R} : n = 3$	0.301 ± 0.081	0.228 ± 0.063	0.159 ± 0.041	0.305 ± 0.103
$\mathcal{R} : n = 10$	$\mathbf{0.103 \pm 0.045}$	0.062 ± 0.011	0.082 ± 0.019	0.094 ± 0.022
$\mathcal{G} : n = 3$	0.249 ± 0.126	0.234 ± 0.045	0.283 ± 0.092	0.255 ± 0.064
$\mathcal{G} : n = 10$	$\mathbf{0.252 \pm 0.169}$	0.195 ± 0.040	0.186 ± 0.092	0.190 ± 0.041
$\mathcal{S} : n = 5$	0.840 ± 0.320	0.911 ± 0.307	0.819 ± 0.300	$\mathbf{0.979 \pm 0.067}$
$\mathcal{S} : n = 10$	0.820 ± 0.722	0.931 ± 0.073	0.596 ± 0.136	$\mathbf{0.959 \pm 0.062}$
$\mathcal{V} : n = 3$	0.972 ± 0.011	0.920 ± 0.005	0.613 ± 0.028	0.552 ± 0.078
$\mathcal{V} : n = 10$	0.998 ± 0.007	0.998 ± 0.001	0.999 ± 0.001	$\mathbf{1 \pm 0}$
$\mathcal{F} : n = 4$	0.0004 ± 0.001	0.0049 ± 0.005	0.0005 ± 0.001	0.0173 ± 0.092
$\mathcal{F} : n = 10$	0.0001 ± 0.001	0.0002 ± 0.001	0.0003 ± 0.001	0.0004 ± 0.001
$\mathcal{R}_{SR} : n = 3$	0.331 ± 0.103	0.231 ± 0.041	0.138 ± 0.051	0.268 ± 0.074
$\mathcal{R}_{SR} : n = 10$	$\mathbf{0.130 \pm 0.039}$	0.087 ± 0.042	0.069 ± 0.019	0.093 ± 0.018
$\mathcal{G}_{SR} : n = 3$	0.0009 ± 0.001	0.0010 ± 0.001	0.0007 ± 0.001	0.0010 ± 0.001
$\mathcal{G}_{SR} : n = 10$	0.0001 ± 0	0.0001 ± 0	0.0001 ± 0	0.0001 ± 0

Table 4.4: MPR saturation values for the $(1 + \lambda)$-Strategy: Mean values and standard deviations over 100 runs. Emphasized in bold-script are winner algorithms with respect to the specified landscape, also in reference to the results of Table 4.3. Landscapes with several winners do not apply bold scripts.

Test-Case	**CMA+**	**M-CMA+**	**S-CMA+**	**MS-CMA+**
$\mathcal{M} : n = 3$	1 ± 0	1 ± 0	1 ± 0	1 ± 0
$\mathcal{M} : n = 10$	0.991 ± 0.003	0.986 ± 0.003	1 ± 0	1 ± 0
$\mathcal{M} : n = 40$	0.975 ± 0.008	0.980 ± 0.007	1 ± 0	1 ± 0
$\mathcal{A} : n = 3$	0.989 ± 0.026	$\mathbf{0.999 \pm 0.009}$	0.930 ± 0.030	0.937 ± 0.159
$\mathcal{A} : n = 10$	0.946 ± 0.017	$\mathbf{0.987 \pm 0.019}$	N.A.	N.A.
$\mathcal{L} : n = 3$	0.959 ± 0.033	$\mathbf{0.962 \pm 0.036}$	0.819 ± 0.079	0.919 ± 0.065
$\mathcal{L} : n = 10$	0.454 ± 0.116	0.373 ± 0.115	0.423 ± 0.108	0.432 ± 0.090
$\mathcal{R} : n = 3$	0.528 ± 0.118	$\mathbf{0.552 \pm 0.107}$	0.163 ± 0.072	0.250 ± 0.089
$\mathcal{R} : n = 10$	0.102 ± 0.040	0.077 ± 0.027	0.049 ± 0.009	0.053 ± 0.011
$\mathcal{G} : n = 3$	0.326 ± 0.094	0.334 ± 0.101	0.305 ± 0.114	$\mathbf{0.494 \pm 0.234}$
$\mathcal{G} : n = 10$	0.037 ± 0.008	0.053 ± 0.015	0.062 ± 0.019	0.060 ± 0.015
$\mathcal{S} : n = 5$	0.681 ± 0.114	0.897 ± 0.109	0.920 ± 0.073	0.882 ± 0.086
$\mathcal{S} : n = 10$	0.658 ± 0.054	0.957 ± 0.104	0.916 ± 0.311	0.939 ± 0.085
$\mathcal{V} : n = 3$	0.962 ± 0.012	$\mathbf{0.999 \pm 0.001}$	0.815 ± 0.072	0.689 ± 0.114
$\mathcal{V} : n = 10$	0.953 ± 0.016	0.990 ± 0.004	0.996 ± 0.002	0.999 ± 0.001
$\mathcal{F} : n = 4$	0.0007 ± 0.001	0.862 ± 0.385	0.0044 ± 0.002	$\mathbf{0.991 \pm 0.038}$
$\mathcal{F} : n = 10$	0.0001 ± 0.001	0.0001 ± 0.001	$\mathbf{0.0005 \pm 0.001}$	0.0001 ± 0.001
$\mathcal{R}_{SR} : n = 3$	0.486 ± 0.137	$\mathbf{0.563 \pm 0.140}$	0.135 ± 0.051	0.249 ± 0.129
$\mathcal{R}_{SR} : n = 10$	0.081 ± 0.030	0.080 ± 0.018	0.044 ± 0.006	0.041 ± 0.006
$\mathcal{G}_{SR} : n = 3$	0.0009 ± 0.001	0.0007 ± 0.001	0.008 ± 0.001	$\mathbf{0.0012 \pm 0.002}$
$\mathcal{G}_{SR} : n = 10$	0.0002 ± 0	0.0002 ± 0	0.0002 ± 0	0.0002 ± 0

Table 4.5: Niching acceleration values for the $(1, \lambda)$-Strategy: Mean values and standard deviations of the **absolute value** of c over 100 runs.

Test-Case	**CMA**	**M-CMA**	**S-CMA**	**MS-CMA**
$\mathcal{M} : n = 3$	0.068 ± 0.010	0.069 ± 0.002	0.049 ± 0.007	0.060 ± 0.008
$\mathcal{M} : n = 10$	0.038 ± 0.001	0.043 ± 0.002	0.029 ± 0.002	0.032 ± 0.001
$\mathcal{M} : n = 40$	0.014 ± 0.001	0.014 ± 0.001	0.010 ± 0.001	0.010 ± 0.001
$\mathcal{A} : n = 3$	0.133 ± 0.015	N.A.	0.035 ± 0.013	N.A.
$\mathcal{A} : n = 10$	0.063 ± 0.002	N.A.	N.A.	N.A.
$\mathcal{L} : n = 3$	0.179 ± 0.038	0.184 ± 0.048	0.128 ± 0.044	0.167 ± 0.036
$\mathcal{L} : n = 10$	0.174 ± 0.024	0.176 ± 0.025	0.144 ± 0.016	0.153 ± 0.019
$\mathcal{R} : n = 3$	0.043 ± 0.007	0.131 ± 0.109	0.045 ± 0.027	0.125 ± 0.058
$\mathcal{R} : n = 10$	0.043 ± 0.013	0.052 ± 0.012	0.064 ± 0.016	0.081 ± 0.015
$\mathcal{G} : n = 3$	0.079 ± 0.079	0.112 ± 0.033	0.097 ± 0.080	0.152 ± 0.095
$\mathcal{G} : n = 10$	0.001 ± 0.002	0.006 ± 0.002	1.051 ± 6.983	1.120 ± 5.418
$\mathcal{S} : n = 5$	0.004 ± 0.005	0.019 ± 0.009	0.080 ± 0.056	0.072 ± 0.020
$\mathcal{S} : n = 10$	0.004 ± 0.010	0.003 ± 0.005	0.012 ± 0.024	0.005 ± 0.004
$\mathcal{V} : n = 3$	0.004 ± 0.004	0.104 ± 0.010	0.010 ± 0.027	1.023 ± 2.018
$\mathcal{V} : n = 10$	0.004 ± 0.009	0.037 ± 0.024	0.055 ± 0.002	0.061 ± 0.003
$\mathcal{R}_{SR} : n = 3$	0.079 ± 0.068	0.153 ± 0.098	0.031 ± 0.019	0.113 ± 0.042
$\mathcal{R}_{SR} : n = 10$	0.077 ± 0.029	0.087 ± 0.032	0.051 ± 0.011	0.069 ± 0.010
$\mathcal{G}_{SR} : n = 3$	0.147 ± 0.088	0.150 ± 0.076	0.274 ± 0.284	0.129 ± 0.076
$\mathcal{G}_{SR} : n = 10$	0.101 ± 0.046	0.107 ± 0.045	0.204 ± 0.297	0.196 ± 0.276

Table 4.6: Niching acceleration values for the $(1 + \lambda)$-Strategy: Mean values and standard deviations of the **absolute value** of c over 100 runs.

Test-Case	**CMA**	**M-CMA**	**S-CMA**	**MS-CMA**
$\mathcal{M} : n = 3$	0.055 ± 0.007	0.056 ± 0.007	0.046 ± 0.004	0.049 ± 0.005
$\mathcal{M} : n = 10$	0.015 ± 0.001	0.016 ± 0.001	0.015 ± 0.001	0.015 ± 0.001
$\mathcal{M} : n = 40$	0.006 ± 0.001	0.006 ± 0.001	0.004 ± 0.001	0.004 ± 0.001
$\mathcal{A} : n = 3$	0.044 ± 0.004	0.048 ± 0.004	0.016 ± 0.015	0.043 ± 0.016
$\mathcal{A} : n = 10$	0.017 ± 0.001	0.016 ± 0.001	N.A.	N.A.
$\mathcal{L} : n = 3$	0.066 ± 0.015	0.066 ± 0.020	0.053 ± 0.012	0.058 ± 0.012
$\mathcal{L} : n = 10$	0.029 ± 0.011	0.034 ± 0.007	0.040 ± 0.002	0.040 ± 0.002
$\mathcal{R} : n = 3$	0.054 ± 0.005	0.053 ± 0.005	0.041 ± 0.007	0.043 ± 0.014
$\mathcal{R} : n = 10$	0.015 ± 0.002	0.007 ± 0.001	0.019 ± 0.001	0.020 ± 0.001
$\mathcal{G} : n = 3$	0.065 ± 0.009	0.064 ± 0.013	0.061 ± 0.014	0.050 ± 0.017
$\mathcal{G} : n = 10$	0.808 ± 5.670	1.080 ± 10.380	0.748 ± 6.995	2.023 ± 18.077
$\mathcal{S} : n = 5$	0.006 ± 0.008	0.006 ± 0.004	0.030 ± 0.012	0.021 ± 0.004
$\mathcal{S} : n = 10$	0.002 ± 0.001	0.002 ± 0.001	0.009 ± 0.010	0.005 ± 0.003
$\mathcal{V} : n = 3$	0.063 ± 0.008	0.065 ± 0.010	0.015 ± 0.005	0.040 ± 0.010
$\mathcal{V} : n = 10$	0.027 ± 0.002	0.020 ± 0.003	0.025 ± 0.001	0.025 ± 0.001
$\mathcal{R}_{SR} : n = 3$	0.055 ± 0.006	0.056 ± 0.009	0.037 ± 0.010	0.045 ± 0.012
$\mathcal{R}_{SR} : n = 10$	0.021 ± 0.002	0.021 ± 0.002	0.018 ± 0.001	0.018 ± 0.001
$\mathcal{G}_{SR} : n = 3$	0.176 ± 0.150	0.156 ± 0.050	0.152 ± 0.069	0.181 ± 0.206
$\mathcal{G}_{SR} : n = 10$	0.031 ± 0.011	0.031 ± 0.016	0.032 ± 0.013	0.034 ± 0.011

niching acceleration for the Mahalanobis-distance based routines. This result is pretty much intuitive - a more accurate spatial classification, as typically obtained by the Mahalanobis metric, allows the niching mechanism in most cases to form appropriate niches and to converge faster.

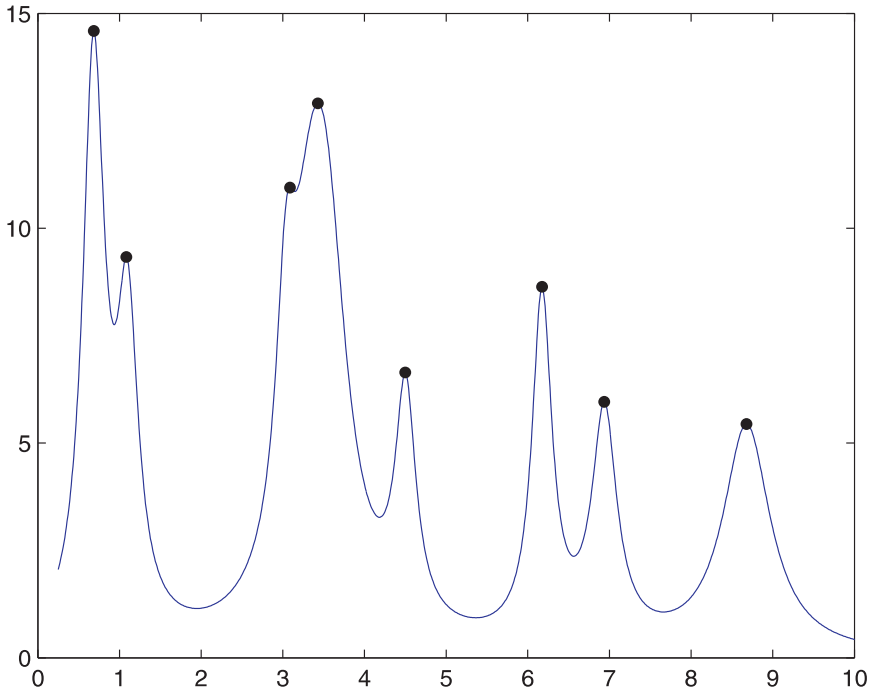

Figure 4.5: Final population of the CMA-$(1, 10)$ with a self-adaptive niche radius on the $1D$ Shekel function.

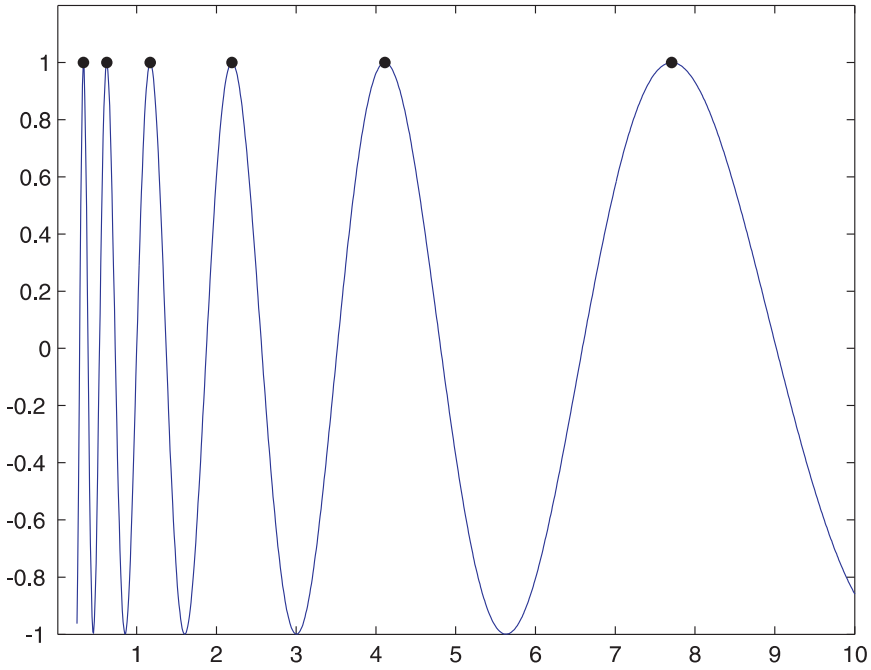

Figure 4.6: Final population of the CMA-$(1, 10)$ with a self-adaptive niche radius on the $1D$ Vincent function.

4.3.2 General Behavior

The proposed self-adaptive niche radius routine performed well on the landscapes with the "deceptive" distribution of optima, i.e., $\mathcal{V}$ and $\mathcal{S}$, and managed to tackle the niche-radius problem successfully. Visualizations of the runs on $\mathcal{V}$ and $\mathcal{S}$ for $n = 1$ are given as Figures 4.5 and 4.6. Figures A.1, A.2, and A.3 illustrate the adaptation of the classification-ellipses by the M-CMA+ routine on the $2D$ Fletcher-Powell, $3D$ Fletcher-Powell, and $3D$ Ackley landscapes, respectively. It can be observed in the Fletcher-Powell case that each niche has its own characteristic matrix and convergence profile, whereas the convergence in the Ackley seems to be simultaneous, as expected from the landscape symmetry.

4.4 Discussion

We have introduced new concepts of adaptive niche-radii and niche-shapes into the framework of niching with the Covariance Matrix Adaptation Evolution Strategy. The main goal was to treat the so-called niche radius problem, and to offer an efficient niching mechanism with no pre-assumptions on the landscape. It was successfully achieved at two levels: The construction of self-adaptive niche-radius, and the employment of the Mahalanobis distance for the adaptation of the niche-shapes. We have described both approaches in detail.

In further detail, given the CMA-ES-$\left(1 \dot{+} \lambda\right)$ routines, 4 variants of niching were considered per routine, and tested on a suite of artificial landscapes. The new approaches were shown to perform in a satisfying manner, on landscapes with evenly and unevenly spread optima. The niche radius problem was tackled successfully by the self-adaptive approach, as demonstrated on landscapes with unevenly spread optima, both separable and non-separable. The application of the Mahalanobis distance achieved its goal in improving the niching process, in terms of obtaining on average higher quality suboptima, subject to higher *niching acceleration*. It does neither seem to improve nor to hamper, on average, the identification of the location of global minimum, as expected.

The careful reader should note that employing the Mahalanobis distance is applicable only when the niching distance is calculated in the decision space. Sometimes this is not the case, and other spaces are used for that (e.g., the *second-derivative space*, for more details see Chapter 8).

We would like to discuss here the important issue of parameters in light of our proposed approaches. The discussion is done at two levels. The first is the relaxation of existing parameters in the fixed-radius CMA niching algorithm, and more specifically the parameter q. The parameter q is reduced in this study, for the first time, from being a critical niching parameter in the fixed-radius approach into being the estimated/desired target number of

niches/peaks in the self-adaptive approaches without any influence on the algorithmic behavior. In essence, a possibly wrong estimation of q would simply be responsible for wasting CPU cycles when too large, or missing good optima when too small. The second level is the introduction of new parameters, i.e., α and γ (Eq. 4.3), for the function of the *learning coefficients* (Eq. 4.2). Although this is an undesired situation, one should keep in mind that by setting only two parameters, we are allowing the application of a niching method to landscapes with a large number of optima with possibly different basin sizes, that would require different niche radii, respectively. We would like to stress that if these parameters had not been introduced, **the application to such landscapes would not have been feasible with the fixed-radius approach, or would have required setting as many parameters as the number of peaks**. Thus, by setting only these two parameters, we achieve a lot. Moreover, the proposed settings apply for a wide range of practically relevant landscapes, and do not have to be chosen for each new problem by means of additional experiments.

Regarding the implementation of the Mahalanobis metric, we have offered here a numerical simplification of the required calculation, which was observed to pay off in terms of computation time. By applying this numerical implementation, the Mahalanobis approach share the same computational complexity as the previously discussed approaches.

We thus present here both the self-adaptive niche-radius CMA-niching as well as the CMA-niching with Mahalanobis distance as state-of-the-art niching techniques within Evolution Strategies, and propose them as solutions to the so-called *niche-radius problem*.

People talk about the middle of the road as though it were unacceptable. Actually, all human problems, excepting morals, come into the gray areas. Things are not all black and white. There have to be compromises. The middle of the road is all of the usable surface.

Dwight D. Eisenhower

Chapter 5

Niching-CMA for Multi-Objective Optimization

This chapter introduces an additional extension to our proposed niching framework of Chapter 3, aiming at constructing a simple algorithm for multi-objective optimization.

5.1 Multi-Objective Optimization

Decision making in real-life is often subject to multiple objectives to be met. In many scenarios, satisfying one objective is typically in conflict with satisfying the other. The field of Multi-Criterion Decision Making (MCDM) aims at developing mechanisms for supporting the decision making process when treating multiple objectives. The idea is to study the nature of the trade-off between the various objectives, to seek a good compromise, and to avoid a lose-lose scenario.

Naturally, we are interested in the optimization perspective of MCDM, and especially in *evolutionary multi-objective optimization* algorithms (EMOA). The latter has developed in the last two decades, and has become a field of intense research.

Next, we briefly review here formally the basic concepts of Multi-Objective Optimization.

5.1.1 Formulation

Given an optimization problem with m objectives, we consider its m-dimensional objective space, also referred to as the *solution space*. By definition, the vector of objectives is in $\mathbb{R}^m$:

$$\vec{f}(\vec{x}) = \left(f_1(\vec{x}), f_2(\vec{x}), \ldots, f_m(\vec{x})\right)^T \tag{5.1}$$

We assume that all objectives are to be minimized. A partial order is defined on the *solution space*, $\mathcal{F} = \vec{f}(\mathcal{X})$, by means of the Pareto domination concept for vectors in $\mathbb{R}^m$, in the following manner:

Definition 5.1.1. Given any $\vec{f}^{(1)} \in \mathbb{R}^m$ and $\vec{f}^{(2)} \in \mathbb{R}^m$, we state that $\vec{f}^{(1)}$ strictly **Pareto** dominates $\vec{f}^{(2)}$, noted as

$$\vec{f}^{(1)} \prec \vec{f}^{(2)},$$

if and only if the following holds:

$$\forall i \in \{1, \dots m\} : f_i^{(1)} \leq f_i^{(2)} \ \wedge \ \exists i \in \{1, \dots, m\} : f_i^{(1)} < f_i^{(2)} \tag{5.2}$$

Note, that in the bi-criteria case this definition is reduced to:

$$\vec{f}^{(1)} \prec \vec{f}^{(2)} :\Leftrightarrow f_1^{(1)} < f_1^{(2)} \wedge f_2^{(1)} \leq f_2^{(2)} \vee f_1^{(1)} \leq f_1^{(2)} \wedge f_2^{(1)} < f_2^{(2)} \tag{5.3}$$

In addition to the strict domination $\prec$, we define further comparison operators:

$$\vec{f}^{(1)} \preceq \vec{f}^{(2)} \Longleftrightarrow \vec{f}^{(1)} \prec \vec{f}^{(2)} \vee \vec{f}^{(1)} = \vec{f}^{(2)} \tag{5.4}$$

Moreover, we state that $\vec{f}^{(1)}$ is incomparable to $\vec{f}^{(2)}$, noted as

$$\vec{f}^{(1)} || \vec{f}^{(2)},$$

if and only if

$$\vec{f}^{(1)} \not\preceq \vec{f}^{(2)} \wedge \vec{f}^{(2)} \not\preceq \vec{f}^{(1)} \tag{5.5}$$

The crucial claim is that **for any compact subset of $\mathbb{R}^m$, say $\mathcal{F}$, there exists a non-empty set of minimal elements with respect to the partial order $\preceq$** (see, e.g., [97], pp. 29).

We can now define **non-dominated points** as follows:

Definition 5.1.2. Non-dominated points are the set of minimal elements with respect to the partial order $\preceq$:

$$\mathcal{F}_N = \{\vec{f} \in \mathcal{F} | \nexists \vec{f'} \in \mathcal{F} : \vec{f'} \prec \vec{f}\} \tag{5.6}$$

where a subscript N will denote from now on a non-dominated set in the context of multi-objective optimization.

Having defined the non-dominated set and the concept of Pareto domination for general sets of vectors in $\mathbb{R}^m$, we are now in a position to relate it to the optimization mission. The aim of Pareto optimization is to obtain the *non-dominated set* for $\mathcal{F} = \vec{f}(\mathcal{X})$ and its pre-image in $\mathcal{X}$, the so-called *Pareto optimal set*, also referred to as the *efficient set*. We may then define the **Pareto front** as the set of all points in the objective space that correspond to the solutions in the Pareto-optimal set.

In many practical applications we are also satisfied with a set of solutions whose image under $\vec{f}$ yields a good approximation to the non-dominated set, though a definition of what is a good approximation is problem dependent. Often, it is desired to achieve a uniform distribution on the Pareto front and a good convergence of all points in the approximation set to some non-dominated solution.

For notational convenience, we shall define a strict pre-order on the *decision space* as follows:

$$\vec{x}^{(1)} \prec \vec{x}^{(2)} \Longleftrightarrow f(\vec{x}^{(1)}) \prec f(\vec{x}^{(2)}) \tag{5.7}$$

Accordingly, we define the pre-order

$$\vec{x}^{(1)} \preceq \vec{x}^{(2)} \Longleftrightarrow f(\vec{x}^{(1)}) \preceq f(\vec{x}^{2}) \tag{5.8}$$

Note, that this is not a partial order, as the *antisymmetry axiom* does not have to be satisfied. This stems from the fact, that two distinct vectors may have the same function value. For the same reasons, it is also possible that the efficient set comprises more members than the Pareto front.

5.1.2 The NSGA-II Algorithm

Due to their robustness and flexibility, Evolutionary Multi-Objective Optimization Algorithms (EMOA) have recently received increased attention as problem solvers for difficult simulator-based optimization problems [98, 99, 100]. Among these methods, the NSGA-II method is one of the most popular, and it has been successfully applied to many real-world problems.

The NSGA-II algorithm has been proposed by Deb [99]. It aims at obtaining a well distributed approximation set of points that are close to the Pareto front. It is a $(\mu + \lambda)$-EA (see Algorithm 1), which employs specific variation operators (for details we refer the reader to [99]), as well as a unique selection operator. We choose to describe the latter in detail.

The NSGA-II selection consists of two phases, that correspond to primary versus secondary selection criteria. At first, a procedure called *non-dominated sorting* is applied, that obtains perfect order on the set of decision vectors. Next, the solutions which share the same rank are sorted by means of the *crowding distance criterion*. Explicitly, non-dominated sorting works as follows: Given a population R, its non-dominated subset $R_1 = R_N$ is extracted. This set forms the best ranked solutions (rank=1). Given the set $R - R_N$, the non-dominated subset $R_2 = (R - R_N)_N$ is then extracted, and so on. This is repeated until the set of solutions is empty. The sets $R_1, \ldots, R_i, \ldots, R_\ell$ are called the non-dominated sets *of rank* i, $i = 1, \ldots, \ell$. Since these sets can possibly contain more than one member, a second criterion is applied in order to sort solutions that share the same rank. This secondary criterion puts emphasis on the diversity of the solutions, and is

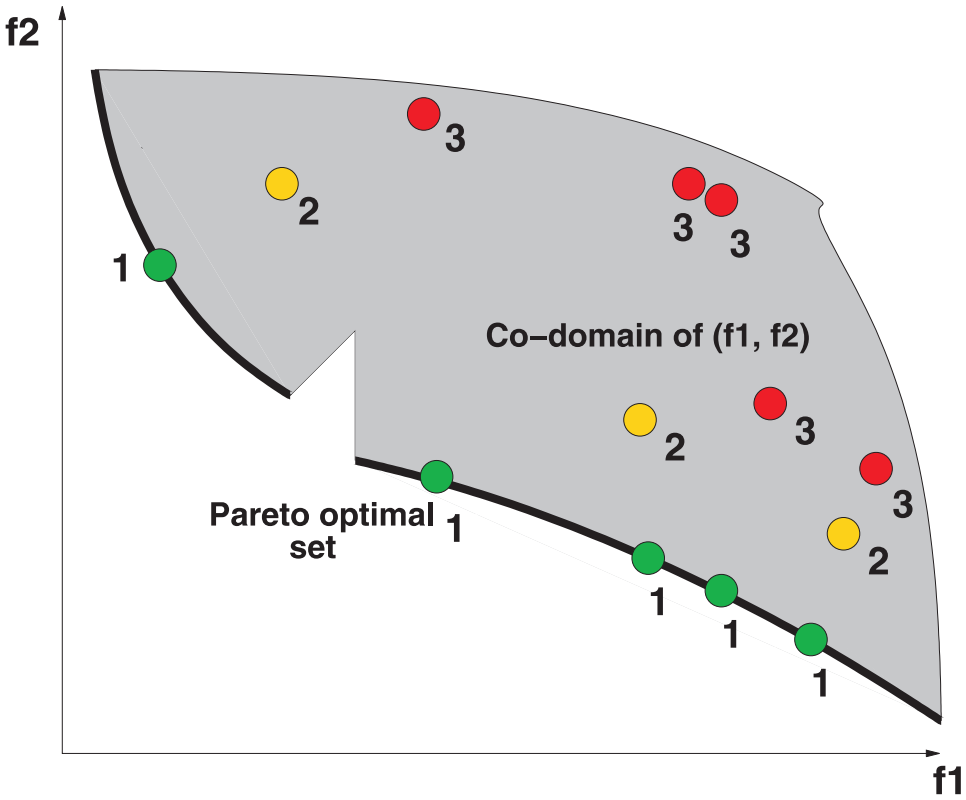

Figure 5.1: Non-dominated sorting. Figure courtesy of Michael Emmerich [101].

called the *crowding distance*: Given a solution $\vec{x}^{(i)} \in \mathbb{R}^n$, we determine the corresponding $\vec{f} = f(\vec{x})$ in the solution space, and then evaluate

$$d(\vec{f}) = \sum_{k=1}^{n} \begin{array}{l} \Big[\min_{\{f_k^{(j)} | j \in \{1,\ldots,|R|\} - \{i\} \wedge f^{(k)} \leq f^{(i)}\}} f_k^{(i)} - f_k^{(j)} + \\ \min_{\{f_k^{(j)} | j \in \{1,\ldots,|R|\} - \{i\} \wedge f^{(k)} \geq f^{(i)}\}} f_k^{(j)} - f_k^{(i)} \Big] \end{array} \tag{5.9}$$

For a visualization of the non-dominated sorting procedure and the crowding distance calculation on a bi-criteria optimization problem we refer to Figures 5.1 and 5.2, respectively.

A comprehensive overview on the NSGA-II and other EMO algorithms can be found in [99]. Recently, an interesting method called the SMS-EMOA [100] was proposed, and was shown to outperform the NSGA-II algorithm on standard benchmarks. However, the NSGA-II can be considered still as the most widely applied EMOA technique in literature, and thus we shall employ it in this study (see Chapter 9).

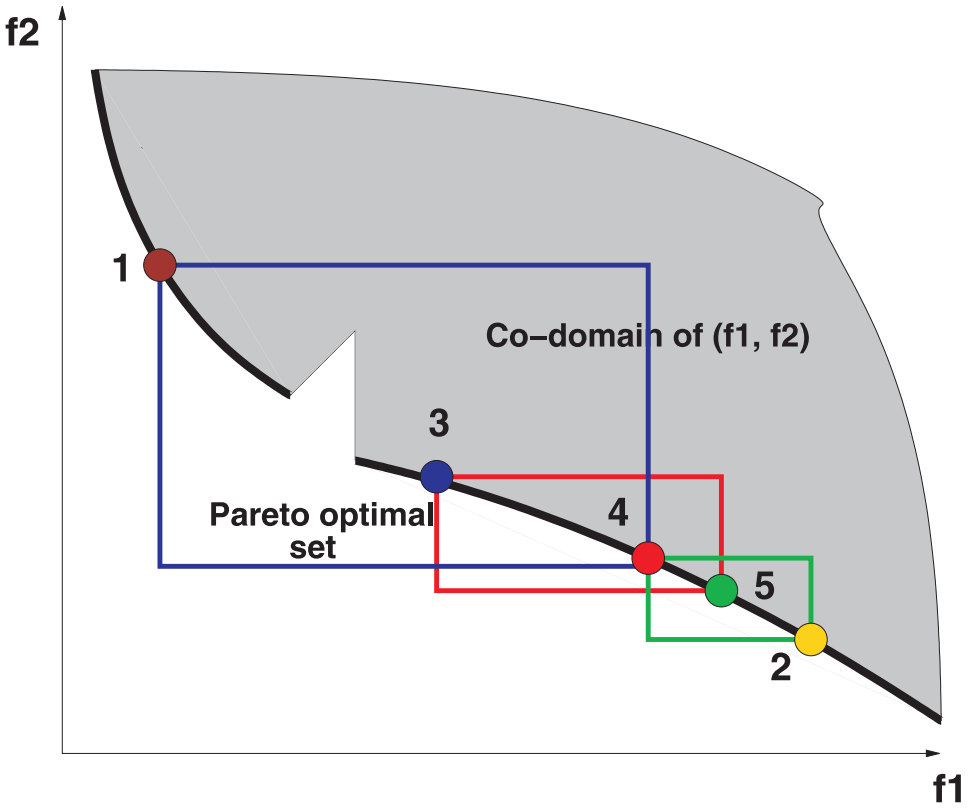

Figure 5.2: Crowding distance. Figure courtesy of Michael Emmerich [101].

5.2 On Diversity in Multi-Objective Optimization

Recently it has been pointed out that not only high diversity of solutions in the objective space but also high diversity of solutions in the efficient set can be of interest for the decision maker [68, 102]. For instance, if a specific point on the Pareto front is selected by the decision maker, it might also be interesting to consider different possible realizations to this solution in the decision space. Hence, if there are two different pre-images of the selected point on the Pareto front in the efficient set, both of them are of potential interest for the decision maker. This situation is illustrated in Figure 5.3.

More precisely, the difference between the classical selection principle to our proposed approach can be formalized as follows. Let $\mathcal{A}$ denote an approximation set on which we would like to apply ranking, and let $\vec{x}_A$ and $\vec{x}_B$ be two solutions in $\mathcal{A}$. In the classical selection method, as employed by the NSGA-II or SMS-EMOA algorithms, a solution $\vec{x}_A$ is preferred to a solution $\vec{x}_B$ if $\vec{x}_A$ has a better dominance rank than $\vec{x}_B$ in $\mathcal{A}$, with respect to non-dominated sorting. Given that $\vec{x}_A$ and $\vec{x}_B$ share the same dominance rank in $\mathcal{A}$, then $\vec{x}_A$ is preferred to $\vec{x}_B$, if and only if $\vec{x}_A$ contributes more

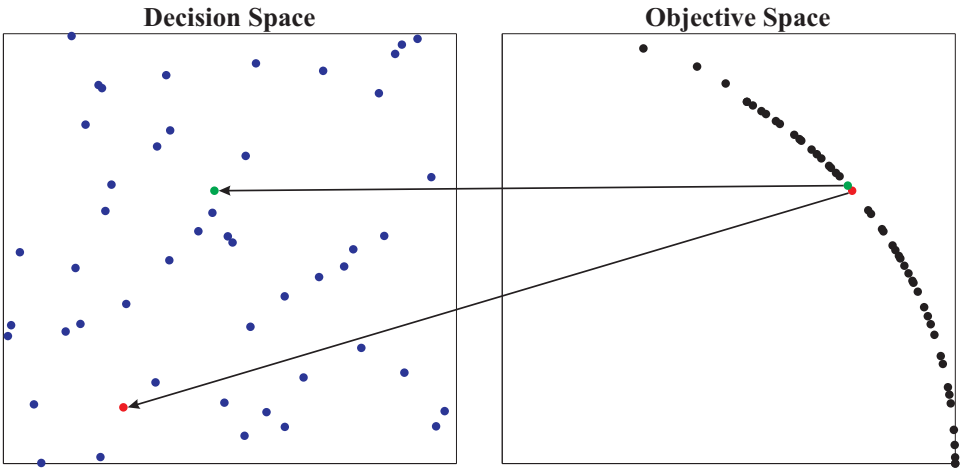

Figure 5.3: Diversity for *decision making*: Illustrative example for a scenario where two adjacent points on the Pareto front are mapped onto two points in two completely different regions in the decision space. Units and scales are arbitrary.

to the diversity of the approximation set in the objective space than $\vec{x}_B$. In the proposed selection principle, $\vec{x}_A$ remains preferable to $\vec{x}_B$, if $\vec{x}_A$ has a better dominance rank than $\vec{x}_B$ in $\mathcal{A}$. However, given that $\vec{x}_A$ and $\vec{x}_B$ share the same dominance rank in $\mathcal{A}$, then $\vec{x}_A$ is preferred to $\vec{x}_B$, if and only if it contributes more to the diversity in the **aggregated space** (i.e., in both objective and decision spaces). This principle can be instantiated in different ways, depending on the diversity measure defined on the aggregated space.

Multi-objective optimization methods aim at maintaining diversity, by their definition, and indeed, one of the popular mechanisms for diversity maintenance is the crowding concept [67], which is also applied, yet differently, as a single-objective niching technique. Thus, the important component of *diversity* is the linking element between the fields of multi-objective and multi-modal optimization. However, in multi-objective optimization the diversity maintenance is typically sought in the objective space, for the sake of obtaining a fair coverage of the Pareto front, while not taken into account for the Pareto optimal set in the decision space.

5.2.1 Related Work

Several different studies treated related topics to the work presented in this chapter. We review them here shortly.

Niching for MOEA: The NPGA Niching techniques have been already used in the multi-objective optimization arena, by being adjusted accordingly. Horn, Nafploitis and Goldberg [103] introduced a niching technique for

multi-objective optimization, known as the Niched-Pareto GA (NPGA). The algorithm was a variant of the *fitness sharing* niching method, whereas the *niching distance metric was set to consider the objective space only.* The selection was based on the so-called *Pareto domination tournaments* or on the minimal niche count, otherwise. The NPGA was a classical example of using an existing single-objective niching technique, in a straightforward manner, for multi-objective optimization - only by redefining the niching distance measure and the selection mechanism. However, its kernel was the simple GA, which typically suffers from limited performance in high-dimensional continuous landscapes, and it lacked any self-adaptation mechanism.

The Omni-Optimizer Deb's so-called Omni-Optimizer [68] is considered to be one of the first and only attempts of introducing a generic optimization routine which aims at covering the four categories of function optimization: Single-objective uni-global, single-objective multi-global, multi-objective uni-global, and multi-objective multi-global problems. Also, it is one of the first attempts to take diversity in the decision space into consideration.

In principle, this algorithm extends the NSGA-II by considering additionally the diversity in the decision space. This is implemented by means of the crowding distance calculation in the decision space for all the individuals. The assigned crowding distance is defined as follows:

$$
\begin{array}{ll}
\textbf{if} & \text{crowd\_dist\_obj}(i) > \text{avg\_crowd\_dist\_obj} \textbf{ or} \\
 & \text{crowd\_dist\_dec}(i) > \text{avg\_crowd\_dist\_dec} \\
\textbf{then} & \text{crowd\_dist}(i) = \max\left(\text{crowd\_dist\_obj}(i), \text{crowd\_dist\_dec}(i)\right) \\
\textbf{else} & \text{crowd\_dist}(i) = \min\left(\text{crowd\_dist\_obj}(i), \text{crowd\_dist\_dec}(i)\right)
\end{array}
$$

i.e., if the individual has above-the-average crowding distance, either in the decision or objective space, the larger of them is assigned to it, otherwise the smaller of the two distances is assigned. This criterion is rather general, and strongly relies on uniform distribution of peaks as well as on their equal fitness values. Also, the scalability of the two different spaces is not treated. We would like to speculate that it is expected to experience difficulties on non-uniform multi-modal landscapes, for instance. From the practical perspective, the algorithm was reported in [68] to be tested only on a single test function, constructed by Deb for this purpose, with uniformly-distributed equi-fitness minima landscape. We shall revisit this test-function in our experimental procedure.

Decision-Space Diversity as an Independent Objective Toffolo and Benini [104] also promoted the issue of genetic diversity in multi-objective algorithms, and proposed their so-called Genetic Diversity Evolutionary Algorithm (GDEA) for multi-objective optimization. The latter considers the

diversity of trial solutions in the decision space, quantified by means of a coverage function, as an independent objective, subject to maximization, in the ongoing multi-objective search. This GA-based approach was shown to outperform the NSGA on a set of $30D$ bi-criteria minimization problems introduced by Zitzler et al. [105].

Self-Adaptation in Multi-Objective Optimization Self-adaptation of strategy parameters [106] has become a fundamental component in the evolutionary optimization routine. Moreover, the self-adaptation of the mutation strategy parameters has been shown to be necessary for efficient single-objective optimization within ES [106].
Self-adaptation is expected to fail in the classical multi-objective optimization routine. This is due to the fact that given conflicting objectives, a successful mutation toward one objective is not necessarily a successful mutation toward the others – and hence should not be selected.
Büche, Müller and Koumoutsakos [107] conducted a pioneering study of self-adaptation in multi-objective optimization. They considered three different classes of multi-objective algorithms - *independent sampling*, *cooperative population search with dominance criterion* and *cooperative population search without dominance criterion*. Three representatives - CMEA, SPEA and SDM - matching the classes respectively, were tested on a multi-objective generalization of the *sphere model*, and compared with respect to each other. Self-adaptation had been plugged-in into the evolutionary core mechanisms of the algorithms, in a limited way (rotation angles, for instance, were not always adapted). The conclusion was that self-adaptation did not work for cooperative population searches which use the dominance criterion in the fitness assignment (SPEA), and this result was reassured by testing more representatives from that class of algorithms, such as the NSGA-II and SPEA2. However, self-adaptation could work for the CMEA and SDM, which do not use dominance, but rather consider a single objective for optimization while the other objectives are treated as constraints. The concluding message was clear – self-adaptation does not work in its classical definition upon considering multiple objectives – as had been speculated.

Recently, the self-adaptation obstacle was treated successfully by using the so-called *hyper-volume indicator* (also known as S-metric) [98] as a selection criterion, similar to [100], in the Multi-Objective CMA-ES [33], to be discussed next. A similar approach, yet employing a simpler ES kernel, was also reported recently in [108].

CMA-ES for Multi-Objective Optimization An algorithm for multi-objective optimization with a CMA kernel was introduced recently [33], employing numerous $(1 + 1)$ parallel search processes that undergo a shared selection phase. The latter is based on non-dominating ranking as a primary

criterion, followed by the maximization of the Pareto front hyper-volume as a secondary criterion. Crowding distance was also considered as an alternative secondary selection criterion. In many ways, this algorithm resembles our niching framework. However, its diversity preservation stems from the outcome of selection with respect to multiple criteria, rather than from the spatial enforcement of speciation by means of a niche definition. It is important to note in this context, that the hyper-volume indicator is well-defined as a measure of diversity and solution-set quality in the objective space, but cannot be applied as an indicator of diversity in the search space.

5.3 Multi-Parent Niching with (μ_W, λ)-CMA

In order to apply a niching algorithm for multi-objective optimization, we would like to design a stable niching kernel, where niches are less dynamic and associated more strongly with their spatial origins. In practice, we aim at fixing an offspring to its spatial niche, or alternatively, at verifying that a selected successor of a niche indeed originates from the same source as the parent as well as the other members. The verification of this condition may be easily incorporated into the niching framework presented in Chapter 3. This condition naturally poses a limitation on the free speciation process. Thus, we would like to boost the performance of this limited niching variant by introducing a multi-parent niching approach, as will be discussed shortly.

The $\left(1 \stackrel{+}{,} \lambda\right)$ niching framework may be extended to a multi-parent niching framework, by employing a (μ_W, λ)-CMA kernel. We propose here the following algorithm. In this extension, the issue to be treated is the identification of the selected set of offspring due to be recombined. Following the $(1, \lambda)$ framework, the niche representative is well defined, i.e., as output from the DPI routine. However, the number of individuals in that niche is unknown a-priori, and moreover, some of the individuals in the current spatial niche might not share the same parent. Thus, we choose to define the rest of the selected offspring as the set of at most $\lfloor \frac{\lambda}{2} \rfloor - 1$ individuals that are within niche radius from the peak individual and share a parent with it. This way, it is guaranteed that the ES mutation distribution evolves continuously, and that the spatial niche is stable.

Since the value of μ is set dynamically every generation, and is likely to vary over time, other auxiliary coefficients must be updated accordingly, such as the recombination weights (see Eq. 1.44). Otherwise, this scheme is not expected to introduce any instabilities into the niching framework. As for the value of λ, we propose to set it to its recommended default value, as in Eq. 1.47:

$$\lambda = 4 + \lfloor 3 \cdot \ln(n) \rfloor$$

A pseudo-code for the *multi-parent-CMA niching routine* is presented in Algorithm 8.

Algorithm 8 Multi-Parent (μ_W, λ) Niching-CMA with a Fixed Niche Radius

for $i = 1 \dots (q + p)$ search points **do**
  Generate λ samples based on the CMA-set of i
end for
Evaluate fitness of the population
Compute the Dynamic Peak Set with the DPI Algorithm
for $j = 1 \dots q$ elements of DPS **do**
  Identify at most $\mu = \lfloor \frac{\lambda}{2} \rfloor$ fittest individuals with $Parent\,(peak(j))$
  Apply weighted recombination on these individuals to yield $\langle \vec{x} \rangle_W^j$, $\langle \vec{z} \rangle_W^j$
  Inherit the CMA-set of $Parent\,(peak(j))$ and update it w.r.t. $\langle \vec{z} \rangle_W^j$
end for
if N_{DPS}=size of $DPS < q$ **then**
  Generate $q - N_{DPS}$ new search points, reset CMA-sets
end if
if $gen \mod \kappa \equiv 0$ **then**
  Resample the $(q+1)^{th} \dots (q+p)^{th}$ search points
end if

Numerical Observation: $(1, \lambda)$-Niching vs. (μ_W, λ)-Niching

We tested the derived multi-parent niching-CMA variant on the suite of artificial multimodal landscapes of Section 3.4. A comparison with its $(1, \lambda)$ sibling clearly shows that the multi-parent variant is inferior in performance on the given landscapes. It seems that the free speciation component in the original $(1, \lambda)$ strategy plays an important role in the niching process. Therefore, we restrict the use of the multi-parent variant to the multi-objective framework, which will be derived next.

5.4 Niching-CMA as EMOA

The idea of the proposed method is to approximate the Pareto front using *niches*, i.e. every niche represents a point in the evolving front. This is achieved by considering the aggregated decision and objective spaces for the distance metric of the niching formation. This method employs the multi-parent niching-CMA routine as it is, with the following modifications:

- Ranking of individuals is based upon non-dominated sorting.
- Distance between niches is evaluated in the aggregated space, as will be explained shortly. Also, the estimation of the niche radius is adjusted.

5.4.1 The Niching Distance Metric

Given the n-dimensional decision vector of individual i, $\vec{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_n^{(i)}\right)^T$, with its assigned m-dimensional objective vector, $\vec{f}^{(i)} = \left(f_1^{(i)}, \ldots, f_m^{(i)}\right)^T$, and given the equivalent decision and objective vectors of individual j, $\left(\vec{x}^{(j)},\ \vec{f}^{(j)}\right)$, the distance between individuals i and j is defined as the Euclidean distance between the two aggregated vectors subject to dimensionality normalization, i.e., norm-2 in the $n+m$ aggregated space. It explicitly reads,

$$d_{i,j} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(x_k^{(i)} - x_k^{(j)}\right)^2 \ + \ \frac{1}{m}\sum_{\ell=1}^{m}\left(f_\ell^{(i)} - f_\ell^{(j)}\right)^2} \tag{5.10}$$

5.4.2 Selection: Non-dominating Ranking

In order to select individuals based on more than a single objective, the existing selection mechanism had to be modified. As outlined previously, the niches are identified based on their ranked quality. In our new multi-objective context, rather than sorting the fitness values, we propose to perform *dominance ranking*, after which the routine will proceed as usual: Starting with rank 1, a greedy identification of the niches will be executed, considering the distance with respect to the aggregated objective and decision spaces. If not all q niches are populated, the routine will proceed to rank 2, and so on.

5.4.3 Estimation of the Niche Radius

Since our method aims to approximate the Pareto front by populating it with a uniform distribution of q niches, we can estimate the niche radius ρ for specific cases. The following derivations are strictly limited to $2D$ decision or objective spaces, but we believe that they could be generalized to n-dimensional spaces.

Consider a connected Pareto front, and assume that we can define its *length*, denoted by l_{FRONT}. Also, let the diameter of the Pareto set be denoted by l_{SET}. Upon considering the aggregated space, and demanding a uniform distribution of niches, one may write:

$$2 \cdot \rho \cdot q = \sqrt{l_{FRONT}^2 + l_{SET}^2} \tag{5.11}$$

Simplified Model One can consider a simplified model for providing an upper and a lower bounds for ρ, by taking into account only the objective space. For this purpose let us consider the *Nadir* objective vector, denoted here as $\vec{\zeta}^{(\mathcal{N})} = (f_{1,\mathcal{N}}, f_{2,\mathcal{N}})^T$. In the general m-dimensional objective space, the *Nadir* objective vector is defined as the vector with the *worst objective*

values of all Pareto optimal solutions (as opposed to the worst objective values of the entire space):

$$\zeta_i^{(\mathcal{N})} = \max \left\{ f_i \,\middle|\, (f_1, \ldots, f_i, \ldots, f_m)^T \in \mathcal{F}_N \right\}. \tag{5.12}$$

The Nadir objective vector can be computed for $m = 2$ by employing single-objective optimization. For $m > 2$, heuristics are available, but the problem is considered to be computationally hard [97].

Without loss of generality, assume that the objectives $\{f_1,\ f_2\}$ are assigned with values in the intervals $\{[f_{1,min}, f_{1,\mathcal{N}}],\ [f_{2,min}, f_{2,\mathcal{N}}]\}$, respectively. The length of the assumably-connected Pareto front has a lower bound of

$$l_{FRONT,min} = \sqrt{\left((f_{1,\mathcal{N}} - f_{1,min})^2 + (f_{2,\mathcal{N}} - f_{2,min})^2 \right)}, \tag{5.13}$$

and an upper bound of

$$l_{FRONT,max} = |f_{1,\mathcal{N}} - f_{1,min}| + |f_{2,\mathcal{N}} - f_{2,min}|. \tag{5.14}$$

Hence, upon assuming a uniformly spaced population of the q niches along the front, one can derive

$$\frac{\sqrt{\left((f_{1,\mathcal{N}} - f_{1,min})^2 + (f_{2,\mathcal{N}} - f_{2,min})^2 \right)}}{2 \cdot q} \leq \rho \leq \frac{|f_{1,\mathcal{N}} - f_{1,min}| + |f_{2,\mathcal{N}} - f_{2,min}|}{2 \cdot q} \tag{5.15}$$

The General Case For the general case, we choose to define the default values as the radii of the decision or the objective spaces, respectively:

$$r_{SET} = \sqrt{\sum_{i=1}^{n} (x_{i,max} - x_{i,min})^2} \tag{5.16}$$

$$r_{FRONT} = \sqrt{\sum_{j=1}^{m} (f_{j,max} - f_{j,min})^2} \tag{5.17}$$

And thus

$$\rho = \frac{\sqrt{\sum_{i=1}^{n} (x_{i,max} - x_{i,min})^2 + \sum_{j=1}^{m} (f_{j,max} - f_{j,min})^2}}{2 \cdot q} \tag{5.18}$$

The niche radius is essentially a crucial parameter of this method, and its estimation or tuning is critical for the algorithmic success.

5.5 Numerical Simulations

We outline here our experimental setup for the proposed method.

5.5.1 Test Functions: Artificial Landscapes

We consider a set of artificial bi-criteria landscapes in order to test the algorithmic performance. Following our mission statement, and due to the fact that we have no desire in introducing another standard EMOA, we tend to focus in landscapes with more interesting decision space characteristics, and provide the reader with a *proof of concept* for the proposed approach. Next, we describe the four different landscapes to be considered:

1. **Deb's Omni-Test** As mentioned earlier, Deb constructed a bi-criteria multi-global landscape for testing his Omni-Optimizer [68]. Explicitly, it reads:

$$\begin{aligned} f_1(\vec{x}) &= \sum_{i=1}^{n} \sin(\pi x_i) \longrightarrow \min \\ f_2(\vec{x}) &= \sum_{i=1}^{n} \cos(\pi x_i) \longrightarrow \min \end{aligned} \tag{5.19}$$

   where $\forall i \; x_i \in [0, 6]$.

2. **EBN** The EBN family of functions [100] introduced a very basic set of test-problems for multi-objective algorithms. Explicitly, it reads:

$$\begin{aligned} f_1^{(\gamma)}(\vec{x}) &= \left(\sum_{i=1}^{n} |x_i| \right)^{\gamma} \cdot n^{-\gamma} \longrightarrow \min \\ f_2^{(\gamma)}(\vec{x}) &= \left(\sum_{i=1}^{n} |x_i - 1| \right)^{\gamma} \cdot n^{-\gamma} \longrightarrow \min \end{aligned} \tag{5.20}$$

   The shape of the Pareto front can be controlled by means of the parameter γ, and it is defined by the following equation:

$$y_2 = \left(1 - y_1^{1/\gamma}\right)^{\gamma}, \quad y_1 \in [0, 1] \tag{5.21}$$

   Thus, the shape of the front will be a concave, linear, or convex arc for the cases of $\gamma < 1$, $\gamma = 1$, or $\gamma > 1$, respectively.

   The main purpose of studies employing this set of problems is characterizing the EMOA distribution points on a Pareto front of different elementary shapes. The EBN problems are attractive in the context of efficient set approximation, as the pre-images of points in the objective space are not single points, but rather line segments on the diagonals of $[0, 1]^n$, excepting the extremal points $(0, 1)^T$ and $(1, 0)^T$ (see, e.g., [101]). In our study we shall consider the case of $\gamma = 1$.

3. **"Two-on-One"** This test-case was originally introduced in an interesting study of the Pareto-optimal set [109], which has been to some extent one of the origins to the study presented in this chapter. It is a two-dimensional function, with a 4^{th}-degree polynomial with two minima as f_1 versus the sphere function as f_2:

$$\begin{aligned} f_1(x_1, x_2) &= x_1^4 + x_2^4 - x_1^2 + x_2^2 - cx_1x_2 + dx_1 + 20 \longrightarrow \min \\ f_2(x_1, x_2) &= (x_1 - k)^2 + (x_2 - l)^2 \longrightarrow \min \end{aligned} \tag{5.22}$$

   We consider the asymmetric case, with $c = 10$, $d = 0.25$, $k = 0$, and $l = 0$ (case number 3 as reported in [109]).

4. **Lamé Superspheres** We consider a multi-global instantiation of a family of test problems introduced by Emmerich and Deutz [110], the Pareto fronts of which have a spherical or super-spherical geometry. In contrast to the EBN problem, the set of pre-images of a point on the Pareto front for this instance is finite, and solutions are placed on equidistant parallel line-segments, each of them being a pre-image of a local Pareto front.

   Let $d = \frac{1}{n-1}\sum_{i=2}^{n} x_i$, and $r = \sin^2(\pi \cdot d)$,

$$\begin{aligned} f_1 &= (1 + r) \cdot \cos(x_1) \longrightarrow \min \\ f_2 &= (1 + r) \cdot \sin(x_1) \longrightarrow \min \end{aligned} \tag{5.23}$$

   with $x_1 \in \left[0, \frac{\pi}{2}\right]$, and $x_i \in [1, 5]$ for $i = 2, \ldots, n$.

5.5.2 Modus Operandi

We carried out numerical simulations on the bi-criteria landscapes introduced in the previous section in order to test the algorithmic performance of the proposed method. We chose to apply three additional algorithms as reference methods: the NSGA-II, the Omni-Optimizer, and a variant of the NSGA-II which considers an aggregated space in the crowding calculations. The latter routine is meant to assess the importance of the aggregation concept for attaining decision space diversity. The idea was to approximate the Pareto front by means of $q = 50$ points, and allocate a fixed number of $NumEval_{\max} = 50,000$ function evaluations per run. We are aware that these are not the optimal settings for the reference methods; The Omni-Optimizer, for instance, was reported in [68] to employ a population of $1,000$ individuals. However, our goal here is also to exploit the advent of modern derandomized Evolution Strategies, which offer optimization with minimal settings.

In order to assess the boost of diversity in the decision space, we would like to introduce here a quantifier for that. Let $d_{A,B}$ denote the Euclidean

Table 5.1: Hypervolume values of the Pareto fronts of the 4 different algorithms on the 4 test-cases: Average and standard-deviation over 20 runs.

Hypervolume	Niching-CMA	NSGA-II	NSGA-II-Agg	Omni-Opt.
Omni-Test	30.27 ± 0.05	30.17 ± 0.034	29.80 ± 0.23	29.75 ± 0.18
EBN	3.283 ± 0.042	3.289 ± 0.088	2.87 ± 0.182	2.064 ± 0.057
Two-on-One	173.4 ± 0.26	173.7 ± 1.56	172.7 ± 1.78	150.2 ± 28.6
Superspheres	3.176 ± 0.038	3.203 ± 0.001	3.117 ± 0.080	2.457 ± 0.372

distance between individual $\vec{x}_A$ and individual $\vec{x}_B$:

$$d_{A,B} = \|\vec{x}_A - \vec{x}_B\| \tag{5.24}$$

We then **define** the population diversity of the Pareto optimal set as the mean value of the $\frac{\mu_N(\mu_N-1)}{2}$ distance measures between all the individuals, normalized by the diameter of the decision space, denoted by **diam**:

$$D = \frac{2}{\text{diam} \cdot \mu_N(\mu_N - 1)} \cdot \sum_{A \neq B} d_{A,B} \tag{5.25}$$

This scalar should give us an indication to what degree the *final* population is diverse.

5.5.3 Numerical Observation

We present the numerical results by means of plots of typical runs of the resulting approximated Pareto-set and Pareto-front (i.e., all the non-dominated individuals in the last generation). The plots present the outcome of the different algorithms both in the decision and the objective spaces, per landscape. Note that the decision space is represented by plotting x_1 versus x_2, except for the Superspheres test-case where x_1 is plotted versus $\frac{1}{(n-1)} \cdot \sum_{i=2}^{n} x_i$. These plots are given in Figures 5.4, 5.5, 5.6, and 5.7.

Table 5.1 presents the calculations of the S-metric, as a performance criterion in the objective space, averaged over 20 runs. Moreover, Table 5.2 presents the calculations of the decision space diversity, as defined in Eq. 5.25, averaged over 20 runs.

Generally speaking, the proposed algorithm performed in a highly satisfying manner, obtaining good Pareto-sets with high diversity in the decision space, which are mapped onto well-approximated Pareto-fronts. In terms of the performance criterion in the objective space, the S-metric (hypervolume), Niching-CMA and the NSGA-II performed equally well, while the NSGA-II with aggregation and the Omni-Optimizer typically performed slightly worse. Regarding the diversity in the decision space, the proposed algorithm accomplished its goal: it attained higher decision space diversity in comparison to

Table 5.2: Decision-space diversity, as defined in Eq. 5.25, of the 4 different algorithms on the 4 test-cases: Average and standard-deviation over 20 runs.

Diversity	Niching-CMA	NSGA-II	NSGA-II-Agg	Omni-Opt.
Omni-Test	0.256 ± 0.060	0.205 ± 0.079	0.222 ± 0.070	0.030 ± 0.002
EBN	0.483 ± 0.008	0.410 ± 0.023	0.356 ± 0.028	0.011 ± 0.010
Two-on-One	0.295 ± 0.01	0.136 ± 0.036	0.116 ± 0.031	0.106 ± 0.054
Superspheres	0.413 ± 0.024	0.239 ± 0.049	0.307 ± 0.046	0.062 ± 0.056

the other method on all landscapes. This result can also be clearly observed in the decision space plots. On the Omni-Test landscape, Niching-CMA performed very well, while typically obtaining 4 Pareto subsets, in comparison to one or two subsets for each of the other routines. On the EBN landscape, Niching-CMA attained a quasi-uniform distribution in the decision space. On the "Two-on-One" landscape, the proposed algorithm managed to explore both branches of the so-called *propeller-shaped Pareto-set* [109], while the other algorithms typically explored either one of the two branches. On the Super-Spheres landscape, Niching-CMA performed extremely well, while obtaining a good distribution of typically 3 Pareto subsets. The other methods, nevertheless, usually obtained a single Pareto subset. This is clearly observed in Figure 5.7, where the final population of the these algorithms is mostly concentrated along a single line-segment, corresponding to a single Pareto subset. Hence, in multi-globality terms, Niching-CMA clearly outperformed the other methods on these landscapes.

It should be noted that introducing the aggregation component into the NSGA-II did improve the attained decision space diversity to some extent on two landscapes, but did not have a considerable contribution. We conclude that considering the aggregated space by itself does not seem to be sufficient for attaining high diversity in the decision space. We rather consider it as a *bridge* for niching to multi-objective domains. We would like also to point out the poor performance of the Omni-Optimizer in terms of the attained decision space diversity. It is likely that its performance was hampered due to the small population size employed here.

Discussion

The constructed algorithm required rather mild adjustments to the new arena of multi-global multi-objective optimization. Due to the fact that it is niche-radius based, we proposed a way to approximate this parameter. The algorithm was applied to a testbed of conventional artificial bi-criteria landscapes, of various dimensions, and compared to the classical GA-based EMOAs: The NSGA-II, the Omni-Optimizer and an aggregated-space vari-

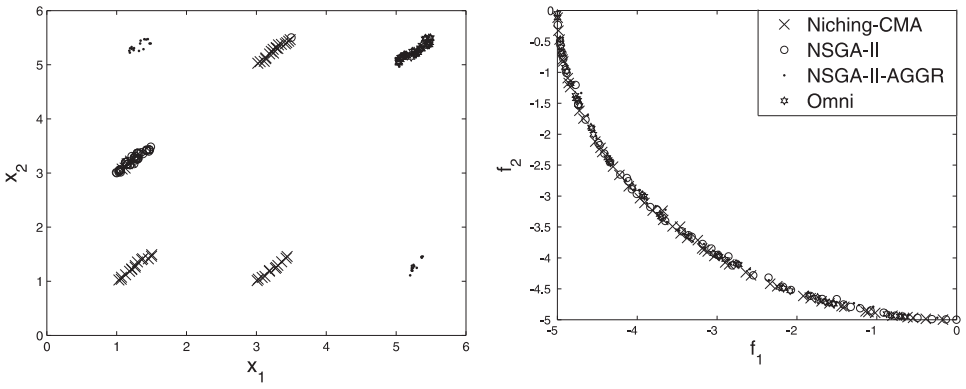

Figure 5.4: $5D$ Omni-Test landscape (Eq. 5.19): Final populations of the four routines (see legend). Left: Decision space; Right: Objective space.

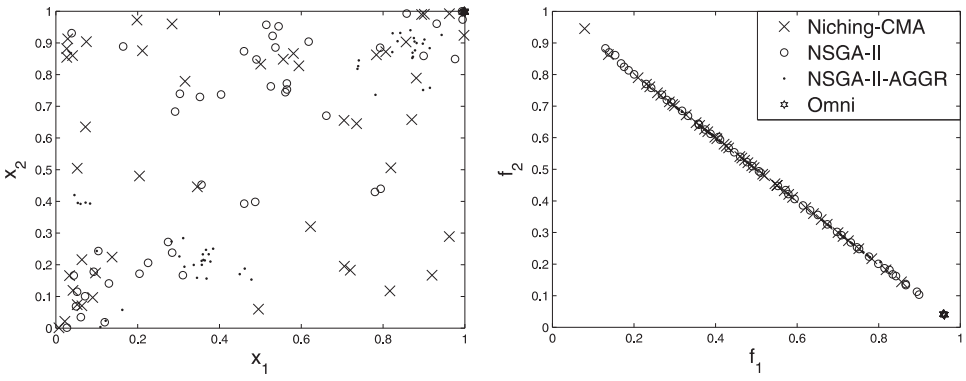

Figure 5.5: $10D$ EBN landscape (Eq. 5.20): Final populations of the four routines (see legend). Left: Decision space; Right: Objective space.

ant of the NSGA-II. The observed numerical results were highly satisfying, where in all cases not only the Pareto front, but also the efficient set, were better covered in comparison to the existing approaches. This outcome provided us with the desired *proof of concept* for the proposed method. It should be noted that the GA-based methods performed poorly, likely due to the small population sizes that are typically employed by ES-based algorithmic kernels.

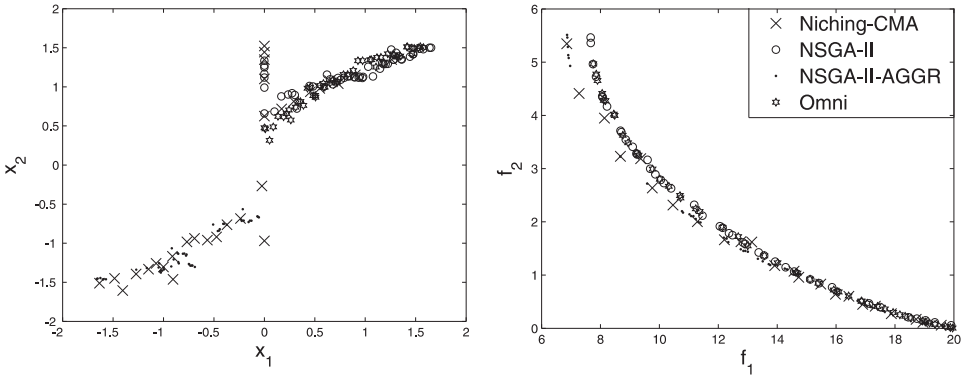

Figure 5.6: $2D$ Two-on-One landscape (Eq. 5.22): Final populations of the four routines (see legend). Left: Decision space; Right: Objective space.

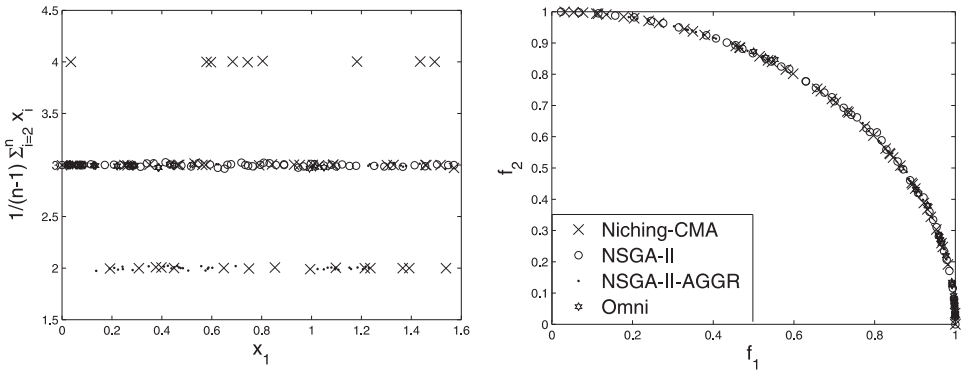

Figure 5.7: $4D$ Super-Spheres landscape (Eq. 5.23): Final populations of the four routines (see legend). Left: Decision space; Right: Objective space.

Part II

Quantum Control

The dream is alive!
Herschel Rabitz

Chapter 6

Introduction to Quantum Control

Controlling the motion of atoms and molecules has been a dream since the early days of Quantum Mechanics. Although this quest initially met with failure, the foundation of the Quantum Control (QC) field in the 1980s, throughout the development of various approaches [111, 112, 113], has finally brought this dream to fruition. Quantum Control, sometimes referred to as Optimal Control or Coherent Control, aims at altering the course of quantum dynamics phenomena for specific target realizations. There are two main threads within Quantum Control, *theoretical* versus *experimental* control, as typically encountered in Physics. They have experienced an amazing increase of interest during the past 10 years, in parallel to the technological developments of ultrafast laser pulse shaping capabilities, that obviously made it possible to turn the dream into reality. For a broad field review see [114, 115, 116].

The list of successfully closed-loop quantum controlled systems in Physics and Chemistry is practically endless. Examples of early work contain successful applications in *fluorescence spectrum manipulation* [117], *control of quantum wavefunctions* [118], *vibrational excitation tailoring in polymers* [119], *molecular rearrangement selectivity* [120], *chemical discrimination* [121], *ultrafast solid-state optical switching* [122], and *photosynthetic bacteria energy transfer* [123].

In this chapter we review the fundamental principles of Quantum Control, both in theory and in experiments. Should the reader choose to explore this chapter, an understanding of the basic quantum mechanics principles is assumed, as well as being familiar with the Dirac notation. The reader who wishes to abstract from the physics details could simply view the Quantum Control applications in this study as a non-linear high-dimensional set of problems with real-world applications.

6.1 Optimal Control Theory

Optimal Control Theory (OCT) [124, 125] aims at manipulating the quantum dynamics of a *simulated system* by means of an external control field, $\epsilon(t)$, which typically corresponds to a temporal electromagnetic field arising from a laser source. The objective to be met in this control process is defined by means of a given physical observable, whose yield is subject to maximization. A quantum control landscape is thus defined as the functional dependence of an observable yield on the control variables, and may be visualized as a surface over the space of all possible controls.

This section is mainly based on [126] (definitions) and on [127, 128] (QC derivations).

6.1.1 The Quantum Control Framework

Formally, we consider quantum systems which are described by Hamiltonians of the form

$$\mathcal{H}(t) = \mathcal{H}_0 - \vec{\mu} \cdot \vec{\epsilon}(t) \tag{6.1}$$

with $\mathcal{H}_0$ as the free-field Hamiltonian, $\vec{\mu}$ the dipole moment operator, and $\vec{\epsilon}(t)$ the electric field, within the so-called *electric dipole approximation*. The electric field is often reduced to a scalar, due to the common assumption of a linear polarization. In practice, a finite number N of states is considered, and thus **the Hilbert infinite-dimensional space is practically reduced to an N-dimensional space, and therefore the Hamiltonian is typically an $N \times N$ Hermitian matrix**.

Given some initial quantum state $|\psi(t=0)\rangle = |\psi_0\rangle$, the time evolution of the quantum state $|\psi(t)\rangle$ is dictated by the time-dependent Schrödinger equation:

$$i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle = \mathcal{H}(t) |\psi(t)\rangle \tag{6.2}$$

Equivalently, the time propagation operator, typically referred to as the *propagator*, acts on quantum states in the following manner:

$$|\psi(t)\rangle = \mathcal{U}(t,t') \left|\psi(t')\right\rangle \Leftrightarrow \left|\psi(t')\right\rangle \rightsquigarrow |\psi(t)\rangle \tag{6.3}$$

and has the form:

$$\mathcal{U}(t,t') = \mathcal{T} \exp\left(-\frac{i}{\hbar} \int_{t'}^{t} \mathcal{H}(t')\, dt'\right) = \exp(i\mathcal{A}(t)) \tag{6.4}$$

where $\mathcal{T}$ is Dyson's *time-ordering operator*, and $\mathcal{A} = \mathcal{A}^\dagger$ is an $N \times N$ Hermitian matrix. Figure 6.1 provides an illustration for the concept of multiple quantum pathways from an initial state to a final state.

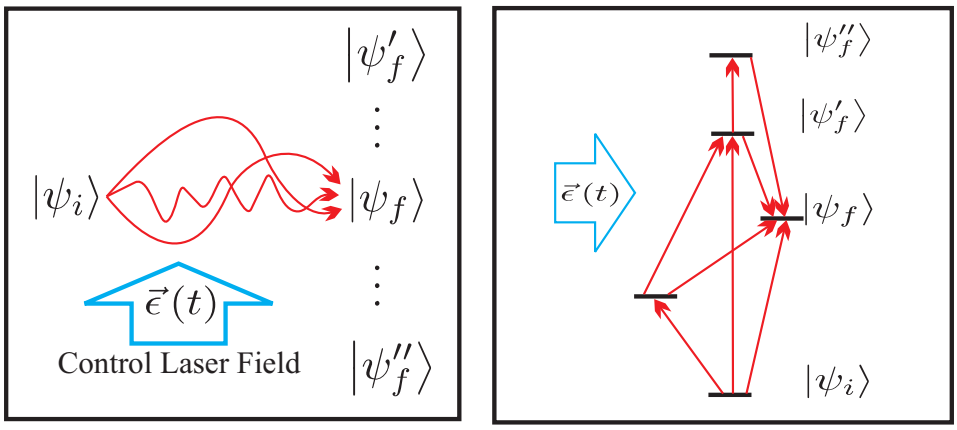

Figure 6.1: **[Left]** Given a quantum system with an initial state $|\psi_i\rangle$, the Quantum Control process aims at steering the system into a desired target state, $|\psi_f\rangle$, by means of the control laser field $\vec{\epsilon}(t)$. Coherent control relies on the existence of multiple quantum pathways between the two states, as illustrated, which result in interference; The goal is thus obtaining constructive interference in the desired final state, and destructive interference elsewhere. **[Right]** The *quantization* of the multiple quantum pathways picture; The transition from the initial state to the target state may be attained in multiple pathways.

Let the target observable operator be $\mathcal{O}$, then the yield of the control process for a pure quantum state is defined as the expectation of the observable operator at time $t = T$:

$$\mathcal{J} = \langle \mathcal{O} \rangle_T = \langle \psi_T | \mathcal{O} | \psi_T \rangle = \left\langle \psi_0 \middle| \mathcal{U}^\dagger \mathcal{O} \mathcal{U} \middle| \psi_0 \right\rangle = \langle \psi_0 | \mathcal{O}_T | \psi_0 \rangle \tag{6.5}$$

while referring from now on to $\mathcal{U}$ as $\mathcal{U}(T, 0)$, unless specified otherwise.

Let $\mathcal{O}_T$ be diagonalized and spanned by means of its eigenvectors:

$$\mathcal{O}_T = \mathcal{U}^\dagger \mathcal{O} \mathcal{U} = \sum_j \sigma_j |\phi_j\rangle \langle \phi_j|, \tag{6.6}$$

then the highest eigenvalue σ_{max} corresponds to the maximal attainable observable value.

When an ensemble of quantum states is under investigation,

$$|\Psi(t)\rangle = \sum_j p_j(t) |\psi_j\rangle,$$

it is characterized by the density operator $\rho(t) = |\Psi(t)\rangle \langle \Psi(t)|$. The dynamics of the ensemble is then dictated by the **von Neumann equation** for the

density operator $\rho(t)$:

$$i\hbar\frac{\partial\rho(t)}{\partial t} = [\mathcal{H}(t), \rho(t)] \tag{6.7}$$

where $[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$.

An observable is measured by $\mathrm{Tr}\,(\rho\mathcal{O})$, and the Quantum Control yield is defined respectively by:

$$\mathcal{J} = \langle\mathcal{O}_T\rangle = \mathrm{Tr}\,(\rho_T\mathcal{O}) = \mathrm{Tr}\left(\mathcal{U}\rho_0\mathcal{U}^\dagger\mathcal{O}\right) \tag{6.8}$$

where

$$\rho_T = \rho(T) = \mathcal{U}\rho_0\mathcal{U}^\dagger$$

Additional auxiliary costs may be imposed on the controls due to constraints, e.g., minimal fluence, and construct respectively a quantum control cost functional of the form:

$$\mathcal{J}' = \mathcal{J} - \lambda\int_0^T g\left(\epsilon(t)\right)dt \tag{6.9}$$

However, in this chapter we restrict our treatment to quantum optimal control problems in the absence of these constraints.

Critical Points: Kinematic Treatment At a *critical point* the differential of the control landscape with respect to $\mathcal{U}$ vanishes. This is the so-called *kinematic treatment* of the critical point analysis, and it reads:

$$\frac{\delta\mathcal{J}}{\delta\mathcal{U}} = 0 \tag{6.10}$$

Since $\mathcal{U}^\dagger\mathcal{U} = \mathbf{I}$, we get

$$\delta\mathcal{U}^\dagger\mathcal{U} + \mathcal{U}^\dagger\delta\mathcal{U} = 0$$

for any $\delta\mathcal{U}$. Eq. 6.10 may be rewritten now as

$$\begin{aligned}\frac{\delta\mathcal{J}}{\delta\mathcal{U}} &= \mathrm{Tr}\left(\delta\mathcal{U}\rho_0\mathcal{U}^\dagger\mathcal{O} + \mathcal{U}\rho_0\delta\mathcal{U}^\dagger\mathcal{O}\right) = \mathrm{Tr}\left(\delta\mathcal{U}\rho_0\mathcal{U}^\dagger\mathcal{O} - \mathcal{U}\rho_0\mathcal{U}^\dagger\delta\mathcal{U}\mathcal{U}^\dagger\mathcal{O}\right) = \\ &= \mathrm{Tr}\left(\left[\rho_0, \mathcal{U}^\dagger\mathcal{O}\mathcal{U}\right]\mathcal{U}^\dagger\delta\mathcal{U}\right) = \left\langle\mathcal{U}\left[\mathcal{U}^\dagger\mathcal{O}\mathcal{U}, \rho_0\right], \delta\mathcal{U}\right\rangle = 0\end{aligned} \tag{6.11}$$

leading to the important result that at a critical point

$$[\mathcal{O}_T, \rho_0] = \left[\mathcal{U}^\dagger\mathcal{O}\mathcal{U}, \rho_0\right] = 0 \tag{6.12}$$

Hence, $\mathcal{O}_T$ and ρ_0 commute, and thus are simultaneously diagonalizable, according to this kinematic treatment.

Critical Points: Dynamic Treatment The *dynamic treatment*, which considers the differential of the observable with respect to the control field $\epsilon(t)$, is typically based on the chain rule:

$$\frac{\delta \mathcal{J}\left[\vec{\epsilon}(t)\right]}{\delta \vec{\epsilon}(t)} = \frac{\delta \mathcal{J}}{\delta \mathcal{U}} \cdot \frac{\delta \mathcal{U}}{\delta \vec{\epsilon}(t)} \tag{6.13}$$

The dynamic picture is more complex, and is subject to a more delicate treatment, accordingly. At a critical point, it could be shown [127] that this differential yields:

$$\frac{\delta \mathcal{J}}{\delta \vec{\epsilon}(t)} = \mathrm{Tr}\left(\left[\mathcal{O}_T, \rho_0\right] \mathbf{B}(t)\right) = 0, \tag{6.14}$$

where $\mathbf{B}(t) = (i/\hbar)\mathcal{U}^{\dagger}(t,0)\nabla_{\vec{\epsilon}}\mathcal{H}(t)\mathcal{U}(t,0)$.

The crucial assumption which is made by the *dynamic treatment* states that the matrix $\mathbf{B}(t)$ forms a set of N^2 linearly independent functions for all time $0 \leq t \leq T$. This assumption obviously leads to $[\mathcal{O}_T, \rho_0] = 0$, as in Eq. 6.12, and to the conclusion that the observable and the density matrix commute in the dynamic picture as well.

When diagonalizing the density matrix, the same eigenvectors of the observable (Eq. 6.6) are used:

$$\rho_0 = \sum_j \lambda_j \left|\phi_j\right\rangle \left\langle\phi_j\right|$$

The control yield now reads:

$$\begin{aligned} \mathcal{J} &= \mathrm{Tr}\left(\sum_i \sum_j \sigma_i \lambda_j \left|\phi_i\right\rangle \left\langle\phi_i\right| \left.\phi_j\right\rangle \left\langle\phi_j\right|\right) = \mathrm{Tr}\left(\sum_j \lambda_j \sigma_{\pi(j)} \left|\phi_j\right\rangle \left\langle\phi_j\right|\right) = \\ &= \sum_j \lambda_j \sigma_{\pi(j)} \end{aligned} \tag{6.15}$$

where $\pi(j)$ denotes a permutation, out of $N!$ possible permutations of these eigenvalues, assuming that there is no degeneracy.

Special Case: $P_{i\to f}$ A special state-to-state case is commonly considered, where the transfer of a pure initial state $|i\rangle$, into a desired final state $|f\rangle$, is subject to maximization. It is expressed accordingly through pure density projectors: A density matrix $\rho_0 = |i\rangle\langle i|$, and an observable $\mathcal{O} = |f\rangle\langle f|$. This *population transfer* problem has a simpler theoretical treatment, and moreover, is also commonly encountered in real-world applications. More explicitly, let us consider the time evolution operator by its matrix element,

$$\mathbf{U}_{if} = \left\langle i\right| \mathbf{U} \left|f\right\rangle \tag{6.16}$$

being a functional of the control field, $\mathbf{U} = \mathbf{U}\left[\epsilon(t)\right]$. Then the quantum control *population transfer problem* is posed as maximizing the probability

$$\mathcal{P}_{i\longrightarrow f} = \left|\mathbf{U}_{if}\right|^2 \tag{6.17}$$

6.1.2 Controllability

By assessing the *controllability* of the quantum system we aim at attaining the existence of a control field which obtains the maximal target yield, without studying the nature of the landscape. This is essentially different from *optimality analysis*, which aims at locating extrema on the landscape, without necessarily conducting controllability assessment.

A powerful aspect of Quantum Control theoretical landscapes is the ability to assess perfect controllability of the system, with hardly any assumptions on the quantum system, as presented in the following theorem:

Theorem 6.1.1. *Assuming controllability of the system, the only extrema values for Quantum Control of population transfer corresponds to perfect control:*

$$\mathcal{P}_{i\longrightarrow f} = 1$$

In the following we shall outline the principal steps of the proof for this claim, following [129, 130]. For simplicity, we choose to consider the special case of $\mathcal{P}_{i\longrightarrow f}$, subject to *dynamic treatment*. Note that $\mathcal{P}_{i\to f} = |\mathbf{U}_{if}|^2$.

Proof Idea A dynamic treatment of a landscape extremum reads:

$$\frac{\delta \mathcal{P}_{i\to f}}{\delta \epsilon(t)} = 0 \tag{6.18}$$

Using the identity

$$\langle i| \mathbf{U} |f\rangle = \langle i| \exp(i\mathbf{A}) |f\rangle ,$$

where $\mathbf{A} = \mathbf{A}^\dagger$ is an $N \times N$ Hermitian matrix, Eq. 6.18 may be rewritten as

$$\frac{\delta \mathcal{P}_{i\to f}}{\delta \epsilon(t)} = \sum_{p,q} \frac{\partial |\mathbf{U}_{if}|^2}{\partial \mathbf{A}_{pq}} \frac{\delta \mathbf{A}_{pq}}{\delta \epsilon(t)} = 0 \tag{6.19}$$

The same crucial assumption made regarding Eq. 6.14 is made here, reducing the dynamic picture into the kinematic picture: The uniqueness of the functional dependence of the matrix elements $\mathbf{A}_{pq}[\epsilon(t)]$ on $\epsilon(t)$ is implied by the assumed controllability of the system.

Eq. 6.18 can now be satisfied by

$$\frac{\partial |\mathbf{U}_{if}|^2}{\partial \mathbf{A}_{pq}} = \frac{\partial}{\partial \mathbf{A}_{pq}} |\langle i| \exp(i\mathbf{A}) |f\rangle|^2 = \mathbf{U}^*_{if} \frac{\partial \mathbf{U}_{if}}{\partial \mathbf{A}_{pq}} + \mathbf{U}_{if} \frac{\partial \mathbf{U}^*_{if}}{\partial \mathbf{A}_{pq}} = 0 \quad \forall p \forall q \tag{6.20}$$

Further examination of this equation (see *Supplemental Online Material* of [129]) leads to the following conclusion:

$$\mathbf{U}_{if} = \exp(i\alpha), \quad \alpha \in \mathbb{R} \tag{6.21}$$

and thus $|\mathbf{U}_{if}| = 1$, and the claim is satisfied accordingly:

$$\mathcal{P}_{i\longrightarrow f} = 1 \tag{6.22}$$

The most general case would be the dynamical treatment of the extrema of $\mathcal{J} = \mathrm{Tr}\left(\rho_T \mathcal{O}\right)$. An equivalent theorem, stating that the extrema of such landscapes would correspond to perfect control or to no-control, exists and is proven in [127]. Furthermore, the latter article presents important results regarding the nature of the landscape, which we choose to review here briefly:

1. **The Slope** An upper bound of the gradient reads:

$$\left|\frac{\delta \mathcal{J}}{\delta \epsilon(t)}\right| \leq \frac{2}{\hbar} \|\mathcal{O}\| \times \|\vec{\mu}\| \tag{6.23}$$

   where the linear polarization of the electric field was assumed for simplicity. In practical realizations, it is reasonable to expect that the landscape slope up to the global maximum will have no steep regions, suggesting that the **optima are robust**.

2. **Hessian at the Global Maximum** The Hessian matrix has typically at most $(2N - n_p - 1)$ non-zero negative eigenvalues (n_p is the number of non-zero eigenvalues of ρ_0), where the rest correspond to the null space, which is spanned by their eigenfunctions. Thus, there exist saddle points, but they do not introduce any obstacle toward locating the global maximum.

3. **Robustness** The trace of the Hessian matrix at the top of the landscape suggests a robust global maximum in any practical realization, and gets more robust as the dimensionality N increases.

We conclude this section by stating the following corollary:

Corollary 6.1.2. *Quantum Control landscapes have extrema that correspond to perfect control or to no-control. Furthermore, given a controllable quantum system, there is always a trap-free pathway up to the top of the control landscape from any location, allowing the location of the global maximum with first-order (gradient) information.*

6.1.3 Control Level Sets

Given the results obtained in the previous section, stating that the gradient of the yield function vanishes only at the top of the landscape, it is possible to draw an important conclusion regarding the existence of **level sets**[1] in the landscape.

[1] This important concept, which was discussed previously in the context of global minimum definition (see Eq. 1.2 and Theorem 1.1.1) or the basin definition (see Definition 2.3.1), is revisited here in the context of success-rate.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be under investigation, with a point in the landscape which satisfies:

$$f^* = f(\vec{x}^*), \quad \nabla f(\vec{x}^*) \neq 0$$

The so-called *Implicit Function Theorem* states that there exists an $(n-1)$-dimensional *manifold* near $\vec{x}^*$ with the same function value of f^*, and its *tangent plane* at $\vec{x}^*$ is perpendicular to $\nabla f(\vec{x}^*)$.

This theorem can be applied directly to Quantum Control landscapes, due to the results presented previously. While climbing up the QC landscape, every associated yield value along the way has a corresponding manifold, which can potentially be explored by continuous trajectories.

Obviously, we cannot apply the same theorem in order to draw an equivalent conclusion regarding the existence of a level set at the top. However, it is possible to show that a denumerably infinite number of solutions exists at the top of the landscape. Under mild assumptions, it was shown in [131, 132] that in the absence of constraints an infinite number of solutions will exist for a general Quantum Control problem. The proof is based on functional analysis treatment, subject to perturbation formulation, and is beyond the scope of this study.

We may conclude that Quantum Control landscapes are not only easy in terms of the location of its maxima, i.e., optimal controls, as suggested previously, but also **offer a rich diversity of multiple solutions**.

The careful reader should note that the above conclusions are valid only for Theoretical Quantum Control landscapes, where no constraints whatsoever are posed. In the context of our work on Quantum Control optimization, to be presented in the following chapters, the landscapes under study will always be underposed by multiple constraints, and thus the degree to which these theorems are applicable is generally unknown. However, possible corroboration of the given Quantum Control landscape analysis might be identified in our work, and will be discussed.

The D-MORPH Algorithm Standard algorithms for the optimization of optimal control are designed for climbing-up the control landscape and locating its extrema at the top, but are not capable of examining the level-sets of the landscape.

A special algorithm for exploring control fields on a given landscape level-set was designed by Rothman et al. [133, 134], aiming to produce trajectories throughout control fields which correspond to a preserved observable. This algorithm is referred to as Diffeomorphic Modulation under Observable-Response-Preserving Homotopy (D-MORPH), and it allows an examination of various control fields which attain the same yield, but may have different physical properties, e.g., fluence.

The basic idea of the D-MORPH algorithm is to constrain the quantum dynamics such that the observable is preserved for all control fields at a given time. It is convenient to introduce a dummy exploration variable s, and present the quantum dynamics accordingly ($0 \leq s \leq 1$):

$$\begin{aligned} \epsilon(s,t) &\longleftarrow \epsilon(t) \\ \mathcal{H}(s,t) &= \mathcal{H}_0(s) - \vec{\mu}(s) \cdot \vec{\epsilon}(s,t) \\ i\hbar \frac{\partial}{\partial t} |\psi(s,t)\rangle &= \mathcal{H}(s,t) |\psi(s,t)\rangle \\ \langle \mathcal{O}(s) \rangle_T &= \langle \psi(s,T) | \mathcal{O} | \psi(s,T) \rangle \end{aligned} \tag{6.24}$$

Given the desired target observable value at time T, denoted by C_T, the D-MORPH algorithm aims at locating control fields $\epsilon(s,t)$ that satisfy the following non-linear equation:

$$F(s) = \langle \mathcal{O}(s) \rangle_T - C_T = 0 \tag{6.25}$$

A homotopy path can then be obtained by solving the following differential equation:

$$\frac{dF(s)}{ds} = \frac{d \langle \mathcal{O}(s) \rangle_T}{ds} = 0 \tag{6.26}$$

We only outline the D-MORPH algorithm above, while omitting most of the explicit derivations of the integration process to be followed. We refer the reader to [133, 134] for those details.

We conclude this section with the following corollary:

Corollary 6.1.3. *A general controllable Quantum Control problem has a rich landscape with an infinite number of optimal solutions, corresponding to perfect control. Climbing-up to the top of the landscape reveals control level-sets at every yield value, with manifolds which can be explored with continuous trajectories. The latter may be obtained by means of the D-MORPH algorithm.*

6.1.4 Computational Complexity

The framework of this study is global optimization, where the focus here is on optimal control of theoretical quantum systems, by means of optimally determining a control field parameterized by n function values. As such, studying its computational complexity aspect would traditionally consider the resources required for the optimization algorithm as a function of the dimensionality of the search space, denoted here by n.

Due to the special nature of quantum systems, studying the time complexity of OCT optimization algorithms with respect to the Hilbert space dimensionality N is of considerable interest. In fact, when considering the computational expense of resources for a given OCT optimization problem,

the propagation of the Schrödinger equation is far more substantial than the scalability of the control field to be optimally determined. Accordingly, the underlying optimization challenge seems to stem from the size of the quantum system N, rather than from the number of the electric field function values to be optimally determined, n.

Hence, OCT computational complexity research focuses on the Hilbert space dimensionality N. It should be noted that *kinematic optimization treatment* of OCT, which is typically not of this study's focus, considers Hermitian matrices of dimension $\mathcal{O}\left(N^2\right)$ as the control. Thus, in the latter case the time complexity anyway has to be treated in terms of N.

We review here briefly a single test case.

Time Complexity of a Pure-State Quantum System

Following Corollary 6.1.2, we know that an OCT search can be algorithmically implemented by means of gradient-based steps. It is thus convenient to consider the *gradient flow*, which is defined as the trajectory followed by the algorithm when the step update follows

$$-\nabla_{\mathcal{U}}\mathcal{J}(\mathcal{U})$$

The latter is based upon the *kinematic treatment* (see Eq. 6.10 and its derivations). It is then possible to estimate an upper bound for the required time for convergence into an ε-neighborhood of the global maximum for the class of observable maximization problems [135]. The upper bound for a pure initial state system, $\rho_0 = |i\rangle\langle i|$, then reads:

$$\tau_{max} \leq \frac{1}{2\left(\sigma_1 - \sigma_{k+1}\right)}\left[\ln\left(\frac{2Nk}{\varepsilon^2}\right) + 2\cdot\ln\left(\frac{\left(N-k-2\right)\sigma_{k+1}}{k\left(\sigma_1-\sigma_{k+1}\right)}\right)\right] \tag{6.27}$$

where N is the Hilbert space dimension, $\sigma_1 > \sigma_{k+1} > \ldots > \sigma_N$ are the eigenvalues of the observable $\mathcal{O}$, and k is the degeneracy of the maximal eigenvalue, σ_1.

OCT optimization has a polynomial number of variables in terms of N, and given the estimation of Eq. 6.27 we may conclude that it has a *logarithmic time complexity*. It thus belongs to the complexity class **CLOG** (continuous log) in the context of the relevant complexity literature (see, e.g., [136]).

OCT computational complexity research is still in its early days, and is currently under promising study. It includes the investigation of other test cases, subject to theoretical as well as empirical approaches.

6.2 Optimal Control Experiments

Optimal Control Experiments (OCE) [116, 137] consider the realization of Quantum Control in the real-life laboratory, aiming at employing a learning process for altering the course of quantum dynamics phenomena of specific target-applications. Here, the yield, or the success-rate, is obtained by a physical measurement of the target application, whereas numerical modeling of the system's Hamiltonian is not required.

Initially, there were several qualitatively different quantum control schemes. Brumer and Shapiro proposed the use of multi-color interference to control quantum systems [112, 138]: Combinations of harmonic light fields were used to control the total and differential cross-sections of photo-ionization and dissociation processes. That approach focused on the frequency-domain description of the quantum system, and it was followed by a proposed Quantum Control approach by Tannor and Rice, based on exploiting the time-evolution of wave packets that are produced when quantum systems interact with short laser pulses [111, 139]. Finally, Rabitz introduced the important concept of *feedback control*, where phase-, amplitude- and/or polarization shaping subject to a closed learning loop are used to guide a quantum system toward a desired final state [113]. Rabitz's approach has been successfully applied in numerous applications, and practically became the common experimental routine in the field. We shall focus in this study on the feedback control approach.

The remainder of this section will review experimental Quantum Control, while focusing in computational and optimization aspects. We do not discuss the technical realization of the actual laser pulse. This part is mainly based on [116, 140], as well as on personal lecture notes[2].

6.2.1 Femtosecond Laser Pulse Shaping

As presented earlier, the control field in OCT corresponds to the electric field, which is tuned in the temporal domain in a straightforward manner by the optimization routine. However, the realization in OCE dramatically differs [116].

When considering laser pulses in the duration of *femtoseconds*[3], it is not yet possible to shape pulses in the temporal domain: State-of-the-art electro-optic switches can currently modulate only in the order of *picoseconds*[4]. Hence, the pulse shaping in OCE is typically implemented by means of "slow" manipulation of the spectrum, subject to a realization of the Fourier

[2] Notes were taken in the course "Quantum Control" of Prof. Herschel Rabitz (CHM509), Princeton University, Fall 2007.

[3] $1\text{fs} = 10^{-15}\text{s}$, i.e., 1 millionth of 1 billionth of a second.

[4] $1\text{ps} = 10^{-12}\text{s}$, i.e., 1 trillionth of a second.

transform. We denote the *experimental* electric field by $E(t)$,

$$E(t) \sim \mathbb{R}\left\{\int_{-\infty}^{\infty} E(\omega)\exp(i\omega t)\ d\omega\right\}$$

where $E(\omega)$ is the spectral field. Pulse shapers allow independent *amplitude* as well as *phase* modulations, and the spectral field may be modeled accordingly:

$$E(\omega) = A(\omega)\exp\left(i\phi(\omega)\right)$$

with $A(\omega)$ as the spectral amplitude, and $\phi(\omega)$ as the spectral phase.

Time vs. Frequency The transition between time to frequency domains is obtained by the Fourier transform, $\mathcal{F}$, whose action can be summarized as follows:

$$\begin{aligned} E(\omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\tilde{E}(t)\exp\left(-i\omega t\right)\ dt = \mathcal{F}\left[\tilde{E}(t)\right] \\ \tilde{E}(t) &= A(t)\exp\left(i\Phi(t)\right) = \int_{-\infty}^{\infty} E(\omega)\exp\left(i\omega t\right)\ d\omega = \mathcal{F}^{-1}\left[E(\omega)\right] \end{aligned} \tag{6.28}$$

where $A(t)$ is the temporal amplitude and $\Phi(t)$ is the temporal phase. In practice, the modeling of the experimental electric field is **real**, and it reads:

$$E(t) = \mathbb{R}\left\{\int_{-\infty}^{\infty} A(\omega)\exp(i\phi(\omega))\exp(i\omega t)\ d\omega\right\} \tag{6.29}$$

The Fourier transform also determines the reciprocal relation between the spectral width to the temporal width, which is another form of the *uncertainty principle*. Given the temporal full-width-half-maximum (FWHM) pulse width, $\Delta\tau_{laser,FWHM}$, and the FWHM spectral width, $\Delta\omega_{laser,FWHM}$, the *time-bandwidth relation* reads:

$$\Delta\omega_{laser,FWHM}\cdot\Delta\tau_{laser,FWHM} \geq 2\pi c_B \tag{6.30}$$

where $c_B \leq 1$ depends on the profile of the spectral amplitude $A(\omega)$.

It is important to distinguish between the *temporal intensity* of the field,

$$I(t) = \left|\tilde{E}(t)\right|^2 \tag{6.31}$$

and the *spectral intensity* of the field,

$$I(\omega) = \left|E(\omega)\right|^2 \tag{6.32}$$

which are strictly not directly related, due to the loss of the phase information.

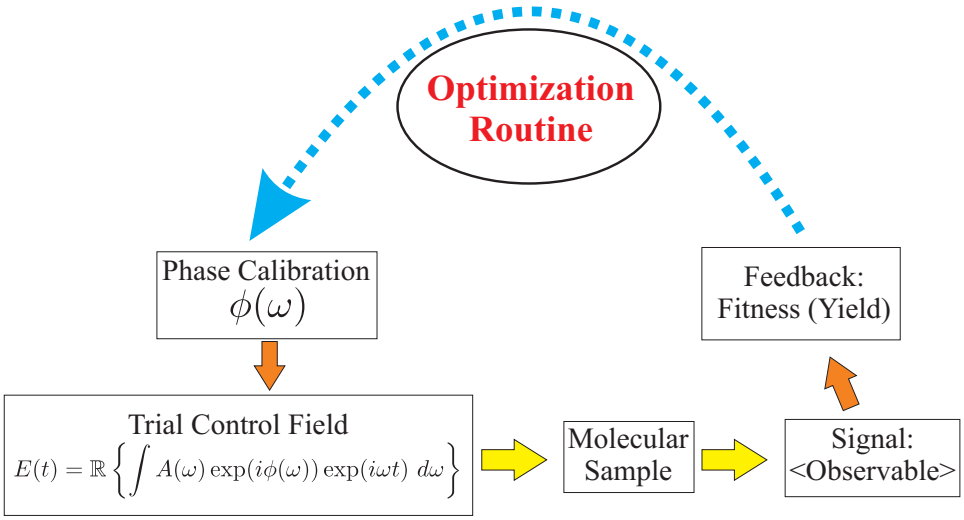

Figure 6.2: The Quantum Control experimental learning loop.

The Control Phase Generally speaking, the control function in spectral modulation consists of the spectral amplitude function $A(\omega)$ as well as of the spectral phase function $\phi(\omega)$. Most Quantum Control processes are more sensitive to the phase than to the amplitude, and phase-only shaping is typically sufficient for attaining optimal control. We thus choose to restrict our study to phase modulation, and to consider the spectral function $A(\omega)$ as fixed. The latter is then well-approximated by a Gaussian which determines the bandwidth, or the pulse duration, accordingly. Note that shaping the pulse with phase-only modulation guarantees the conservation of the pulse energy.

We thus consider only $\phi(\omega)$ as our control function: It defines the spectral phase at n frequencies $\{\omega_i\}_{i=1}^n$, that are equally distributed across the spectrum of the pulse. These n values $\{\phi(\omega_i)\}_{i=1}^n$ correspond to n pixels of the pulse shaper, and they would become the decision parameters to be optimized in the experimental learning loop:

$$\phi(\omega) := (\phi(\omega_1), \phi(\omega_2), ..., \phi(\omega_n)) \tag{6.33}$$

Figure 6.2 illustrates the closed learning loop experimental Quantum Control process.

6.2.2 Laboratory Realization: Constraints

The realization of the quantum system in the laboratory poses constraints on the quantum dynamics, and may lead to a different OCE search landscape, in comparison to its equivalent OCT landscape. The OCT theorems which guarantee a trap-free pathway to perfect control from any location in the

landscape, with gradient-based steps and in logarithmic time complexity, may no longer be valid in OCE landscapes. Generally speaking, it is not clear how do Quantum Control landscapes appear in the laboratory.

We discuss here briefly several aspects of laboratory experiments which are likely to be translated into constraints in the OCE landscape [140].

The crucial component of laser pulse shaping process is the phase modulation, which is typically exposed to waveform distortion effects (for a comprehensive study see [141]). We outline here several modulation components.

Pixelation and Replica Pulses In practice, the pulse shaping process is implemented by a so-called Spatial Light Modulator (SLM), which is typically based on Liquid Crystal Display (LCD). This approach considers individual pixels subject to rectangle-activation-functions, $\mathrm{squ}(\nu)$, ideally sharply-defined and with no gaps between each other. This is referred to as the *staircase approximation*. The time modulation of these step-functions is attained by means of their inverse Fourier transform,

$$\mathcal{F}^{-1}\left[\mathrm{squ}(\nu)\right] \sim \mathrm{sinc}(\tau)$$

where the width of $\mathrm{sinc}(\tau) = \frac{\sin(\tau)}{\tau}$ is inversely proportional to the pixel width. Explicitly, the resulting temporal electric field in this pixelization can be described as follows:

$$e(t) = \sum_n \tilde{e}(t - n\tau) \cdot \mathrm{sinc}\left(\frac{\pi t}{\tau}\right), \tag{6.34}$$

with $\tilde{e}(t)$ as the *desired* electric field, and where $\tau = \frac{1}{\Delta\nu}$ is the inverse frequency spacing per pixel.

Practically, step-function gaps between SLM electrodes are responsible for the construction of so-called *parasitic replica pulses* in the temporal domain, which are located at the zeros of the **sinc** envelope function.

Pulse Break-Up A *linear phase function* results in the time shift of the temporal pulse. This can easily be derived by a change of variables, or by the application of the so-called Fourier Shift Theorem (see, e.g., [142]). The influence of the replica pulses becomes more substantial when they are moved from the zeros of the envelope **sinc** function, by breaking-up the pulse energy into multiple parasitic replica pulses. This is equivalent to the following statement: *The steeper the linear phase, the more pronounced become the replica pulses, which generally result in lower suboptimal yields* [140].

Phase Range: Wrapping Phases that differ in 2π *radians* are mathematically equivalent. This periodic nature of the phase in $[0, 2\pi]^n$ practically

poses periodic boundary conditions on the modulator. Given $0 < \varepsilon \leq 2\pi$, the so-called *phase wrapping* operator is implemented as follows:

$$\begin{array}{lll} \phi_i = 2\pi + \varepsilon & \longrightarrow & \tilde{\phi}_i := \varepsilon \\ \phi_j = -\varepsilon & \longrightarrow & \tilde{\phi}_j := 2\pi - \varepsilon \end{array} \tag{6.35}$$

or simply as $\tilde{\phi}_i := \phi_i \bmod 2\pi$.

From an optimization perspective, this means that the search space is practically an n-dimensional hypercube spanning a length of 2π in each dimension. It is likely to have implications on the optimization routine in use. In terms of constraints, wrapped phases may be exposed to singularity effects ($0 - 1$ jumps), but it is not considered to be a significant effect. Thus, we consider it here more as a mathematical feature of the search space, rather than a constraint.

Resolution The number of pixels, n, determines the **control resolution**, and poses a direct constraint on the shaped-pulse in the temporal domain: Due to the reciprocal nature of the Fourier transform with respect to frequency versus time, spectral resolution determines the upper bound for temporal resolution. For instance, typical laboratory realizations currently consider $n = 128$ pixels with spectral resolution of 0.25 nm/pixel, which allow a shaped pulse with maximum temporal length of $8.5ps$ at FWHM bandwidth of $10nm$.

We hereby summarize the main laboratory constraints in a typical quantum system realization:

1. **Temporal or spectral resolution of the field** Limited spectral resolution in the realized shaper implies limited pulse temporal resolution. State-of-the-art LCD pulse shapers contain 640 pixels to be tuned.

2. **Limited field fluence, limited field intensity** Potential *damage* to different experimental components restricts in practice the applied field fluence and its intensity.

3. **Limited spectral bandwidth or pulse duration** State-of-the-art commercial lasers can produce nowadays pulses at the duration of $\sim 20\ fs$.

4. **Proper basis** The actual representation of the control phase, e.g., pixel basis, polynomial expansion basis, etc., poses by itself an additional constraint on the landscape.

5. **Noise** Existence of laboratory noise, by definition, poses constraints on the landscape.

6.3 Experimental Procedure

In this study we are interested both in numerical modeling of quantum systems, as well as in their real laboratory experiments. The numerical modeling is typically driven by a known Hamiltonian, but designed in a laboratory-oriented manner, as will be described shortly. Essentially, it is OCT combined with some OCE characteristics.

In our calculations, we choose to restrict this study mostly to noise-free simulations, as we are interested in the physics of the system, rather than conducting an actual simulation of a real laboratory experiment. On this note, we consider the absence of noise in our calculations as a blessing, as it allows for clean interpretation of the physics of the system. In one particular case, we will carry out simulations with noise.

Generally speaking, considering the various quantum systems under investigation in this study, the goal that we would like to achieve in our *experimental work* is three-fold, and may be outlined as follows:

1. A preliminary part of our work on each quantum system is devoted to a large extent to an investigation of the performance of specific derandomized Evolution Strategies, as well as parameterizations, with respect to the given optimization task. As suggested in Section 1.4.4, this would include the comma-strategy DES variants.

2. After having identified the routines which perform best on our problems, further work would typically concentrate on the physical interpretation of the obtained optimal solutions, when applicable to the system under study. In particular, we will aim at clarifying why certain pulse structures perform better than other trial solutions. This will also be accompanied with investigation of pulse-intensity, field scalability, and other defining features.

3. Finally, we will be interested in applying miscellaneous optimization techniques, at the level of decision making: *multi-objective optimization*, and the application of *niching*.

Next, we provide technical details concerning the two classes of experimental work conducted in this study: numerical simulations and laboratory experiments.

6.3.1 Numerical Simulations

We present here the numerical modeling of our laser pulse shaping framework, which is in essence valid for all the numerical calculations conducted in this work, unless specified otherwise. The idea is to simulate the experimental pulse shaping process, in terms of control definition, physical limitations, etc.

As discussed earlier, in our calculations the control is solely the phase function $\phi(\omega)$. It defines the phase at n frequencies $\{\omega_i\}_{i=1}^{n}$ that are equally distributed across the spectrum of the pulse. These n values $\{\phi(\omega_i)\}_{i=1}^{n}$ are the decision parameters to be optimally determined. Upon their calibration they are numerically interpolated into $\tilde{n} = 2^{14}$ points, using the `spline()` procedure [143], for the calculation of the electric field in Eq. 6.29. The latter is implemented by means of the `FFT()` procedure [143].

The numerical resolution is naturally underposed to a conflict with the expected optimization efficiency. In order to achieve a good trade-off between the two, i.e., keeping both resolution and optimization efficiency as high as possible, the value of $n = 80$ turned to be a good compromise. The search space is therefore an 80-dimensional hypercube spanning a length of 2π in each dimension.

The spectral function $A(\omega)$ is taken to be a Gaussian, centered at $800nm$, with a width chosen such that the *full-width-at-half-maximum* (FWHM) length of the *Fourier transform limited* (FTL) pulse (obtained by setting $\phi(\omega) \equiv 0$) is $\Delta\tau \approx 100fs$.

Most of the simulations were run with FORTRAN code, as written and provided by Prof. Marc Vrakking, of Amolf-FOM, Amsterdam[5]. This was later combined with a MATLAB version of the original code, as implemented by the author. For the two-photon processes reported in Chapter 7 we used a LabView simulator of Princeton University, coded by Jonathan Roslund.

6.3.2 Laboratory Experiments

The laboratory experiments reported in this work were all conducted at the Frick Laboratory, Rabitz Group, Chemistry Department, Princeton University[6]. The laser source was a Ti:sapphire femtosecond system, with a Tsunami oscillator and a $1kHz$ $1.8mJ$ Spitfire amplifier. A pulse was centered at $\sim 800nm$, with a bandwidth of $\Delta\lambda \approx 10nm$, yielding $\Delta\tau \approx 100fs$ pulse duration at FWHM. The employed SLM consisted of 128 pixels (phase-only modulation, liquid-crystal), but the experiments typically used 64 pixels, by coupling together pairs of adjacent pixels, unless specified otherwise. All algorithms were coded in LabView.

Reference Routine in the Lab: Genetic Algorithm

Genetic Algorithms (GAs) are the most common optimization routines in QC experiments in the vast majority of physics laboratories, likely due to

[5] Dedicated training was given by **Marc Vrakking** and **Christian Siedschlag**, and I thank them both for that.

[6] All experiments were conducted under the dedicated supervision of **Jonathan Roslund** of the Rabitz Group, whose support in running the experiments has been priceless.

historical reasons. As a reference to specific derandomized ES that we apply in our experiments, we shall also report on the GA performance.

The Traditional GA We use the traditional GA [22], with bitstring representation of $l = 6$ bit resolution per pixel. It employs a fixed population of $\mu = 30$ individuals. The mutation rate for a bit-flip is $p_m = 0.005$, and the selection mechanism keeps the fittest offspring, as well as the single best individual of the previous generation (*elitism*). It should be noted that these parameters were collectively optimized to allow sufficient resolution so as to arrive at the highest quality solution with the fastest convergence.

You should understand the physics, write down the correct equations, and let nature do the calculations.
Peter Debye

Chapter 7

Two Photon Processes

7.1 Introduction

The field of *non-linear optics* describes optical phenomena which are observed when high intensity light passes through media. The non-linearity is due to the interaction between the light, typically a laser field, and a dielectric media, whose field-induced polarization responds non-linearly to the incident electric field.

Given the *temporal intensity* of the electric field, $I(t)$, its non-linear signal of the k^{th} order is modeled for $k > 1$ as:

$$Signal_{NL}^{(k)} \propto \int_{-\infty}^{\infty} I^k(t)dt, \tag{7.1}$$

corresponding to the interaction of k photos.

The field of non-linear optics offers a variety of popular Quantum Control applications. Second-order variants, which correspond to two-photon processes, are particularly attractive because of their easy implementation in the laboratory, as well as their known mathematical formulation. Two-photon processes can be utilized to explore experimental Quantum Control landscapes, and also can form a realistic testbed for global optimization algorithms.

This chapter is devoted to the formal definition of two-photon processes, their mathematical description, and to the application of optimization routines to their signal-maximization problems in the laboratory.

7.2 Second Harmonic Generation

Second harmonic generation (SHG) or *frequency doubling* is a two-photon process in which an electric field interacts non-linearly with a material and generates an output photon with double the energy of two input photons. The total energy of the output light is proportional to the integrated squared

intensity of the primary pulse, as expected from a second-order non-linear process.

The time-dependent profile of the laser field is exactly as given in Eq. 6.29. The SHG signal is then defined by:

$$SHG_t \equiv S_t = \int_{-\infty}^{\infty} I(t)^2 \, dt = \int_{-\infty}^{\infty} |E(t)|^4 \, dt, \tag{7.2}$$

i.e., integration over time of the intensity. SHG is a process that turns out to be a good test case in the laboratory, and its investigation contributes to the understanding of other processes. This is because the SHG is a measure of the pulse duration, and this property is useful as an auxiliary characteristic. From the theoretical point of view, the SHG is a simple test function, with some interesting mathematical properties that will be fully derived here, but yet not an easy optimization task for global optimizers.

7.2.1 Total SHG

In order to gain a better insight into the problem, we provide here the reader with some of its mathematical properties. Especially, we would like to derive the equivalence between time and frequency pictures. The following section is mainly based on Bracewell [142].

Definition 7.2.1. Given the spectral amplitude equipped with the complex phases, $E(\omega) = A(\omega)\exp(i\phi(\omega))$, consider *its autocorrelation (convolution) function* $E_2(\omega)$:

$$E_2(\omega) = E(\omega) * E(\omega) = \int_{-\infty}^{\infty} E(\Omega) \cdot E(\omega - \Omega) d\Omega$$

We would like to show how this *autocorrelation* function in the frequency domain is linked to the time domain:

Theorem 7.2.2. *The autocorrelation function of the spectral amplitude,* $E_2(\omega)$*, is proportional to the Fourier transform of the squared time-dependent electric field, i.e.:*

$$E_2(\omega) \propto \int_{-\infty}^{\infty} \tilde{E}(t)^2 \exp\left(-i\omega t\right) dt \tag{7.3}$$

Proof.

$$
\begin{aligned}
E_2(\omega) &= \int_{-\infty}^{\infty} E(\Omega) \cdot E(\omega - \Omega) d\Omega = \\
&= \int_{-\infty}^{\infty} \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{E}(t) \exp\left(-i\Omega t\right) dt \right] \cdot \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{E}(\tau) \exp\left(-i(\omega - \Omega)\tau\right) d\tau \right] d\Omega = \\
&= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{E}(t)\tilde{E}(\tau) \exp\left(-i\Omega(t-\tau)\right) \cdot \exp\left(-i\omega\tau\right) d\Omega \; dt \; d\tau = \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{E}(t)\tilde{E}(\tau)\delta(t-\tau) \exp\left(-i\omega\tau\right) dt \; d\tau = \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{E}(t)\tilde{E}(t) \exp\left(-i\omega t\right) dt = \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{E}(t)^2 \exp\left(-i\omega t\right) dt
\end{aligned}
$$

where $\delta(x - \tilde{x})$ is the *Dirac delta function.* □

Theorem 7.2.3. *(***Plancherel's Theorem***) Given $f(x)$, which has the Fourier transform $F(s)$, the integral over the squared modulus of $f(x)$ is equal to the integral over the squared modulus of its spectrum $F(s)$:*

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \int_{-\infty}^{\infty} |F(s)|^2 ds$$

See [142]. Thus, we can conclude from Theorems 7.2.2 and 7.2.3 that

$$\int_{-\infty}^{\infty} |E_2(\omega)|^2 d\omega = \int_{-\infty}^{\infty} |E(t)|^4 dt$$

and, equivalently, in terms of the intensities

$$S_t = \int_{-\infty}^{\infty} I_2(\omega) d\omega = \int_{-\infty}^{\infty} I(t)^2 dt \tag{7.4}$$

where $I_2(\omega) = |E_2(\omega)|^2$.

Global Maximum

Theorem 7.2.4. *The Total-SHG signal is maximized by the phase being any linear function of frequency, and in particular by the constant phase:*

$$\mathrm{argmax}_{\phi(\omega)} \left\{ \mathrm{S_t} \left(\phi \left(\omega \right) \right) \right\} \equiv \mathrm{a} \cdot \omega + \mathrm{b}$$

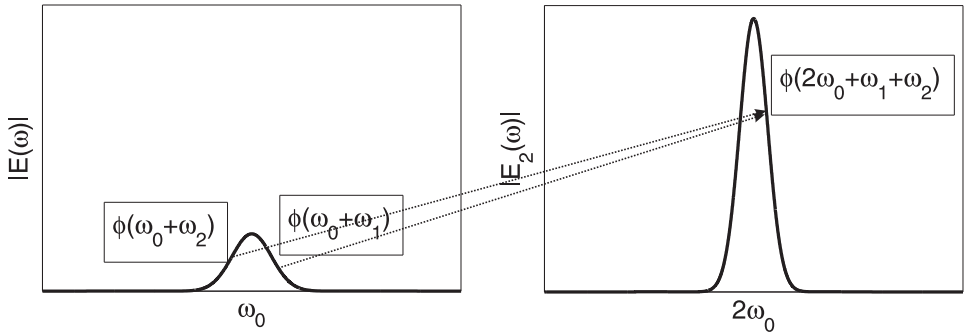

Figure 7.1: An illustration of the frequency doubling effect in Second Harmonic Generation. Construction of $E_2(\omega)$ out of $E(\omega)$.

An important remark should be made concerning the existence of a single optimal solution for the SHG maximization problem: Due to the use of second-order perturbation theory, the **constant phase is a point in the control space** (the generalization to a linear phase stems from symmetry), i.e., the level-set collapses into a single point. In higher-order corrections for SHG the maximally attained yield can be obtained by various other phase profiles.

Figure 7.1 provides the reader with an illustration for the so-called frequency doubling effect - the contribution of two phase points around the central frequency ω_0 at $E(\omega)$, $\phi(\omega_0+\omega_1)$ and $\phi(\omega_0+\omega_2)$, to the construction of $\tilde{E}(\omega)$ with $\phi(2\cdot\omega_0+\omega_1+\omega_2)$. Note the shift in the central frequency, and the scaling of the Gaussian.

7.2.2 Filtered SHG

We consider another second-order quantum optical system, which could be considered as a *filtered* case of the SHG system. It corresponds to a *two photon absorption* (TPA) process, whose model describes, within the limits of second-order time-dependent perturbation theory, the probability of making a transition from a ground state $|g\rangle$ to an excited state $|e\rangle$, upon the activation of the laser field. Thus, a specific *transition frequency* is considered here, ω_{eg}, which practically filters the signal,

$$SHG_f \equiv S_f(\omega_{eg}) = \int_{-\infty}^{\infty} \delta(\omega_{eg}-\omega)\, I_2(\omega) d\omega,$$

by means of the Dirac delta function $\delta(\Omega-\Omega')$. It explicitly reads

$$S_f(\omega_{eg}) = \left| \int_{-\infty}^{\infty} E(\omega) E(\omega_{eg}-\omega)\, d\omega \right|^2 \tag{7.5}$$

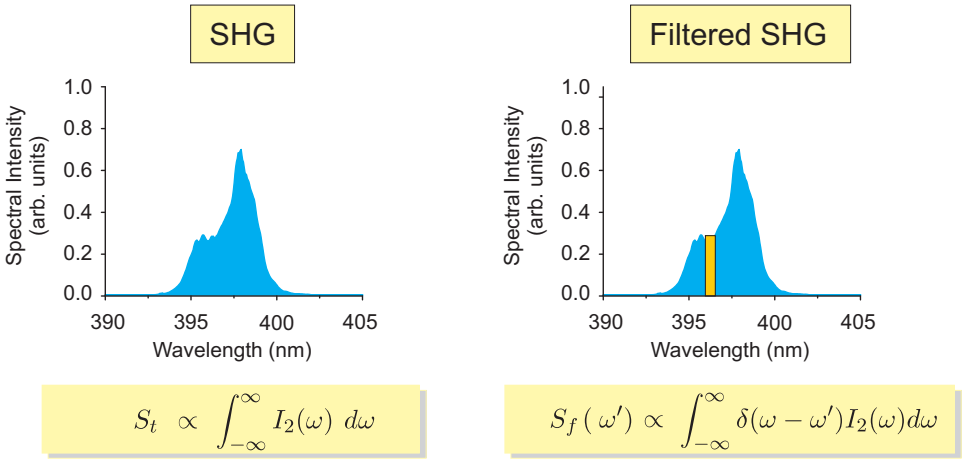

Figure 7.2: A spectral illustration for the total-SHG (left) versus the filtered-SHG (right) signals. Figure courtesy of Jonathan Roslund.

Global Maximum

Theorem 7.2.5. *The filtered-SHG signal is maximized by the phase being any odd function of frequency antisymmetric about* $\frac{\omega_{eg}}{2}$*, i.e., spectral phases of the form* $\phi(\frac{\omega_{eg}}{2} - \omega) = -\phi(\frac{\omega_{eg}}{2} + \omega)$.

See [144, 145]. Figure 7.2 provides an illustrative comparison between the two SHG variants considered here.

Problem Difficulty: Numerical Assessment

In order to assess the optimization difficulty of the Second Harmonic Generation maximization problems, we considered numerical simulations of the two SHG problem variants and conducted the following simple statistical test. We considered phase functions pixelized by $n = 64$ function values, which are randomly initialized in the interval $[0, 2\pi]^{64}$. We then gradually transformed the given random phases into a *zero-phase* in two different routines: **(1)** Setting function values to zero when consistently indexing from right to left, or **(2)** Setting function values to zero in random permutation of indices, with no repetition. Both routines eventually obtain zero-phases, which attain the maximal yield of 1 for both SHG problem variants.

Figure 7.3 presents typical runs for the two routines when applied to both SHG problem variants. It is observed in these plots that approximately 50% of the function values must be set to zero in order to enhance the yield value, for all cases. Once this threshold is exceeded, the yield value increases consistently until it reaches the value of 1. The actual profiles of routine (1) versus routine (2) differ, for both SHG variants. More variables are required to be set to zero in the random indexing routine, in comparison to

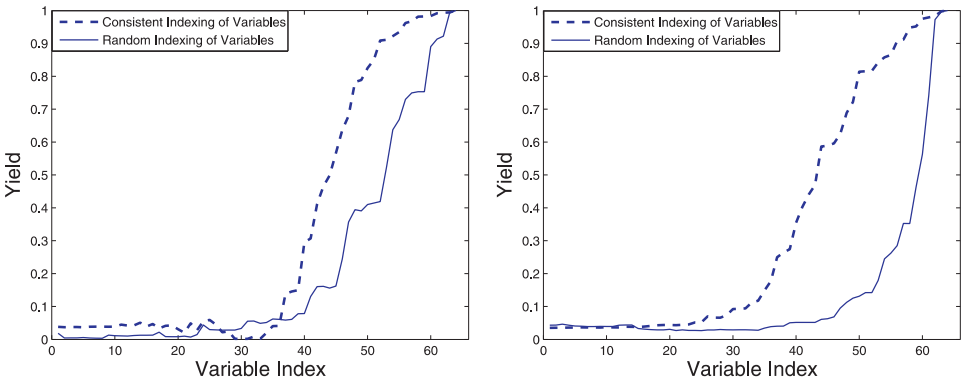

Figure 7.3: Transforming randomly-initialized phases into a zero-phase, pixel-by-pixel, either by **(1)** Consistently indexing the phase function from right to left, or by **(2)** Randomly selecting phase function indices, without repetition. The attained yield per index-step is recorded for each test-case. Typical runs are presented for the two routines applied to the SHG problem variants. Left: Filtered-SHG system; Right: Total-SHG system.

the consistent indexing. This is due to the shape of the weighting function (i.e., a Gaussian), which limits the contribution to the yield value from pixels which are not in the proximity of the central frequency.

This statistical test reveals that the SHG problems under investigation are non-separable upon following the formal definition.

7.3 Numerical Simulations

We present here results of the four derandomized ES comma-variants when applied to numerical simulations of second-order photon processes: The maximization of the Total-SHG as well as the Filtered-SHG signals.

7.3.1 Preliminary ES Failure: Stretched Phases

When applied to both SHG simulations, the derandomized ES variants suffered from pre-mature convergence to sub-optimal solutions of low yield. Upon examination of the attained optimized phases in the decision space, they were always observed to be *highly steep linear phases*. We offer the following explanation for that.

The ES is not subject to any restrictions concerning its decision parameters, in particular in the context of the periodic nature of the phase. It seems that an unrestricted search, as employed by the ES variants in hand, is likely to stretch the candidate phases, with no way to reverse it. It suffers accordingly from convergence to highly steep linear phases with sub-optimal

Table 7.1: Derandomized Evolution Strategies optimizing the Total-SHG simulation: Mean and standard-deviation of attained yield over 100 runs for the three procedures – unrestricted, wrapped and bounded.

Algorithm	**Unrestricted**	**Wrapped**	**Bounded**
DR1	0.208 ± 0.072	0.873 ± 0.187	0.574 ± 0.189
DR2	0.181 ± 0.064	0.967 ± 0.019	0.725 ± 0.185
DR3	0.457 ± 0.198	0.718 ± 0.274	0.529 ± 0.278
CMA	0.581 ± 0.136	1 ± 0	0.997 ± 0.002

Table 7.2: Derandomized Evolution Strategies optimizing the Filtered-SHG simulation: Mean and standard-deviation of attained yield over 100 runs for the three procedures – unrestricted, wrapped and bounded.

Algorithm	**Unrestricted**	**Wrapped**	**Bounded**
DR1	0.257 ± 0.087	0.666 ± 0.247	0.713 ± 0.152
DR2	0.248 ± 0.091	0.804 ± 0.195	0.908 ± 0.125
DR3	0.539 ± 0.162	0.762 ± 0.209	0.554 ± 0.173
CMA	0.487 ± 0.134	0.990 ± 0.008	0.964 ± 0.052

yield values, as outlined earlier in Section 6.2.2. By implementing *periodic boundary conditions* into the ES algorithms, by means of coupling the *wrapping operator* (Eq. 6.35) to the mutation operator, this problem was solved. This procedure will be referred to as the **wrapped procedure**.

As a third procedure, we also considered the application of a boundary operator that fixes an exceeded value to the lower or upper bounds. Given $\varepsilon > 0$, it reads:

$$\begin{aligned} \phi_i = 2\pi + \varepsilon \quad &\longrightarrow \quad \tilde{\phi}_i := 2\pi \\ \phi_j = -\varepsilon \quad &\longrightarrow \quad \tilde{\phi}_j := 0 \end{aligned} \tag{7.6}$$

It is referred to as the **bounded procedure**.

7.3.2 Numerical Observation

Tables 7.1 and 7.2 summarize the numerical results of the application of the four derandomized ES comma-variants to the total-SHG and filtered-SHG simulation problems, respectively, subject to the three specified procedures, with $n = 64$ decision parameters. There are two clear observations from the given calculations:

1. The *wrapping operator* seems to be an essential component for the unrestricted ES optimization, and should be implemented into ES when

optimizing "phase" variables on a QC landscapes. This is an expected conclusion, given the nature of the search space. However, it is interesting to note the relatively high standard deviations for the results obtained subject to wrapping for the filtered-SHG case for the first three DES variants. Also, it is observed that the bounded approach works better for the DR2 on the filtered-SHG landscape.

2. The CMA outperformed the other algorithms on these two landscapes, with consistent winning performance. The DR2 was second-best, and it performed in a highly satisfactory manner. We thus hold two DES variants, each representing first- or second-order information approach, respectively, which performed well on these QC landscapes.

Intermediate Discussion

We found that employing the ES variants with default settings unrestrictively on the given QC landscapes resulted in pre-mature convergence to sub-optimal phases with highly sloped linear profiles. We analyzed this effect, and introduced the wrapping operator into the ES framework. The latter solved the observed problem.

7.4 Laboratory Experiments

We report here on laboratory experiments where we aimed at optimizing the two quantum control systems described in Section 7.2. Due to the tremendous effort and time which are required for a reliable experiment, we had no choice but to restrict ourselves to a limited number of experiments as well as optimization routines.

We chose to employ three optimization routines in the laboratory:

- DR2: First-order DES.
- CMA: Second-order DES.
- GA: Laboratory reference.

Concerning the technical details, for total-SHG signal, S_t, the amplified pulses are delivered to a 100 μm type-I BBO crystal, and the time integrated SHG signal is recorded with a photodiode and boxcar integrator. For the filtered-SHG signal, S_f, unamplified seed pulses are focused onto a 100 μm type-I BBO crystal, and the resultant up-converted light is analyzed with a spectrometer. Regarding the actual yield values recorded by us, we choose to normalize the FTL signal as yield 1.0 for both systems.

It should be noted that the SHG optimization problems have been widely investigated at several levels, including at laboratory experiments [146], where it was shown to have a highly complex landscape.

Table 7.3: Laboratory SHG Optimization: Performance Evaluation. The experimental results of the two SHG systems, averaged over 10 experiments. The final yield (averaged over the last 50 iterations) and the number of evaluations required to cross a yield threshold of 0.90 are considered here.

Routine	Filtered-SHG		Total-SHG	
	Avg. Yield	0.9 Eval	Avg. Yield	0.9 Eval
GA	0.95	4665	0.95	5557
DR2	0.93	2159	0.72	NA
CMA	0.95	841	0.98	766

ES Failure Revisited: Stretched Phases When applied to the experimental setup, the derandomized ES variants initially suffered from premature convergence to sub-optimal solutions of yield ≈ 0.75, where the maximum value is 1.0. Upon examination of the attained optimized phases in the decision space, the stretching effect as reported in Section 7.3.1 was observed. Thus, we used the *wrapping operator* in the two DES variants in all the reported experiments. The GA, on the other hand, did not typically locate highly-steep linear phases since the $[0, 2\pi]$ bounds are implicitly implemented by means of the *phenotypic mapping* (see, e.g., [22]).

7.4.1 Performance Evaluations

Table 7.3 presents the results of the two reported systems, averaged over 10 experiments. We consider the final yield (averaged over the last 50 iterations), as well as the number of evaluations required to cross a yield threshold of 0.90, as the performance criteria per experiment. Figure 7.4 presents averaging of the runs, with attained yield as a function of the required number of function evaluations. Note that this averaging procedure takes into account all 10 runs, whereas the convergence data shown in Table 7.3 considers only the relevant runs that exceeded the 0.90 yield threshold. Figure 7.5 presents histograms for the different algorithms with final yield versus the number of runs.

As reflected from the experimental results, the CMA performed best on the given experimental systems, both in terms of final yield as well as convergence speed. We would like to emphasize the extraordinary boost of convergence speed provided by the CMA relative to the GA, which is significant in the laboratory. Moreover, the CMA has a sharp and rapid convergence profile, in contrast to the inefficient hill-climbing capability of the GA. This profile is easy to identify as there is no ambiguity about convergence, and thus it is another attractive feature for the laboratory user.

Next, we discuss the experimental results and the algorithmic behavior.

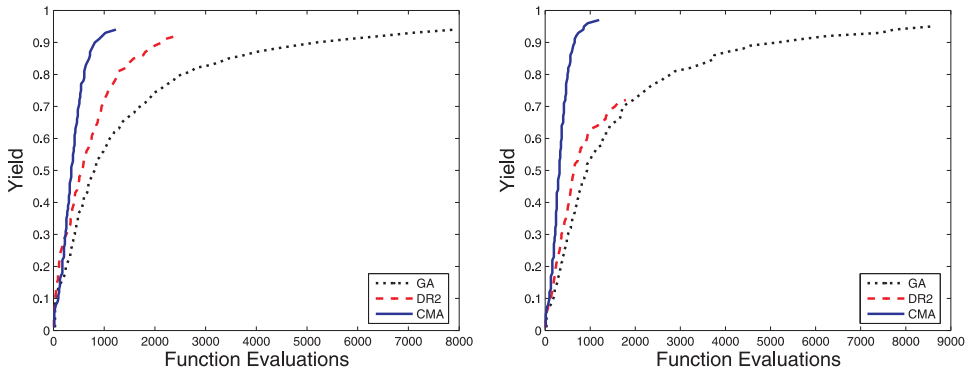

Figure 7.4: Averaged runs of the algorithms over 10 runs. Left: Filtered-SHG system; Right: Total-SHG system.

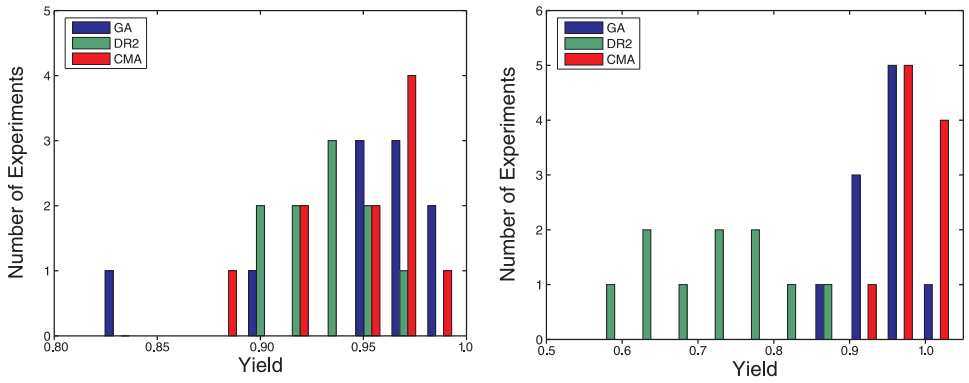

Figure 7.5: Success-rate (yield) histograms. Left: Filtered-SHG system; Right: Total-SHG system.

Diversity of Solutions

As mentioned earlier in Section 7.2.2, the filtered SHG system possesses a family of nontrivial phases that correspond to global maxima. Interestingly, each run for the filtered SHG case converged to a distinct antisymmetric phase. This collection of different solutions provided a practical perspective concerning the richness of QC landscapes and their underlying level sets.

Sensitivity to Noise

The CMA-ES and the GA performed in a satisfactory manner on the given control problems and did not seem to be significantly impaired by the existence of noise in the experimental system. The DR2, on the other hand, suffered from high-sensitivity to the initial step-size. Its performance was disappointing, in particular in comparison to noise-free calculations that were

reported in the past [147, 148]. A proposed explanation for this behavior could be the lack of recombination, which has been shown to be a crucial ES component in noisy environments (see, e.g., [149]).

Covariance Learning

Recording the CMA data during the optimizations allows an analysis of the evolutionary search process. Upon examination of the data, it is found that the covariance matrix remains diagonal during the search (Eq. 1.41), or equivalently, the CMA does not utilize its second-order mechanism (i.e., *rotations*) when climbing up the landscape. This is not a surprising result, but rather an important piece of experimental evidence toward the corroboration of the OCT landscape analysis as outlined in Corollary 6.1.2.

Figure 7.6 presents a typical CMA run for the optimization of total-SHG in the laboratory and shows the yield and step-size upon function evaluations. Figure 7.7 presents the square-roots of the covariance matrix eigenvalues as a function of the number of experiments as well as the Euclidean distances between the best phase variables of successive iterations, i.e.,

$$d^{(g+1)} = \|\vec{\phi}_{best}^{(g+1)}(\omega) - \vec{\phi}_{best}^{(g)}(\omega)\|, \tag{7.7}$$

where $\vec{\phi}_{best}(\omega)$ is as in Eq. 6.33.

We conducted an equivalent test in a **noise-free simulator** for the total-SHG problem[1]. Figure 7.8 presents a typical CMA run on the simulator. The convergence profile on the simulator is observed to be similar to the laboratory experiment, i.e., rapid climbing-up of the landscape without utilizing the second-order mechanism. However, upon approaching the *top of the landscape*, one of the covariance matrix eigenvalues dramatically grows, as shown in Figure 7.9. This behavior was observed to be typical in all runs. The corresponding eigenvector is always a flat phase, suggesting that the CMA discovers the invariance of a constant phase on the total-SHG signal. The phase Euclidean trajectories are plotted as well in Figure 7.9, showing some minor activity during this growth stage, corresponding to super-fine tuning of the spectral phase. The yield values, nonetheless, do not seem to be further improved during this process, at least in the precision available. In practice, the parameter adaptation during this fine-tuning stage produces fitness variations below that of the system noise in the laboratory, which explains its absence in laboratory optimizations.

Simulations: Zeroth-Order CMA

Given the experimental observation reported in the previous section, we were interested in testing the CMA while removing its covariance learning

[1] The simulator was implemented in LabView with the Lab2 package.

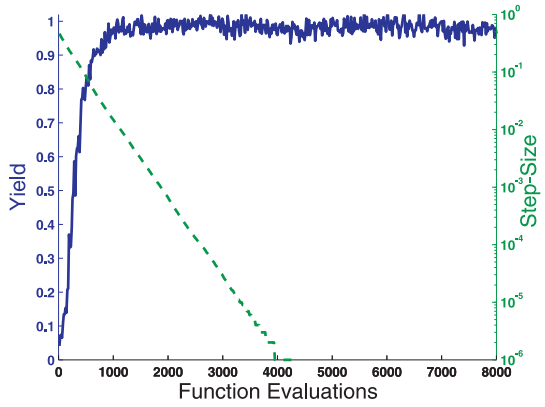

Figure 7.6: CMA optimization of the Total-SHG in the **laboratory**. Yield (solid line, left axis) and step-size (dashed line, right log-scaled axis), versus function evaluations.

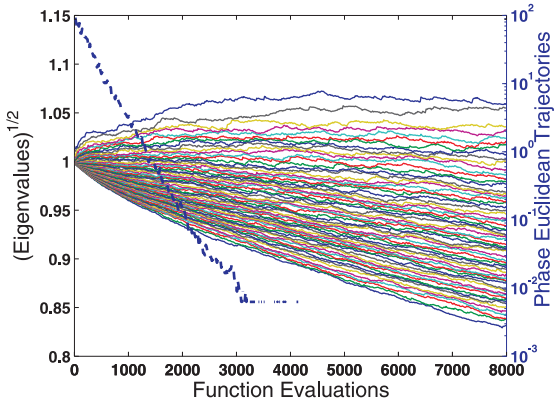

Figure 7.7: CMA optimization of the Total-SHG in the **laboratory**. Square-root of the 64 eigenvalues of the covariance matrix (solid thin lines, left axis), and phase Euclidean trajectories (bold points, right log-scaled axis), versus function evaluations. Missing trajectory points correspond to zero values.

components. In essence, we leave the CMA only with the step-size as a strategy parameter, and fix the covariance matrix as an identity matrix. This is a zeroth-order ES with normal mutations subject to hyperspheres as the equidensity probability surfaces. In order to assess the zeroth-order CMA behavior on the given QC systems, we conducted additional simulations with two variants of the algorithm:

- (μ_W, λ)-CMA with $\mathbf{C} = \mathbb{I}$.
- $(1, \lambda)$-CMA with $\mathbf{C} = \mathbb{I}$.

The simulations were conducted for both systems - total-SHG as well as filtered-SHG - both with a noise-free simulator and a simulator with noise.

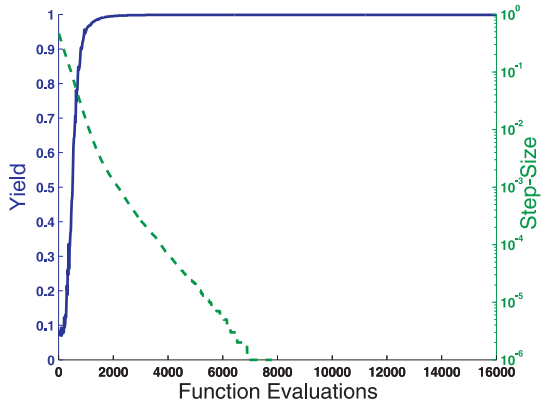

Figure 7.8: CMA optimization of the Total-SHG on a **noise-free simulator**. Yield (solid line, left axis) and step-size (dashed line, right log-scaled axis), versus function evaluations.

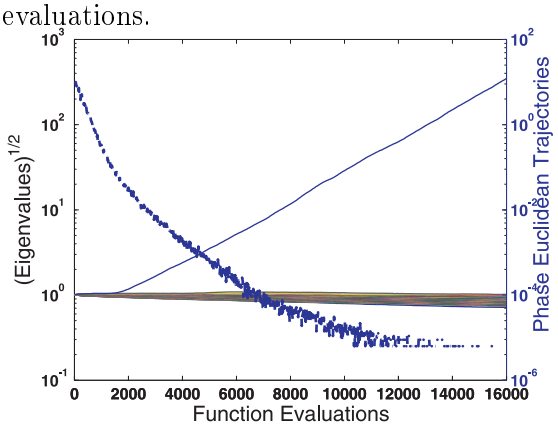

Figure 7.9: CMA optimization of the Total-SHG on a **noise-free simulator**. Square-root of the 64 eigenvalues of the covariance matrix (solid thin lines, left log-scaled axis), and phase Euclidean trajectories (bold points, right log-scaled axis), versus function evaluations. Missing trajectory points correspond to zero values. The single exploding eigenvalue can easily be identified in this scale.

The results of the simulations show that the CMA performance is not hampered at all on both systems when removing its covariance learning components: the (μ_W, λ)-CMA with $\mathbf{C} = \mathbb{I}$ performs as well as the original CMA, in terms of final attained yield and convergence speed. This observation is valid for noise-free as well as for noisy simulations. However, when the weighted recombination operator was removed, the $(1, \lambda)$-CMA with $\mathbf{C} = \mathbb{I}$ did not converge, nor did it even climb-up from the initial yield at the bottom of the landscape. We thus conclude that it is possible to optimize the given simulated QC landscapes by a zeroth-order ES, as long as the weighted-recombination operator is kept.

7.4.2 Discussion

We presented a survey of derandomized Evolution Strategies and a Genetic Algorithm to a set of Quantum Control systems in the laboratory. As far as we know, this was one of the first applications of *derandomized* ES to experimental QC in general, and the first study to conduct a comparison between ES to GA as well as to explore the evolutionary path of the CMA, in particular. We would like to mention, however, two studies [150, 151] that applied Evolution Strategies to OCE, and explored a specific QC system both in experiments and simulations. The latter studies concluded that the employed Evolution Strategies were promising optimization routines.

While the QC systems examined here possess easily understood global optima, the search is conducted over a highly complex, curvilinear control landscape, which provides a good testbed for optimization algorithms. From the practical point of view, these systems are relatively easy for implementation in the laboratory.

We found that employing the ES variants with default settings unrestrictively on the given QC landscapes resulted in pre-mature convergence to sub-optimal phases with highly sloped linear profiles. We analyzed this effect, and introduced the wrapping operator into the ES framework. The latter solved the observed problem.

The CMA-ES outperformed the other algorithms in terms of final yield as well as in convergence speed. It introduced a significant increase in convergence speed to the typical performance of the GA in the laboratory and is a promising tool for future laboratory experiments. While analyzing its behavior, it was experimentally confirmed that its second-order mechanism was not utilized when climbing-up the landscape. This may be considered as an experimental corroboration of the OCT landscape analysis.

We also conducted *noise-free simulations* of the CMA-ES applied to the systems. The latter calculations revealed interesting behavior of the covariance matrix, upon approaching the top of the landscape. A single eigenvalue consistently explodes with a corresponding eigenvector of a flat phase. We suggest that this is due to the fact that the CMA successfully learned the invariance of a constant phase in these problems. Furthermore, we considered zeroth-order versions of the CMA in simulations, where the covariance learning component was removed. The latter performed extremely well, as long as the weighted-recombination operator was kept.

It is the theory that decides what can be observed.
Albert Einstein

Chapter 8

The Rotational Framework

The main Quantum Control application of this study is *dynamic molecular alignment*, which will be presented in the next chapter. The current chapter considers the *rotational* framework of molecules, as a preparation for the alignment application. We describe here the formal numerical modeling basis, and present calculations for the optimization of population transfer. Finally, we apply our niching algorithms to the population transfer problem.

8.1 Numerical Modeling

We consider here Hamiltonians that consist of a molecular part $\mathcal{H}_0$, while the interaction with the *semi-classical* laser field subject to the *dipole approximation* is expressed by V:

$$\begin{aligned} \mathcal{H}(t) &= \mathcal{H}_0 - V \\ V &= \mu E(t)\cos(\omega t) \end{aligned} \tag{8.1}$$

The envelope of the laser field, which completely determines the dynamics, is exactly as introduced in Eq. 6.29:

$$E(t) = \mathbb{R}\left\{\int_{-\infty}^{\infty} A(\omega)\exp(i\phi(\omega))\exp(i\omega t)\ d\omega\right\}$$

8.1.1 Preliminary: Two Electronic States Systems

We start by outlining the fundamental details of a *two-electronic-state system*. This section is mainly based on [152].

Consider a system with two *electronic states*: The ground state $|g\rangle$, and an off-resonant excited state $|e\rangle$ with energy $\hbar\omega_0$. Its wavefunction may be described as follows:

$$|\Psi(t)\rangle = \alpha_g(t)\,|g\rangle + \alpha_e(t)\exp(-i\omega_0 t)\,|e\rangle \tag{8.2}$$

Upon applying the Schrödinger equation,

$$i\hbar\frac{\partial\left|\Psi(t)\right\rangle}{\partial t}=\mathcal{H}\left|\Psi(t)\right\rangle, \tag{8.3}$$

by using a Hamiltonian of the form of Eq. 8.1, two coupled differential equations are obtained:

$$\begin{aligned} i\hbar\dot{\alpha}_g(t) &= -\exp(-i\omega t)E(t)\left\langle g\right|\mu\left|e\right\rangle\alpha_e(t) \\ i\hbar\dot{\alpha}_e(t) &= -\exp(i\omega t)E(t)\left\langle e\right|\mu\left|g\right\rangle\alpha_g(t)-\hbar\Delta\alpha_e(t) \end{aligned} \tag{8.4}$$

where $\Delta=\omega-\omega_0$ is the so-called *detuning*.

In order to keep the description as general as possible, the *peak field strength* is not fixed explicitly; Instead, we set the peak Rabi frequency $\Omega(t)$ for the transition between the electronic states $|g\rangle$ and $|e\rangle$, which is proportional to the product of peak field strength and the coupling matrix element between $|g\rangle$ and $|e\rangle$:

$$\Omega(t)=\frac{\left\langle g\right|\mu\left|e\right\rangle\tilde{E}(t)}{2\hbar}, \tag{8.5}$$

where we used the complex form of the electric field, $\tilde{E}(t)$ (see Eq. 6.28). Also, it is convenient to note:

$$\Omega_{ge}=\frac{\left\langle g\right|\mu\left|e\right\rangle}{2\hbar} \tag{8.6}$$

The differential equations for the expansion coefficients of the wavefunction may be written now in a matrix notation as follows:

$$i\begin{pmatrix}\dot{\alpha}_g(t)\\ \dot{\alpha}_e(t)\end{pmatrix}=-\begin{pmatrix}0 & \Omega(t)\\ \Omega^*(t) & \Delta\end{pmatrix}\begin{pmatrix}\alpha_g(t)\\ \alpha_e(t)\end{pmatrix} \tag{8.7}$$

The Rabi frequency thus determines the interaction strength in our framework.

8.1.2 Rotational Levels

We proceed by describing the rotational framework of the molecules. This section is mainly based on [153]. We consider a model of diatomic linear molecules that populate rotational levels in a given temperature T. The molecules are characterized by their rotational quantum number, J, as well as by the projection of the angular momentum on the laser polarization axis, M. We take the molecule to be a *rigid rotor*, which allows a description of its wavefunction solely in terms of the rotational *eigenstates* $|JKM\rangle$, where $K=0$ for a diatomic molecule. We take into account the two *electronic states*, as presented earlier: Ground state $|g\rangle$ and off-resonant excited state $|e\rangle$. The wavefunction, for a given M, is thus expanded as follows:

$$\left|\Psi_M(t)\right\rangle=\sum_{J=M}^{N_{rot}}\alpha_{JM}^{(g)}(t)\left|gJM\right\rangle+\exp(-i\omega_0 t)\alpha_{JM}^{(e)}(t)\left|eJM\right\rangle \tag{8.8}$$

The molecular component of the Hamiltonian can be divided into two parts,

$$\mathcal{H}_0 = \mathcal{H}_{elec} + \mathcal{H}_{rot}, \tag{8.9}$$

that correspond to the following eigenstates:

$$\begin{aligned} \mathcal{H}_{elec} \left|gJM\right\rangle &= 0 \\ \mathcal{H}_{elec} \left|eJM\right\rangle &= \hbar\omega_0 \left|eJM\right\rangle \end{aligned} \tag{8.10}$$

$$\begin{aligned} \mathcal{H}_{rot} \left|gJM\right\rangle &= B_g J(J+1) \left|gJM\right\rangle \\ \mathcal{H}_{rot} \left|eJM\right\rangle &= B_e J(J+1) \left|eJM\right\rangle \end{aligned} \tag{8.11}$$

with B_g and B_e as the *rotational constants of the molecule.*

The time dependence description of the molecular wavefunction is given by:

$$i\hbar \frac{\partial \left|\Psi_M(t)\right\rangle}{\partial t} = \mathcal{H} \left|\Psi_M(t)\right\rangle \tag{8.12}$$

The laser field induces transitions between the rotational states which, in the off-resonant case, occur via subsequent Raman processes. The transitions between $|g\rangle$ and $|e\rangle$ are assumed to proceed via the selection rules of the quantum numbers $\Delta J = \pm 1, \Delta M = 0$.

The derivation concludes with the following differential equations for the expansion coefficients of the wavefunction:

$$\begin{aligned} &\dot{\alpha}_J^{(g)}(t) = -\frac{i}{\hbar} B_g J(J+1) \alpha_J^{(g)}(t) + i\Omega(t) \left\langle J\right| \cos\theta \left|J+1\right\rangle \alpha_{J+1}^{(e)}(t) + \\ &+ i\Omega(t) \left\langle J\right| \cos\theta \left|J-1\right\rangle \alpha_{J-1}^{(e)}(t) \\ &\dot{\alpha}_J^{(e)}(t) = \left[i\Delta - \frac{i}{\hbar} B_e J(J+1) \right] \alpha_J^{(e)}(t) + i\Omega^*(t) \left\langle J\right| \cos\theta \left|J+1\right\rangle \alpha_{J+1}^{(g)}(t) + \\ &+ i\Omega^*(t) \left\langle J\right| \cos\theta \left|J-1\right\rangle \alpha_{J-1}^{(g)}(t) \end{aligned} \tag{8.13}$$

where

$$\begin{aligned} \left\langle J\right| \cos\theta \left|J+1\right\rangle &= \sqrt{\frac{(J+1)^2}{(2J+3)(2J+1)}} \\ \left\langle J\right| \cos\theta \left|J-1\right\rangle &= \sqrt{\frac{J^2}{(2J+1)(2J-1)}} \end{aligned} \tag{8.14}$$

8.2 Population Transfer: Optimization

We consider here the problem of population transfer within the rotational framework as an optimization problem, subject to the numerical modeling for diatomic molecules presented earlier. The objective to be met is defined as the probability to populate a specific target rotational level, given the initial ground state:

$$\mathcal{J} := \mathcal{P}_{i \longrightarrow f}, \quad \left|i\right\rangle = \left|gJ=0\right\rangle, \ \left|f\right\rangle = \left|gJ_{target}\right\rangle, \tag{8.15}$$

where possibly $J_{target} \in \{0, 2, 4, 6, 8, \ldots, N_{rot}\}$. In our calculations, the yield subject to maximization is simply $\left|\alpha^{(g)}_{J_{target}}(T)\right|^2$, in terms of the notation introduced earlier. Also, by definition, $M = 0$.

We consider $N_{rot} = 20$, where this expansion was confirmed to give converged results in the present calculations. The molecule under investigation has a rotational constant of $B_{rot} = B_g = B_e = 5\text{cm}^{-1}$.

Solving the defining differential equations for the population transfer problem (Eq. 8.13) is obviously computationally expensive. In practice, given an electric field, a single evaluation of the resulting wavepacket has the duration of approximately 5s on a single P4-HT $2.6GHz$ processor. We are thus interested in optimization procedures with as minimal function evaluations as possible.

8.2.1 Experimental Procedure

There are several defining parameters in the present calculations. Some of them are critical, as they pose direct constraints on the quantum system at hand, and practically determine its controllability. In our model, such parameters are the peak Rabi frequency, which plays the equivalent role of the laser intensity, as well as the pulse duration. Setting these two parameters defines the simulated physical system. Given the target rotational level, it is then possible to aim at steering the system toward it. Thus, we choose to consider the population transfer as a function of these two defining parameters, where the focus will be on specific values that reflect best state-of-the-art laboratory experiments.

From the algorithmic perspective, we choose to restrict our calculations to the DR2 and the CMA algorithms, which performed best on the Two-Photon Process problems. They both employ small populations, and consider first-order and second-order information, respectively.

Preliminary Runs Preliminary calculations revealed a clear picture, which could have been predicted by intuition[1]. These preliminary calculations were consisted of 10 runs per algorithm on $J_{target} = \{0, 2, 4, 6, 8\}$ with the following peak Rabi frequencies:

$$\Omega_{ge} = \{40, 60, 80, \ldots, 160, 180\} \times 10^{12}\text{s}^{-1}.$$

Given a Rabi frequency of $\Omega_{ge} = 160 \times 10^{12}\text{s}^{-1}$, the quantum system could easily be steered into perfect control for low J values ($J = \{0, 2, 4\}$). This task became infeasible for higher J values with the given Rabi frequency. However, when the latter was increased, e.g., $\Omega_{ge} = 180 \times 10^{12}\text{s}^{-1}$, it became

[1] As much as intuition exists for Quantum Mechanics; "My batting average on intuition is close to zero in quantum control, and I wear that zero average proudly" (Herschel Rabitz, *private communications*).

feasible. Hence, there is a trend of controllability as a function of the laser intensity, especially for the higher rotational levels. As far as the algorithmic performance was concerned, the DR2 and the CMA performed equally well on the given systems. Most importantly, there was never a situation where the DR2 obtained controllability on a given system on which the CMA did not, nor vice versa.

We consider the case of a target rotational level of $J = 4$ as an interesting case-study. This is due to the fact that it allows perfect control at $\Omega_{ge} = 160 \times 10^{12}\mathrm{s}^{-1}$, but yet it is a challenging task for the optimization routines. Also, the effect of decreasing the peak Rabi frequency while losing controllability can be observed relatively easily.

8.2.2 Numerical Observation: $J = 0 \longrightarrow J = 4$

We applied the DR2 algorithm to the optimization of the population transfer problem from $J = 0$ to $J = 4$. These optimizations were performed for three values of the peak Rabi frequency:

$$\Omega_{ge} = \left\{80 \times 10^{12}\mathrm{s}^{-1},\ 120 \times 10^{12}\mathrm{s}^{-1},\ 160 \times 10^{12}\mathrm{s}^{-1}\right\}.$$

All calculations were carried out with 80 runs, limited to $10,000$ function evaluations per run. These calculations obtained qualitatively different results for the three intensities considered. For $\Omega_{ge} = 80 \times 10^{12}\mathrm{s}^{-1}$ the optimizations were unable to accomplish the transfer from $J = 0$ to $J = 4$ with unit efficiency. The best efficiency obtained was $\approx 32\%$. For $\Omega_{ge} = 120 \times 10^{12}\mathrm{s}^{-1}$ and for $\Omega_{ge} = 160 \times 10^{12}\mathrm{s}^{-1}$ the transfer efficiency approached 100% in most of the calculations.

Aiming at comparing the results of individual optimization runs, we **define** a correlation coefficient that compares pulse-shapes attained in two runs i and j, by means of their field intensities:

$$c_{i,j} = \frac{\max_{\Delta t}\left\{\sum_t I_i(t) I_j(t+\Delta t)\right\}}{\left[\sqrt{\sum_t I_i^2(t)}\sqrt{\sum_t I_j^2(t)}\right]} \tag{8.16}$$

where $I_i(t)$ and $I_j(t)$ are the field intensities of the pulses obtained in runs i and j, respectively. Taking the maximum as a function of Δt is due to the fact that pulse-shapes attained by the optimization may be shifted with respect to each other. The sums are over the discrete time steps, as conducted in the numerical calculation. Eq. 8.16 thus yields $c_{i,i} = 1$, and $c_{i,j} = 0$ if pulses i and j do not overlap at all.

Case 1: $\Omega_{ge} = 80 \times 10^{12}\mathrm{s}^{-1}$ Figure A.4 presents the correlation coefficient for the 80 optimization runs of the $\Omega_{ge} = 80 \times 10^{12}\mathrm{s}^{-1}$ test-case. The runs are sorted based on their success-rate (see top panel in the plot). From Figure A.4 we conclude that all solutions that approach the maximum observed

population are highly correlated. Upon examination of the actual calculations, it is observed that all of these solutions are very close to a single FTL pulse. Deviations from the FTL pulse do not only lead to a drop in the correlation coefficient, but also in the population transfer yield.

Case 2: $\Omega_{ge} = 120 \times 10^{12}\mathrm{s}^{-1}$ In Figure A.5 the correlation coefficient is plotted for the 80 optimization runs that were performed for the $\Omega_{ge} = 120\times10^{12}\mathrm{s}^{-1}$ test-case. Here, the laser pulse energy was sufficient to transfer population from $J = 0$ to $J = 4$ with near-unit efficiency. The best solutions, which have a population transfer efficiency of 99.982% and 99.98%, were only weakly correlated to each other, and were only weakly correlated to most of the other solutions. Specifically, there were only 9 solutions among the set of 80 that share a correlation coefficient larger than 0.95 with the best solution (indexed as 1). Many of the remaining solutions are strongly correlated with the 3^{rd}-best solution, which has a population transfer yield of 99.975%: As many as 41 solutions shared a correlation coefficient larger than 0.95 with that solution (indexed as 3). While the three good solutions 1, 2, and 3 are rather different from each other, they contain most of the dominant features of the identified optimized solutions.

Solutions 1-3 are presented in Figure 8.1. Despite their different characteristics, all three solutions in Figure 8.1 are dominated by a series of peaks with a separation of 4.79×10^{-13}s. This corresponds to the beating period of a coherent superposition of $J = 2$ and $J = 4$ ($\Delta E = 14B$). Additional good solutions likely exist, possibly continuously connected on a common *level set*, and further special numerical methods are needed to explore this possibility, such as the D-MORPH algorithm (Section 6.1.3).

Case 3: $\Omega_{ge} = 160\times10^{12}\mathrm{s}^{-1}$ Figure A.6 presents the correlation coefficient for 80 optimization runs of the $\Omega_{ge} = 160 \times 10^{12}\mathrm{s}^{-1}$ test-case. While the degree of population transfer is very high in almost all the runs at this intensity, the correlation between the various solutions is very limited. Clearly, a large number of solutions that transfer the population with unit efficiency coexist, with very little commonality between them. Indeed, inspection of the actual pulse shapes obtained in these runs reveals highly complicated pulses, with few regular features, and an absence of the peak arising from coherence between $J = 2$ and $J = 4$ in the Fourier transform power spectrum.

8.2.3 Intermediate Discussion

Upon increasing the intensity from $\Omega_{ge} = 80\times10^{12}\mathrm{s}^{-1}$ to $\Omega_{ge} = 160\times10^{12}\mathrm{s}^{-1}$ we find that population transfer is accomplished with an ever increasing number of distinguishable solutions.

The results presented here can be viewed as additional experimental corroboration to the results outlined in Corollary 6.1.2, where it was concluded

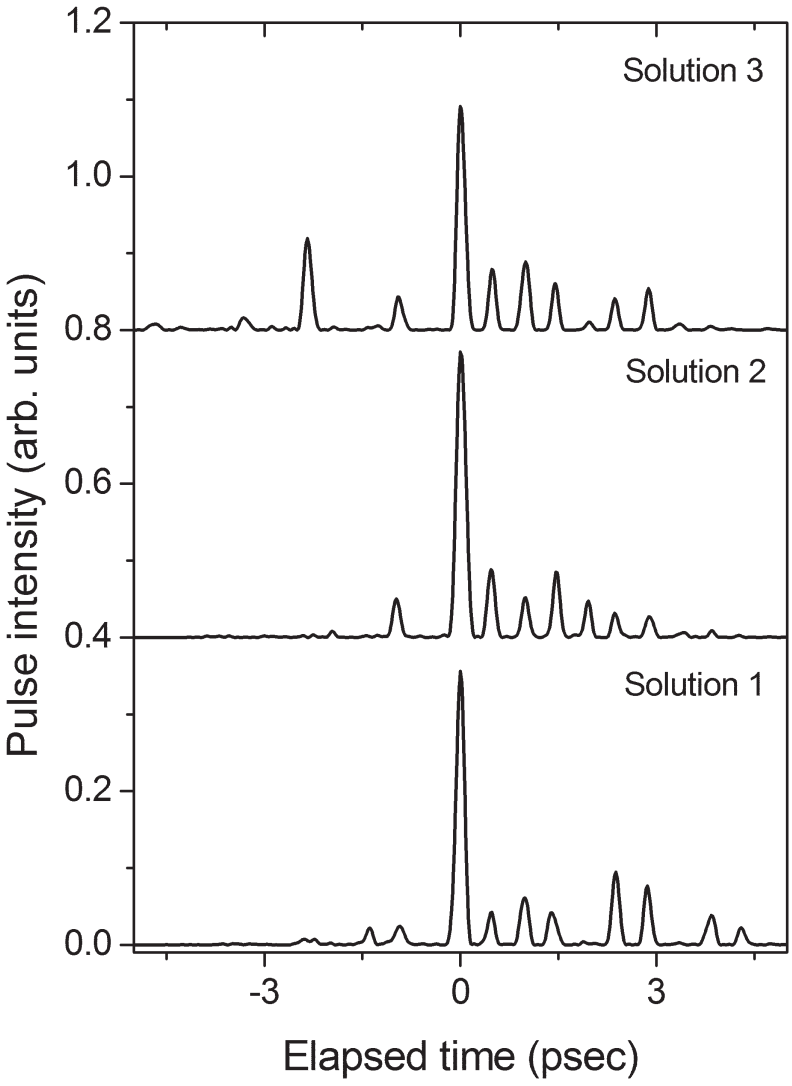

Figure 8.1: Comparison of the 3 best-performing pulse shapes that were obtained in 80 runs of the DR2 for the population transfer problem of $J = 0 \longrightarrow J = 4$ at $\Omega_{ge} = 120 \times 10^{12}\mathrm{s}^{-1}$. All solutions consist of trains of pulses with a spacing of 4.79×10^{-13}s, which corresponds to the beating period between $J = 2$ and $J = 4$.

that controllable quantum systems with no constraints placed on the controls only have extrema that correspond to perfect control, or to no control at all; Additional analysis revealed the fundamental nature of control level sets (see Corollary 6.1.3) at the absolute extrema and at sub-optimal control yields.

A striking aspect of the results is the evidence that the number of independent solutions produced by an optimization seems to critically depend on the difficulty of the problem. In the current population transfer calculations we observed that at low intensity, where reaching the target is a hard problem with less than perfect yield, the trials invariably converge onto one and the same solution, whereas at higher intensity, where this represents an easier problem, a wide variety of solutions are encountered.

8.3 Application of Niching

Motivation: Landscape Richness The numerical observation of the previous section, as summarized in the intermediate discussion, provides us with the strong motivation to apply niching to the problem. The revealed *richness* of the landscape, as predicted by OCT theorems but assessed here on our constrained OCE/OCT-combined landscape, is considered by us as a welcoming invitation for the niching framework.

8.3.1 Preliminary: Distance Measure

Upon applying niching to Quantum Control landscapes, we are required to define an appropriate *distance metric*. Although Eq. 8.16 already provides us with a possible diversity measure, we would like to select a distance metric which is as close as possible to the decision parameters, i.e., the control phase space. We shall then apply Eq. 8.16 for assessing the diversity of the attained solutions.

When considering the decision frequency space, one should keep in mind that the attained field calculations are invariant under the following transformations:

- $\tilde{\phi}(\omega) = \phi(\omega) + \phi_0$: This would add a multiplication constant after the Fourier transform is calculated.
- $\tilde{\phi}(\omega) = \phi(\omega) + c \cdot \omega$: This would simply shift the entire pulse with respect to the time origin and therefore has no observable effect.

These invariance properties must be taken into account when defining a distance measure between two individuals in the decision space, $\phi_i(\omega)$ and $\phi_j(\omega)$, as it is clear that using the straightforward approach of the Euclidean distance would not accomplish the desired goal: Due to the fact that $\phi(\omega)$

is invariant under the specified transformations, calculating the distance between two feasible solutions, $\phi_i(\omega)$ and $\phi_j(\omega)$, would not guarantee that the derived pulse-shapes, $I_i(t)$ and $I_j(t)$, respectively, would have different profiles. Thus, a new distance measure that would remove this degeneracy is much needed here.

Our proposed solution is to apply the distance metric in the **second-derivative space** of $\phi(\omega)$, where the invariance properties vanish. Explicitly, given that the discretization is to n function values, the distance between $\phi_i(\omega)$, $\phi_j(\omega)$ is defined as follows:

$$d_{i,j} = \sqrt{\sum_{k=1}^{n} \left(\left(\frac{\partial^2 \phi_i(\omega)}{\partial \omega^2} \right)_k - \left(\frac{\partial^2 \phi_j(\omega)}{\partial \omega^2} \right)_k \right)^2} \tag{8.17}$$

8.3.2 Numerical Observation

We consider here three niching strategies:

1. The $(1, \lambda)$-DR2 - as a representative of first-order information approach.
2. The $(1, \lambda)$-CMA - as a representative of second-order information approach.
3. The $(1 + \lambda)$-CMA - as a representative of elitist strategies.

We conduct 10 runs per method, searching for $q = 3$ niches, subject to phase-function parameterization of $n = 80$. Each run was limited to $10,000$ function evaluations per niche.

The results of our calculations are discussed at several levels.

Niche-Radius

Numerically, the derivative is simply implemented by means of the MATLAB command `diff`. Thus, after the double-application of `diff` to the original phase-vector of dimension $n = 80$, the modified vector $\vec{y}$ is reduced to dimension $n^* = n - 2 = 78$. Given the original upper and lower bound values of the decision parameters,

$$x_{k,min} = 0, \; x_{k,max} = +2\pi \qquad k = 1..80,$$

the first application of `diff` will make new bound values of

$$\tilde{x}_{k,min} = -2\pi, \; \tilde{x}_{k,max} = +2\pi \qquad k = 1..79,$$

and the second application will make it

$$y_{k,min} = -4\pi, \; y_{k,max} = +4\pi \qquad k = 1..78.$$

Table 8.1: Three niches obtained in 10 runs – averaged yield values (in parentheses - best value attained) – for the three employed niching strategies.

Ranked-Niches	**DR2**	**CMA**	**CMA+**
Best niche	0.9999 (0.9999)	0.9892 (0.9923)	0.9992 (0.9997)
2^{nd}-best niche	0.9745 (0.9910)	0.7391 (0.9797)	0.9982 (0.9995)
3^{rd}-best niche	0.2293 (0.2984)	0.0951 (0.1619)	0.9780 (0.9972)

When plugging this into Eq. 3.5, we obtain:

$$\rho = \frac{\frac{1}{2}\sqrt{78 \cdot (8\pi)^2}}{3^{\frac{1}{78}}} \approx 110 \tag{8.18}$$

The initial setting of the niche-radius, $\rho = 110$, failed to obtain satisfying performance. The DR2 as well as the CMA-comma routines did not succeed in obtaining good solutions. The CMA-plus, however, managed to locate good solutions for the first niche only; the second and third niches were not populated by good solutions. Upon dividing the niche radius by half, i.e., $\tilde{\rho} = 55$, we started to obtain satisfying results, as will be reported here. We shall offer an explanation for this observation in the discussion to follow in the end of this section.

Success-Rate

The averaged as well as maximally attained yield values of the three methods, for the three obtained niches, are presented in Table 8.1. It can be concluded that niching with the CMA-plus kernel typically obtains the best three niches in terms of the population-transfer yield. Niching with the DR2 as well as the CMA-comma kernels always obtain a first niche of high quality. The DR2 typically obtains a very good second niche, but fails in obtaining a third-best niche of high quality. The CMA-comma, on the other hand, typically fails to obtain second- and third-best niches of satisfying quality.

Niches Cross-Correlation

In order to verify that the resulting niches indeed represent sufficiently different pulse shapes, we calculated the cross-correlation coefficients for the obtained pulse-shapes, as defined in Eq. 8.16. The results of these calculations are presented in Table 8.2. In addition, we can state that a correlation value larger than 0.8 was never observed. Based on these findings, we can conclude that the pulse-shapes of the different niches are weakly correlated to one another, as originally desired.

Table 8.2: Niches correlation for the niches obtained in 10 runs – averaged cross-correlation values, as defined in Eq. 8.16 – for the three employed niching strategies.

Niches Correlation	**DR2**	**CMA**	**CMA+**
$c_{1,2}$	0.6583	0.7244	0.6883
$c_{1,3}$	0.6982	0.6835	0.6993
$c_{2,3}$	0.6471	0.7181	0.7154

Discussion

We would like to summarize our numerical observation of the applied niching algorithms to the population transfer problem within the rotational framework. We have identified a degeneracy in the default diversity-measure between candidate solutions, due to some invariance properties of the Fourier transform in the decision space. We offered a problem-specific diversity measure to overcome it. Upon its employment, the latter was shown to be successful, as the obtained pulse-shapes differed considerably. This was also assessed by means of the calculation of the correlation coefficients between the pulse-shapes, which were observed to be low.

The original theoretical calculation of the niche radius was not observed to be successful at the practical level. The results reported here were obtained only after introducing a factor of 0.5 to the original value. We believe that this suggests a landscape with a limited regime of good solutions. Essentially, following the argumentation given in Section 3.5.3, which considered the niche formation process subject to a fixed niche radius as a constrained optimization problem, we argue that introducing a large niche radius would pose a highly constrained problem. This should remind us that the proposed formula for the niche radius is merely an approximation, and moreover, we should keep in mind that the niche radius is a sensitive yet crucial component of this mechanism.

In terms of algorithmic performance, the CMA-plus performed best when obtaining typically three niches of high-quality pulses. The DR2 succeeded in obtaining a first and second good niches, but failed in the third niche. The CMA-comma was observed to typically obtain only a single niche of a high-quality pulse.

We believe that the observed incompetence of the niching framework with the comma-strategy kernels to obtain good results in the secondary niches is due to the landscape properties in general, and the limited regimes of high-quality basins of attraction. Furthermore, we would like to speculate that the failure of the originally employed niche-radius is linked to the failure of the comma-strategies in obtaining good secondary optima.

I can safely say that nobody understands Quantum Mechanics.
Richard Feynman

Chapter 9

Dynamic Molecular Alignment

The Quantum Control application to dynamic molecular alignment [153, 154] is of considerable interest because of its many practical consequences. For instance, many chemical and physical processes, ranging from bimolecular reactions [155] to high harmonic generation [156], are directly influenced by the angular distribution of the molecular sample. Furthermore, in many fundamental molecular dissociation or ionization experiments the interpretation of the collected data will become more efficient if the molecules are aligned with respect to a certain axis. Hence, techniques to generate molecular alignment are needed in practice.

Achieving molecular alignment can be classified into two possible modes:

1. **Pendular State** When the envelope of the field changes slowly compared to the timescale of molecular rotation, typically in the *picosecond* regime, each rotational state of the initial Boltzmann distribution is transformed adiabatically into a *pendular state*. The drawback of this approach is that any alignment produced while the field is turned on will vanish once it is turned off again. Thus, such experiments cannot be carried out subject to field-free conditions.

2. **Impulsive Alignment** Here, the duration of the applied pulses is much shorter than a rotational period [157]. A wavepacket of rotational states is constructed such that *field-free alignment* can be considerably attained.

Both modes aim at constructing a superposition of as many angular momentum eigenstates as possible. Due to the *uncertainty principle*, a broad distribution in *angular momentum* corresponds to a narrow distribution of the *angular position*. However, it is important to note that both the amplitudes and the relative phases of the composite rotational states have to be under control in order to achieve alignment. This requirement is fulfilled for the pendular state case, since it is an eigenstate of the combined molecule-field

Hamiltonian. However, in the general case, a randomly phased superposition of rotational states will not interfere favorably in attaining molecular alignment.

For the *impulsive* case, the evolution of the total wavefunction (after the electric field is turned off) repeats with the **revival time**

$$T_{rev} = \frac{1}{2B_{rot}c} \tag{9.1}$$

where B_{rot} is the rotational constant of the molecule and c is the speed of light. Partial revivals can be observed at $T_{rev}/2$ and, possibly, at $T_{rev}/4$, when one-half or one-quarter, respectively, of the populated rotational levels have undergone an identical number of rotations. Shaped femtosecond laser pulses that lead to a high degree of alignment manage to maximize the number of rotational states that are in phase at these times. However, they have to fulfill an additional requirement: Low field intensities should be applied in order to avoid a scenario in which the molecules are ionized. This aspect also plays a role in keeping the numerical modeling consistent in describing the molecule as a *rigid rotator*, as discussed in Chapter 8. Therefore, one would like to achieve high alignment while keeping the peak laser intensity as low as possible.

On that note, recent publications have focused on finding pulse shapes other than the FTL pulse that create a high degree of alignment. Leibscher el al. [158, 159] have theoretically shown that in the nonperturbative regime a train of pulses lead to better alignment than a single FTL pulse. For asymmetric molecules, *orientation* has been found to be optimized by a sequence of kicks as well [160].

Such pulse sequences can be easily constructed and also optimized with respect to the relatively small number of their control parameters. Therefore, they provide an attractive starting point for more complex optimization schemes, where the electric field is defined by a considerably larger number of control parameters. The task of obtaining high-quality solutions in this high-dimensional search space is nontrivial, already when considering only the ground state in the initial distribution. For finite temperatures, the alignment optimization has to be performed simultaneously for a set of initial rotational states, which, together with the large number of electric field control parameters poses a challenging optimization problem.

9.1 Numerical Modeling

The numerical modeling of the rotational framework, as presented in Chapter 8, is adopted here fully. The remaining task is the definition of the *alignment observable*.

The alignment calculation uses the following components in our basis:

$$\begin{aligned}
&\langle JM|\cos^2\theta\,|JM\rangle = \frac{1}{3}+\frac{2}{3}\left(\frac{J(J+1)-3M^2}{(2J+3)(2J-1)}\right)\\
&\langle JM|\cos^2\theta\,|J+2\ M\rangle =\\
&\frac{1}{2J+3}\sqrt{\frac{(J+M+2)(J+M+1)(J-M+2)(J-M+1)}{(2J+5)(2J+1)}}\\
&\langle JM|\cos^2\theta\,|J-2\ M\rangle = \frac{1}{2J-1}\sqrt{\frac{(J+M)(J+M-1)(J-M)(J-M-1)}{(2J+1)(2J-3)}}
\end{aligned} \tag{9.2}$$

We consider a thermal ensemble of diatomic molecules undergoing irradiation at a finite temperature. The latter is set to $T = 100\ K$, and implemented by means of a Boltzmann averaging which practically corresponds to the density matrix ρ. The molecule under investigation has a rotational constant of $B_{rot} = B_g = B_e = 5\mathrm{cm}^{-1}$. We set the Rabi peak frequency to $\Omega_{ge} = 180 \times 10^{12}\mathrm{s}^{-1}$.

For the sake of attaining high molecular alignment while keeping the peak field intensity as low as possible, due to the rigid rotator approximation, we introduce a constraint to the optimization procedure, by means of a penalty term to pulses that are too intense. It explicitly reads

$$I_p = \int E^2(t)\Theta(E^2(t) - I_{thr})\ dt \tag{9.3}$$

with $\Theta(x)$ as the Heaviside step function.

Thus, the fitness function assigned to a candidate pulse shape is defined by

$$F = \max_{\mathrm{E(t)}} \left\langle \cos^2(\theta) \right\rangle - \beta \mathrm{I_p}. \tag{9.4}$$

By choosing β large enough, I_{thr} can be used to effectively operate the evolutionary search only on a subset of pulses whose maximum peak field intensity approaches the threshold intensity from below. We have typically used $\beta = 1$; Unless otherwise specified, I_{thr} was set to $I_{thr} = 0.36 \cdot\ I_{FTL}$.

Figure 9.1 provides an illustrative overview of the numerical process.

9.1.1 Numerical Simulations: Technical Details

We hereby provide some information about the experimental setup of the dynamic alignment numerical simulation:

- In the absence of a laser field, a random phase should yield on average an alignment value of 0.333, due to the isotropic $3D$ space. In the presence of a laser field a random phase typically obtains alignment values around 0.4.

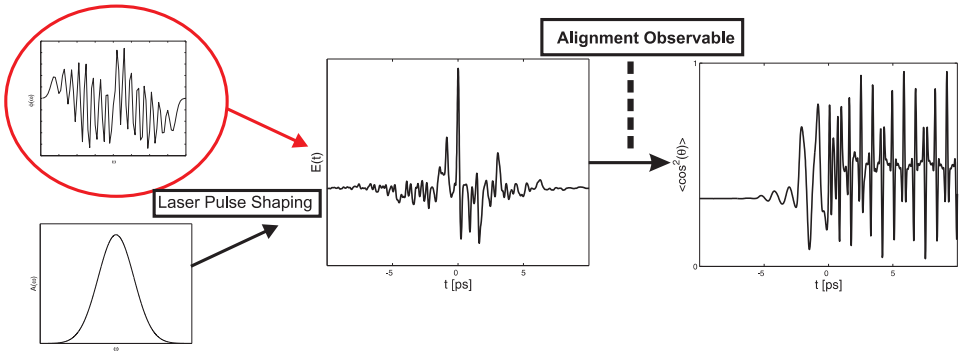

Figure 9.1: An overview of the numerical process. The control function is the phase (circled, top left), the amplitude function is fixed and approximated by a Gaussian (bottom left). The shaping process (Eq. 6.29) generates the electric field, $E(t)$ (center). The "Schrödinger Box" of the alignment observable represents the numerical calculation of the interaction between the electric field with the molecules, based on the quantum dynamics numerical modeling. The revival structure (right) is the observed simulated behavior of the molecules, upon which the yield value is based.

- The penalty term, as introduced in Eq. 9.3 and in Eq. 9.4, can yield fitness values below the value of 0.4. The probability of a randomly generated pulse, with no specific parameterization, to get penalized is extremely low.

- Every fitness evaluation call requires approximately 35s on a single P4-HT 2.6GHz processor.

- Due to the heavy computational cost of a single simulator evaluation, we are limited in granting function evaluations. We are thus encouraged to employ optimization routines with minimal settings. Moreover, we shall apply experiments with a low number of repetitions.

9.2 Experimental Procedure

In order to preliminarily assess the performance of the algorithms on the given problem, we have conducted 10 independent runs for each of the derandomized ES comma-variants with the goal of optimizing the alignment of a sample of generic diatomic molecules undergoing irradiation by a shaped femtosecond laser. We limit each run to $10,000$ function evaluations, due to the computational cost of the simulator.

Algorithm	DR1	DR2	DR3	CMA
AVG-Fitness	0.6399	0.6789	0.6534	0.6261

Table 9.1: Dynamic molecular alignment: Attained fitness values, averaged over 10 runs, for the DES comma variants.

9.2.1 First Numerical Results: Comparison of the Algorithms

Table 9.1 summarizes the numerical results of the runs - the averaged fitness value obtained by each optimization routine. Based on our experience with the problem and the algorithms, the yield differences of Table 9.1 are believed to be significant. Moreover, due to the limited number of simulations we do not provide further statistical analysis of the results.

Roughly speaking, the algorithms were observed to perform equally well, with the exception of the DR2 algorithm that managed to obtain a significantly better optimum than the others. While the DR3 algorithm showed the fastest initial fitness increase, it seemed to get stuck in a sub-optimal local trap after $\approx 2,000$ function evaluations. We have found this behavior to be typical for the DR3 algorithm.
The ranking of the algorithms was qualitatively similar for a number of alignment optimization runs employing different parameter settings.
Figure 9.2 presents the best pulse-shape solution attained, as obtained by the DR2 routine.

9.2.2 The Complete-Basis-Functions Parameterization

In this section we present a new method for learning a function, based on a representation transformation, which can also be referred to as *parameterization*. The so-called *Complete-Basis-Functions Parameterization* was originally derived for the sake of learning the control function of the dynamic alignment problem, i.e. the phase $\phi(\omega)$, but is a general method for learning a generic n-variable function. It can reduce the dimensionality of the search space and possibly boost the convergence speed, respectively, as will be explained in detail.

Appendix B provides the reader with the mathematical background on complete-basis functions, and presents the specific functions that are considered in our study. For the sake of consistency and reading clarity, we specify here our notation for a spanned target function $f(x)$:

$$f(x) = \sum_{k=1}^{K_{max}} c_k \xi_k(x)$$

with c_k as the expansion coefficients, and $\{\xi_k(x)\}_{k=1}^{\infty}$ as the the set of complete-basis functions.

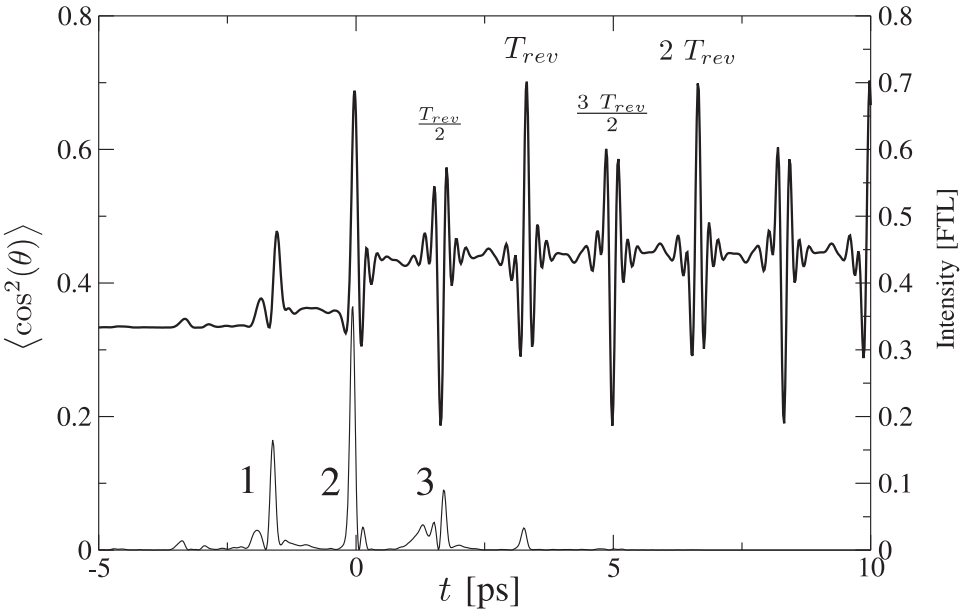

Figure 9.2: Best solution attained by the DR2. Thick line: alignment; thin line: intensity profile of the optimized laser pulse. The solution consists of three main peaks (see labels).

Preliminary: Expanding a Known Function As we will demonstrate here, finding the expansion of a known function by means of a given set of complete-basis-functions, i.e., finding the coefficients of the functions in this basis, is an easy task for a simple evolutionary algorithm, and in particular for the standard-ES. For simplicity, and without loss of generality, let us assume that the task is to approximate a one-variable function using the Fourier series:

$$f(x) = \frac{1}{2}a_0 + \sum_{k=1}^{\infty} a_k \cos\left(\frac{2\pi k}{L} \cdot x\right) + \sum_{k=1}^{\infty} b_k \sin\left(\frac{2\pi k}{L} \cdot x\right)$$

This task can be generalized to functions of higher dimensions, and by using other expansions of complete-basis functions. Following the notation of Appendix B, consider a finite number of the *expansion coefficients* of the cosine and sine functions, $\{a_k\}_{k=0}^{K_a}$, $\{b_k\}_{k=1}^{K_b}$, as the decision parameters to be optimized by the evolutionary search. As a preliminary task in this study, we found that the standard-ES (Schwefel approach) converged easily and quickly to the correct coefficients. This elementary fitting problem was simply defined by means of the *square-error minimization*: The fitness, subject to minimization, was defined respectively as the root-mean-square error function between the original function and its evolving expansion.

Figure 9.3 presents the outcome of learning the *triangle function* with the standard-ES, using only the first 20 frequencies ($K_{max} = K_a + K_b = 40$)

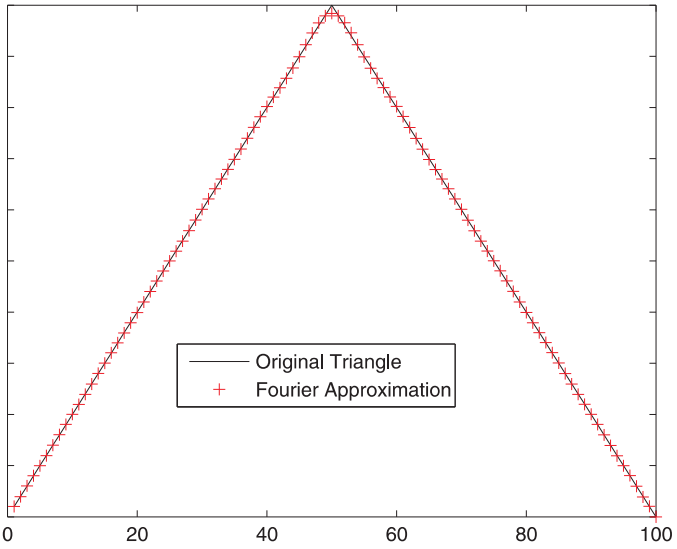

Figure 9.3: Learning the *triangle function* by means of the first 20 *Fourier* frequencies. The plot shows the original triangle function and its Fourier approximation.

of a *Fourier* series as building blocks for a given function discretization of $N = 100$.

Proposed Method: Learning an Unknown Function The idea of spanning a function using a set of complete basis-functions can also be applied for the task of learning an unknown function, represented by N function values, as in our quantum control alignment problem. The inspiration for this method was the initial intuition to the alignment problem, which suggested that the control function should be periodic. Motivated by this intuition, we started to run simulations in which an ES was aiming at learning $\phi(\omega)$ using the *harmonic functions* as building blocks. Rather than learning the interpolated values of the control function, the coefficients of the harmonics (Fourier components) were optimized. Following the success of those experiments, we extended the method to other sets of complete basis functions, and in particular to the sets of functions which are introduced in Appendix B: The *Legendre Polynomials*, the *Bessel Functions*, the *Hermite Polynomials*, and the *Chebyshev* polynomials.
Assuming that the desired discretization is up to a resolution of N points in the interval, we limit the number of elements in the expansion series to K_{max}, where preferably $K_{max} \ll N$. By that we can achieve a dramatic dimensionality reduction of the search space, aiming to boost the convergence speed. The idea is then to apply an evolutionary search to the $n = K_{max}$

coefficients of the expansion functions, where a simple transformation is applied for every fitness evaluation. In practice, the required time for additional computation of this transformation is negligible with respect to the objective function evaluation, in most real-world problems.

An ES employing a Fourier auxiliary function has been proposed in the past, known as the FES method [161]. The FES aims at approximating the fitness landscape, and particularly its small attraction basins, by means of the Fourier series. However, the careful reader should notice that our method is based on a different principle. It uses complete-basis functions as a transformation of the decision parameters themselves, rather than the fitness landscape, which is left untouched. It strongly relies on the fact that these decision parameters represent a continuous function - and this function is due to be approximated.

Preliminary Calculations

Quadratic Phase Functions: The α-Test Since we are about to investigate representations of low-order polynomials, we would first like to address the question whether there exists a trivial extremum which would become a local trap for such phase functions. Hence, we calculated the fitness of constructed quadratic phase functions, centered around the central frequency. Explicitly, we considered the following family of constructed phases:

$$\phi_\alpha(\omega) = \alpha \cdot (\omega - \omega_{central})^2, \tag{9.5}$$

where the continuous parameter α is scanned systematically in the interval $[0, 15]$. Note that these phases are constructed over $n = 80$ function values, and given as input to the dynamic alignment simulator as before.

The results of this so-called *α-test* are presented in Figure 9.4.

As can be clearly seen in the given plot, most of the quadratic phase functions attain extremely low fitness values, due to large penalty terms, and they never exceed the fitness value of 0.45. This eliminates the existence of a trivial quadratic solution for the problem.

The Initial States Density Test We set the number of terms in each expansion to $K_{max} = 40$. The following preliminary experiment is meant to compare the natural initial quality of the different parameterizations with respect to the alignment problem. We applied a so-called *initial states density test*, a statistical fitness measurement of the initialized phase functions in the different parameterizations. For each parameterization in use, i.e., the direct/plain 80-dimensional random phase vector, or the random 40-dimensional coefficient vector for the various polynomials in use, we initialized $1,000$ phase functions and calculated their mean fitness and standard deviation. The numerical results are visualized as histograms in Figures 9.5-

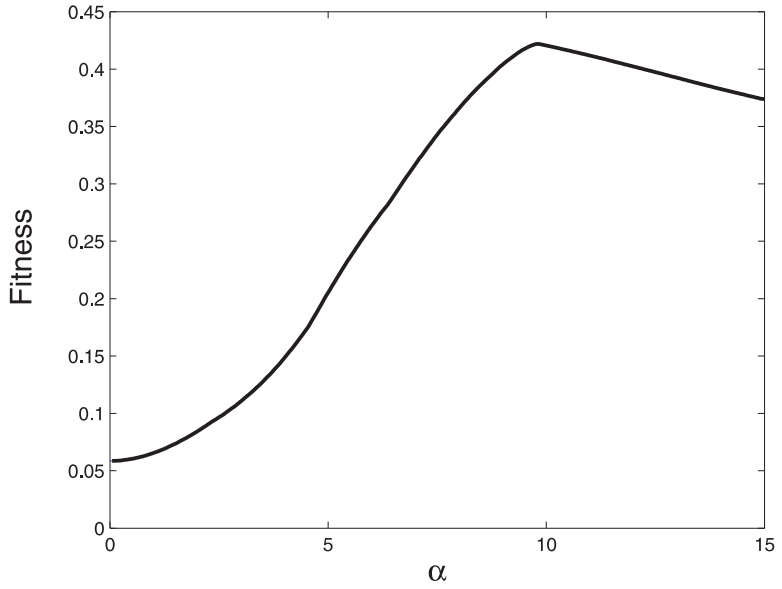

Figure 9.4: The α-test: The fitness of quadratic phase functions, centered around the central frequency, as defined in Eq. 9.5.

Table 9.2: Parameterizations: Averaged Performance

Routine	Direct		Fourier		Legendre	
	Avg. Fit.	0.6 Eval	Avg. Fit.	0.6 Eval	Avg. Fit.	0.6 Eval
(1,10)-DR2	0.6789	2325	0.4494	N.A.	0.6384	629
(1,10)-CMA	0.4676	N.A.	0.4542	N.A.	0.6409	515.1
(μ, λ)-CMA	0.6261	4962.5	0.6171	4475.8	0.6466	194.5
Routine	Bessel		Hermite		Chebyshev	
	Avg. Fit.	0.6 Eval	Avg. Fit.	0.6 Eval	Avg. Fit.	0.6 Eval
(1,10)-DR2	0.6299	1390	0.5944	5610	0.4843	N.A.
(1,10)-CMA	0.6229	2212.9	0.6755	271	0.4979	N.A.
(μ, λ)-CMA	0.6232	2719.5	0.6843	118	0.6225	3770.8

9.10, providing the fitness distributions of the various random initializations. See further discussion below.

Parameterizations: Numerical Results

In this section we present the numerical results for optimizing the dynamic alignment problem with the different parameterizations - the direct/plain parameterization versus the polynomial-based parameterizations with $K_{max} = 40$ terms. Our runs were based on the following algorithmic kernels:

1. $(1, 10)$-DR2

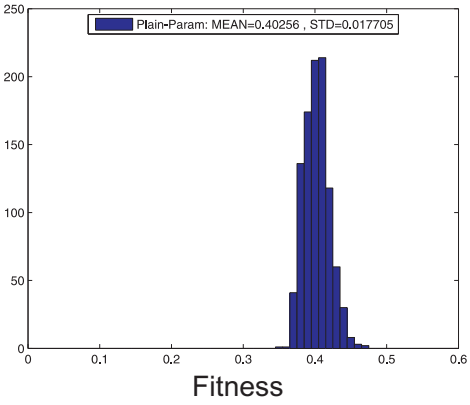

Figure 9.5: Initial states density test for **direct** parameterization.

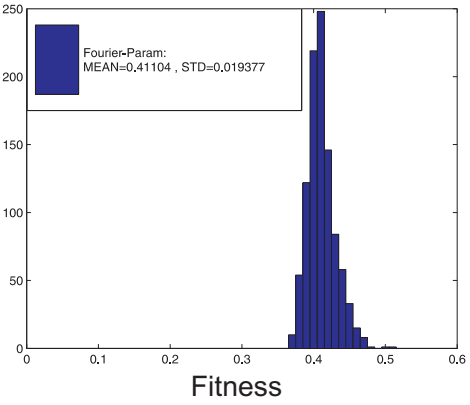

Figure 9.6: Initial states density test for **Fourier** parameterization.

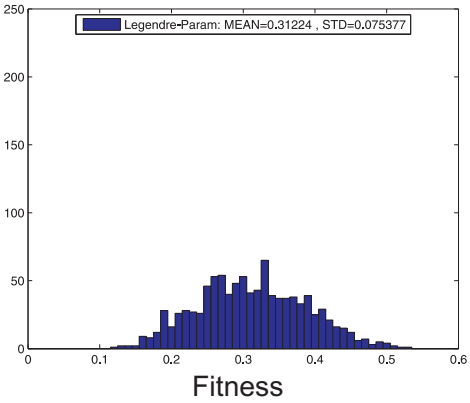

Figure 9.7: Initial states density test for **Legendre** parameterization.

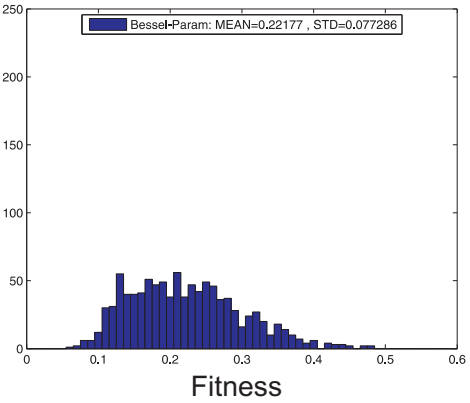

Figure 9.8: Initial states density test for **Bessel** parameterization.

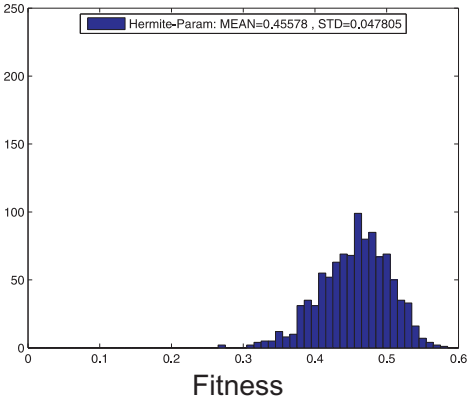

Figure 9.9: Initial states density test for **Hermite** parameterization.

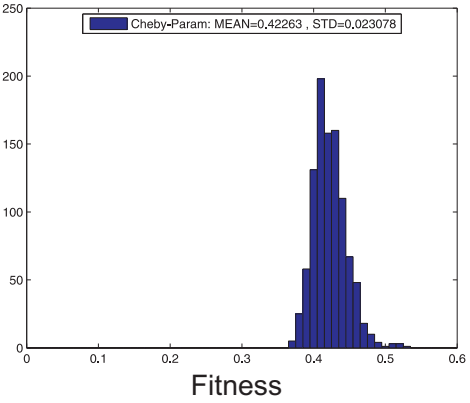

Figure 9.10: Initial states density test for **Chebyshev** parameterization.

Table 9.3: Parameterizations: Summary of Best Results

Parameterization	Best Fitness	0.6 Eval	Routine	Initial States Density
Direct-Param	0.6899	2310	(1,10)-DR2	0.4026 ± 0.018
Fourier	0.6526	1411	(7,15)-CMA	0.4110 ± 0.019
Legendre	0.6487	106	(7,15)-CMA	0.3122 ± 0.075
Bessel	0.6457	61	(7,15)-CMA	0.2218 ± 0.077
Hermite	0.6866	31	(7,15)-CMA	0.4558 ± 0.048
Chebyshev	0.6490	1051	(7,15)-CMA	0.4226 ± 0.023

2. $(1, 10)$-CMA

3. (μ_W, λ)-CMA: Following the recommended settings (Eq. 1.47): $(7, 15)$ for $n = 40$, versus $(8, 17)$ for $n = 80$.

The runs were limited to $10,000$ function evaluations. We conducted 10 runs per method.
We consider the performance criteria of the various methods as the following:

- The mean fitness values per method over the 10 runs.
- The averaged number of evaluations per method until the fitness value of 0.6 was reached during the runs. We consider the yield value of 0.6 as the lower bound of the regime of good solutions.
- The results of the *initial states density test*, as was introduced earlier: The averaged initial fitness values per method, with the standard deviation.

We provide a table of results, which consists of the numerical values of the specified performance criteria per method. It is given as Table 9.2. Table 9.3 summarizes the best results obtained per parameterization.

Analysis and Discussion

An important result that should be pointed out is that **all** the runs in the various parameterizations have converged into a highly fit phase function with at least one optimization routine, i.e., all the given complete-basis functions are capable of spanning a good phase function with $K_{max} = 40$ terms.

Furthermore, we would like to analyze shortly the experimental results of the various parameterizations with respect to the dynamic alignment optimization, as presented in Tables 9.2 and 9.3:

1. **Initial State** The *Hermite* parameterization has clearly the most natural **initial** representation for the phase function for the given problem, among the various cases, as reflected from the *initial states density*

test results (Figures 9.5-9.10 and Table 9.3). Note that the *Legendre* as well as the *Bessel* parameterizations have low initial fitness values, even below the direct parameterization, due to the penalty effect. It should be stressed that the standard deviations of the different fitness distributions are reasonably low.

2. **Fitness Values** The *Hermite* parameterization obtained fitness values as high as the direct parameterization method, though by means of a different algorithm, as will be discussed shortly. As far as we know, the attained yield values in the regime of ≈ 0.69 are the highest *cosine-squared alignment* values which were ever attained for this particular configuration of the problem. Hence, from the optimization perspective, the proposed parameterization does not hamper the feasibility to obtain the maximally-attained yield within the limit of function evaluations.

3. **DR2 vs. CMA** There is a clear trend regarding the two algorithmic kernels. The DR2 obtained the best results for the direct parameterization, but obviously failed to deliver reasonable results for the polynomial-based parameterizations. In most cases, the DR2 does not even converge. The $(7, 15)$-CMA, on the other hand, performed very well with the various polynomial-based parameterizations, and attained fine results also for the direct parameterization. The $(1, 10)$-CMA is clearly inferior with respect to its rank-μ weighted-recombined sibling. Our proposed explanation for this trend is the strong correlations between the polynomials' coefficients, which make the covariance matrix an essential component for successful optimization. On the other hand, it seems that the covariance matrix is not an essential component for the direct parameterization, and may even introduce a barrier, to some degree, to the global search.

   We would like to link this to the conclusions drawn for the QC landscapes of Two-Photon Processes in Chapter 7, where QC landscape analysis stating that first-order information is sufficient for optimizing QC landscapes was experimentally corroborated. The fact that the DR2 algorithm performs so well on the current dynamic alignment landscape, which is a combined OCT/OCE landscape, could be considered as an additional corroboration to this QC landscape analysis.

   We shall further explore the performance of the DR2 versus CMA-ES with respect to the direct versus Hermite parameterizations in Section 9.3.

4. **Boosting Convergence Speed** An immediate conclusion from both tables is that the proposed method achieved a significant boost of the convergence speed for all the different polynomial-based parameteri-

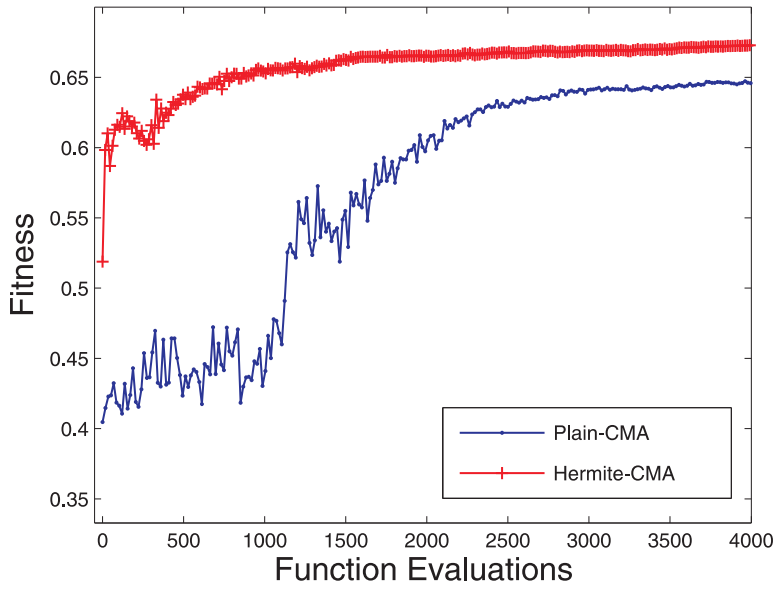

Figure 9.11: The speeding-up effect: Typical convergence profiles of the (μ_W, λ) CMA-ES for the Hermite versus the direct parameterizations.

zations, in comparison to the direct parameterization. The *Hermite* parameterization with the $(7, 15)$-CMA is clearly the fastest routine, and it outperformed the other routines by far. It should be noted that the *Legendre* as well as the *Bessel* parameterizations, which have the lowest initial yield values, manage to compensate for that and reach the regime of good solutions (yield > 0.6) rather quickly.

Typical convergence profiles for Hermite versus direct parameterizations are plotted in Figure 9.11.

5. **Physics Interpretation** Aiming at gaining physics insights into the nature of highly-fit phase functions with respect to the alignment problem, we examined the nature of good solutions in the different parameterizations. The idea was to calculate the distributions of the coefficients, and try to identify dominance of certain components (frequencies in the Fourier case). Unfortunately, such dominance could not be identified within the results. The set of attained optimal phases reveals high complexity, which could not be tackled. This provides us with the motivation to explore a simpler variant of the alignment problem in Section 9.3.

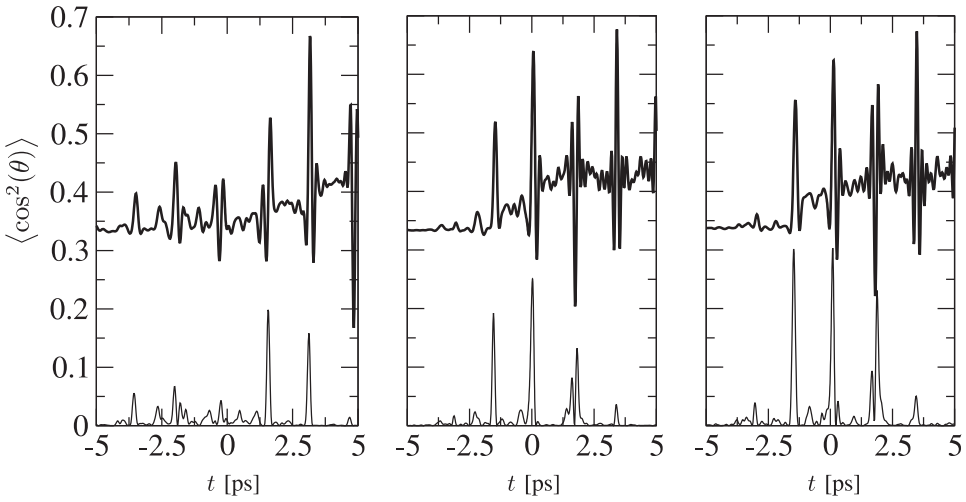

Figure 9.12: Optimized pulses and alignment for $I_{thr} = 0.2 \cdot I_{FTL}$, $I_{thr} = 0.25 \cdot I_{FTL}$ and $I_{thr} = 0.3 \cdot I_{FTL}$. Figure courtesy of Christian Siedschlag [162].

Intensity $[I_{FTL}]$	0.2	0.25	0.3	0.36
$\langle\cos^2(\theta)\rangle$	0.662	0.673	0.6734	0.689

Table 9.4: Best $\langle\cos^2(\theta)\rangle$ values obtained with the DR2 algorithm over five runs for different values of I_{thr} [162].

9.2.3 Further Investigation

We would like to review here briefly additional calculations for this alignment problem, which were carried out by Siedschlag and Vrakking (see, e.g., [162]).

Penalty Strength By decreasing I_{thr}, the search algorithm was shown to look for effective pulses with less available peak intensity. The numerical results of additional optimization runs, carried out by the DR2 algorithm, for $I_{thr} = 0.2 \cdot I_{FTL}$, $I_{thr} = 0.25 \cdot I_{FTL}$ and $I_{thr} = 0.3 \cdot I_{FTL}$ are presented in Table 9.4. Overall, the evolutionary search was able to make up for the smaller peak intensities by redistributing the fluence in a clever way, so to speak. The optimized pulse-shapes for the three lower threshold intensities are presented in Figure 9.12. The three solutions are observed to be remarkably similar.

Constructed Pulse Trains Siedschlag and Vrakking [162] also treated the question whether a simple train of pulses that is constructed by an appropriately designed phase function yields results that are comparable to those achieved by the evolutionary approach. In particular, the question addressed *trains of pulses*, which are generated by oscillatory phase functions.

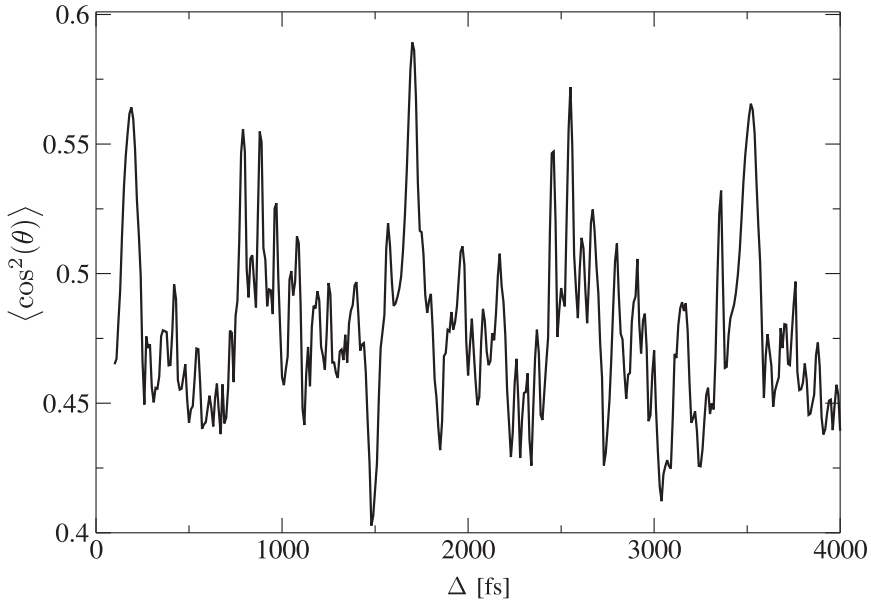

Figure 9.13: A cut through the contourplot of Figure A.7 for $A = 2.26$, for which the largest alignment $\left(\langle\cos^2(\theta)\rangle = 0.589\right)$ in the two-parameter approach under the condition $I < 0.36 \cdot I_{FTL}$ was achieved [162]. Figure courtesy of Christian Siedschlag.

Explicitly, the following family of phases was considered:

$$\phi_{osc}(\omega) = A \cdot \sin(\omega\Delta + \alpha) \tag{9.6}$$

The two relevant parameters, A and Δ, were scanned in a search for the pulse that would produce the best alignment; Figure A.7 presents the outcome of that scan. The magnitude of A controls the distribution of the available intensity over the peaks in the pulse train (and hence the peak intensity with respect to the FTL solution), while Δ corresponds directly to the time delay between two consecutive peaks. Note that the maximally obtained alignment yield in this scan was $A \approx 0.68$ and $\Delta = 1.7$ps; However, its corresponding peak intensity was too high for the model, i.e., $I > 0.36 \cdot I_{FTL}$.

Figure 9.13 presents a cut of the contourplot scan of Figure A.7, at the maximally obtained yield in the *allowed range* (0.589). It was concluded in [162] that this approach was not flexible enough to adapt to the finer details of the time-dependent alignment response.

9.3 Investigation of Optimality: Zero Kelvin

Here we focus in a simplified variant of the original alignment problem, at zero temperature ($T = 0\ K$) and with only a single rotational level at the

initial distribution. The numerical modeling of Eq. 8.8 considers now $M = 0$ and reads:

$$|\Psi(t)\rangle = \sum_{J=0}^{N_{rot}} \alpha_J^{(g)}(t)\,|gJ\rangle + \exp\left(-i\omega_0 t\right)\alpha_J^{(e)}(t)\,|eJ\rangle \tag{9.7}$$

The motivation for this simplification is to allow studying the physical characteristics of the optimal solutions, which would not have been possible for the general case, e.g., tracking the time-dependent population of the rotational levels, given only the ground level at initialization. From the technical perspective, this simplification reduces the simulator evaluation time to approximately 5s on a single P4-HT $2.6GHz$ processor.

We carried out calculations optimizing field-free molecular alignment starting from $J = 0$ for a number of algorithmic approaches and various Rabi peak frequencies. In each case, the same calculation was attempted by means of 20 runs. Each run was limited to $20,000$ function evaluations. We restrict the discussion in this section to the best results obtained in each series of 20 trials.

Figure 9.14 presents a comparison between one optimization of dynamic alignment starting from $J = 0$, performed using the DR2 algorithm under perturbative conditions $\left(\Omega_{ge} = 40 \times 10^{12}\mathrm{s}^{-1}\right)$ and four optimizations performed under non-perturbative conditions $\left(\Omega_{ge} = 160 \times 10^{12}\mathrm{s}^{-1}\right)$ using both the DR2 and the CMA algorithms, with either a direct/plain parameterization of the phase or with the Hermite parameterization, employing the first $K_{max} = 40$ Hermite polynomials. Furthermore, based on our previous observations in this chapter, we employed $(1, 10)$-DR2 or $\{(7, 15), (8, 17)\}$-CMA (the latter depends on the parameterization used).

The obtained result at low laser intensity $\left(\Omega_{ge} = 40 \times 10^{12}\mathrm{s}^{-1}\right)$ is simple: A pulse train is observed where the spacing between the peaks is approximately the rotational period of a coherent superposition state consisting of $J = 0$ and $J = 2$ only $\left(T_{rev_{02}} = \frac{1}{6B_{rot}c} = 1.1\mathrm{ps}\right)$. The time-dependent intensity is given by a train of pulses where the largest pulse reaches an intensity of $0.36 \cdot I_{FTL}$.

The obtained pulse-shapes at high laser intensity $(\Omega_{ge} = 160 \times 10^{12}\mathrm{s}^{-1})$, are considerably more complex and no simple periodicity can be observed. The averaged as well as largest values of $\left\langle \cos^2(\theta) \right\rangle$ attained are shown in Table 9.5.

In consistency with the numerical results of the previous section, the highest alignment yield values attained for this particular system were also obtained by the DR2 with plain parameterization as well as by the CMA with Hermite parameterization. Employing the CMA with plain parameterization or the DR2 algorithm with the Hermite parameterization yields a slightly lower values over 20 trials. Based on our experience with the problem and the algorithms, the yield differences of Table 9.5 are believed to be significant.

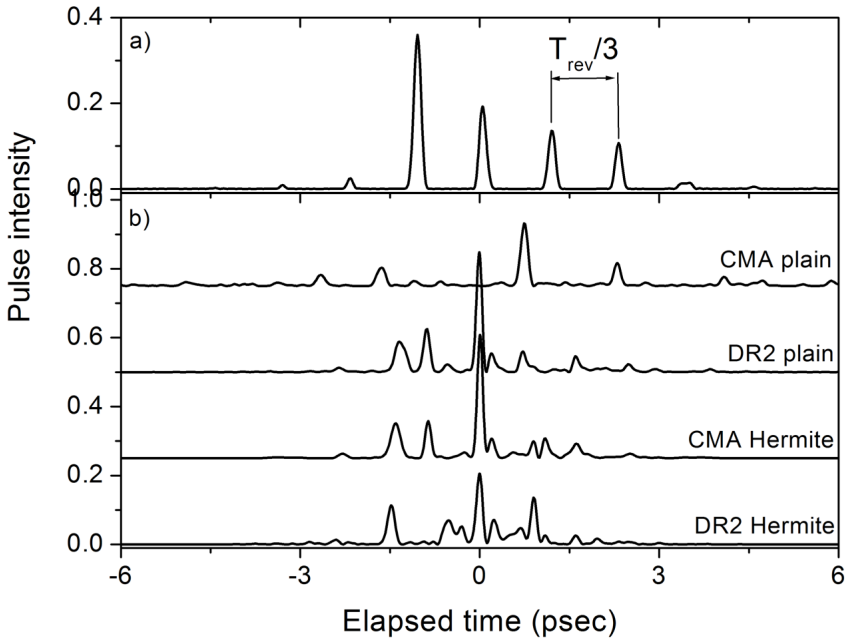

Figure 9.14: **(a)** Comparison of an optimization performed employing the DR2 algorithm with $\Omega_{ge} = 40 \times 10^{12}\mathrm{s}^{-1}$ and **(b)** Four calculations with $\Omega_{ge} = 160 \times 10^{12}\mathrm{s}^{-1}$ employing the DR2 and the CMA algorithms with either a plain or Hermite parameterizations of the control phase function.

	$(1,10)$-DR2	$\{(7,15),(8,17)\}$-CMA
Plain Param.	0.9559 ± 0.007 (0.9622)	0.9413 ± 0.006 (0.9508)
Hermite Param.	0.9501 ± 0.004 (0.9570)	0.9583 ± 0.003 (0.9618)

Table 9.5: Maximizing the cosine-squared field-free molecular alignment starting from $J = 0$ ($T = 0K$) at $\Omega_{ge} = 160 \times 10^{12}\mathrm{s}^{-1}$ over 20 runs with $20,000$ function evaluations per run; Mean and standard-deviation values are given, with the maximal value obtained in brackets.

This is supported by inspection of the pulse shapes shown in Figure 9.14. The two most successful optimizations (CMA/Hermite and DR2/Plain) not only share their yield value of $\langle \cos^2(\theta) \rangle$, but furthermore make use of a pulse shape that is very similar.

9.3.1 Conceptual Quantum Structures

The time-dependent population of the rotational levels can be analyzed in a fairly simple technique, known as the *Sliding Window Fourier Transform* (SWFT), which provides us with a **powerful visual tool**. Given the revival structure of an obtained solution, a sliding time window is Fourier transformed, to produce the frequency picture through the alignment process. This windowing creates a transformation which is localized in time. Due to the *quantization* of the rotational levels, only certain frequencies (or *energy* levels, respectively) are expected to appear.

We applied the SWFT routine to the optimal solutions which were found in the various runs under non-perturbative conditions. Figures A.10, A.11, A.12 and A.13 visualize the typical population process of the rotational levels for four typical solutions of the different optimization procedures (2 parameterizations *times* 2 DES variants). The observed quantum energy levels are indeed as expected from theory.

The results reveal two different conceptual quantum structures, which correspond to optimal and sub-optimal solutions in terms of the alignment yield. The plain-DR2 as well as the Hermite-CMA procedures obtain the best solutions, which share the same structure - they are characterized by the dominant population of the 4^{th} rotational level in the SWFT picture, corresponding to $J = 6$. On the other hand, the plain-CMA and Hermite-DR2 procedures obtain solutions with lower yield, which are characterized by a gradually increasing population of the rotational levels.

The original *revival structures* for two obtained solutions, representing the two conceptual structures, are given in Figures A.8 and A.9. The *optimal* family of solutions (Figure A.8) possesses a dramatic revival structure, with a typical strong pulse in the train which lies on the boundary of the punished regime ($I \approx 0.36 \cdot I_{FTL}$). This strong pulse seems to be essential in giving the molecules the right 'kick', and most likely responsible for the dominant population of the 4^{th} rotational level in the SWFT picture ($J = 6$). The *sub-optimal* family of solutions (Figure A.9) possesses a revival structure with a smooth exponential envelope, and thus has a gradual building-up of the rotational levels in the SWFT picture, respectively. It typically contains a train of medium pulses and lacks a dominant one.

We would like to emphasize the fact that we obtained the same family of optimal solutions, representing a single Quantum structure, from two different optimization approaches: The first employs a first-order DES subject to direct pixelation of the control phase, while the other employs a second-order

DES subject to Hermite expansion of the control phase.

9.3.2 Maximally Attained Yield

While this does not constitute a proof, we speculate that within the constraints in the optimization (i.e., the finite pulse bandwidth and energy, as well as the finite resolution of the phase function), both algorithms have found a solution that approaches the best solution that is possible. However, even if the solutions are optimal within the constraints set by the laser bandwidth, the laser pulse energy and the parameterization of the phase, it is clear that the solutions do not approach the maximum alignment that can be supported by the basis of $N_{rot} = 20$ rotational states (see Eq. 9.7) that were used in the calculation. The maximum alignment supported by this basis is the largest eigenvalue of the observable matrix, which was found to be 0.9863. The corresponding eigenvector will be referred to here as the *maximal eigenvector* or the *maximal wavepacket*.

We ascribe the difference between this maximum value and the values obtained in the optimizations as being largely due to the finite laser bandwidth in our calculations. The bandwidth and the pulse duration of a laser pulse with a Gaussian shape are related by Eq. 6.30, where the spectral amplitude parameter reads $c_B = 0.441$. Thus, for a pulse with a 100fs Fourier-limited duration, the bandwidth is $\Delta\omega_{laser,FWHM} = 0.0182eV = 147\text{cm}^{-1}$. When a molecule undergoes a Raman transition from $J = J_0$ to $J = J_0 + 2$, the energy absorbed from the laser field is $B_{rot} \cdot (4J_0 + 6)$. This absorbed energy is the difference between the pump- and dump-photons involved in the Raman excitation. Consequently, the Raman excitation becomes frustrated when $B_{rot} \cdot (4J_0 + 6) > \Delta\omega_{laser,FWHM}$. In our case, with a rotational constant of $B_{rot} = 5\text{cm}^{-1}$, this threshold occurs for $J_0 \approx 6$.

As Figure 9.15 shows, the rotational wave packet that displays the largest alignment after the optimization contains only limited contributions from $J = 8$ and $J = 10$, and none from rotational levels above $J = 10$. By contrast, the *maximal wavepacket* contains contributions all the way up to $J = 18$. In this respect, it may appear to be surprising that a high yield of 0.962 can be obtained when the optimized wavepacket differs so much from the maximal wavepacket. In order to assess the crucial influence of the bandwidth constraint on the cut-off of accessible J values, additional calculations were performed with the original bandwidth doubled, while the fluence was kept fixed (thus corresponding to a 50fs pulse with $\Omega_{ge} = 226 \times 10^{12}\text{s}^{-1}$). These results are also presented in Figure 9.15 as a reference to the calculations with the original bandwidth. The doubling of the bandwidth permitted populating up to $J = 12$, and thus produced an enhanced alignment yield of 0.975. Note that the distinction between the two families of solutions, corresponding to the two algorithmic classes, as discussed in Section 9.3.1, can be clearly observed in Figure 9.15.

The difference between the maximal wavepacket and optimized wavepacket is also reflected in the *angular probability distribution functions*, as presented in Figure 9.16. These probability distribution functions are respectively constructed from the coefficients of the maximal eigenvector as well as the state obtained from the optimized field, based on Eq. 9.7. Even though at the higher bandwidth the discrepancy between the optimally controlled distribution function and the maximally attainable limit appears to be significant, a high alignment value was still obtained.

The explanation for this excellent behavior, despite considerable differences in the composition of the wavefunction, lies in the *variational principle* (see, e.g., [126]), which states that a first order error in a trial wavefunction (i.e., the wavepacket from the bandwidth limited optimal control field) will produce an extremum eigenvalue (i.e., alignment yield) of second-order error:

$$\frac{\langle\psi|\,\mathcal{H}\,|\psi\rangle}{\langle\psi|\psi\rangle} = \frac{E_n + \langle\delta|\,\mathcal{H}\,|\delta\rangle}{\langle n|n\rangle + \langle\delta|\delta\rangle} = E_n + \mathcal{O}\left(\delta^2\right) \tag{9.8}$$

9.3.3 Another Perspective to Optimality: Phasing-Up

When a molecule is exposed to a shaped, intense laser pulse the optimization has to accomplish two things. First, the optimization has to create a wavepacket consisting of a large number of rotational states that can serve to align the molecule. Second, the optimization has to prepare the wavepacket with the correct phase relationship between the component wavefunctions, so that during its field-free evolution these components would coherently add-up to generate an optimally aligned wavefunction. While there is no criterium available that allows us to ascertain whether the algorithm has optimized the population distribution, it is possible to investigate the phase relationship of the component wavefunctions in the optimized solutions. Maximum alignment occurs if at some point in time the phases of all component wavefunctions differ from each other by 0 (modulo 2π).

Explicitly, given a wavefunction,

$$\psi = \sum_j a_j^{(t)} \cdot |j\rangle \cdot \exp\left(-i\frac{E_j t}{\hbar}\right),$$

the coefficients $a_j^{(t)}$ are complex numbers, and as such can be expressed in their *polar* representation:

$$a_j^{(t)} = r_j^{(t)} \cdot \exp\left(i\varphi_j^{(t)}\right). \tag{9.9}$$

We thus question whether given a certain population - does the optimization routine produce the optimal set of phases $\varphi_j^{(t)}$? In order to answer this question, a simple optimization procedure was implemented in the following manner: It accepts the $a_j^{(t)}$ as input, and aims at optimizing the phases $\varphi_j^{(t)}$

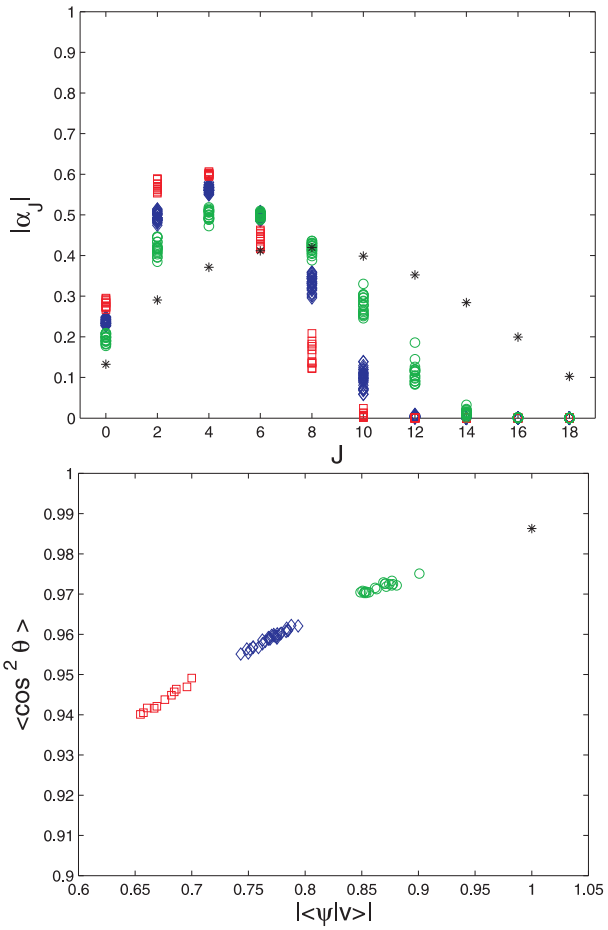

Figure 9.15: TOP: The distribution of the maximal and the best optimized wavepackets over the rotational states. Stars represent the maximal wavepacket in the finite rotational basis (i.e., corresponding to the highest-ranked eigenvector of the observable matrix). Diamonds represent the 1^{st} optimized set of solutions (CMA-Hermite / DR2-Plain), and Squares represent the 2^{nd} optimized set of solutions (CMA-Plain / DR2-Hermite); Circles represent calculations with doubled bandwidth and the same fluence (50fs pulse with $\Omega_{ge} = 226 \times 10^{12}\text{s}^{-1}$), optimized by the DR2 subject to plain parameterization. The figure clearly shows that the limited field bandwidth cuts off the rotational states for the optimized solutions after $J = 10$, when the original bandwidth is used, or after $J = 12$ when the bandwidth is doubled. Furthermore, this plot **illustrates the distinction between the two families of solutions for the original bandwidth** (i.e., Diamonds versus Squares) arising from the different algorithmic approaches. BOTTOM: The alignment as a function of the overlap of the optimized wavepackets $|\Psi\rangle$ with the maximal eigenvector $|V\rangle$. Note that the overlap for the original bandwidth never exceeds 0.8 in magnitude. Also note the **three clusters for the families of algorithmic solutions**.

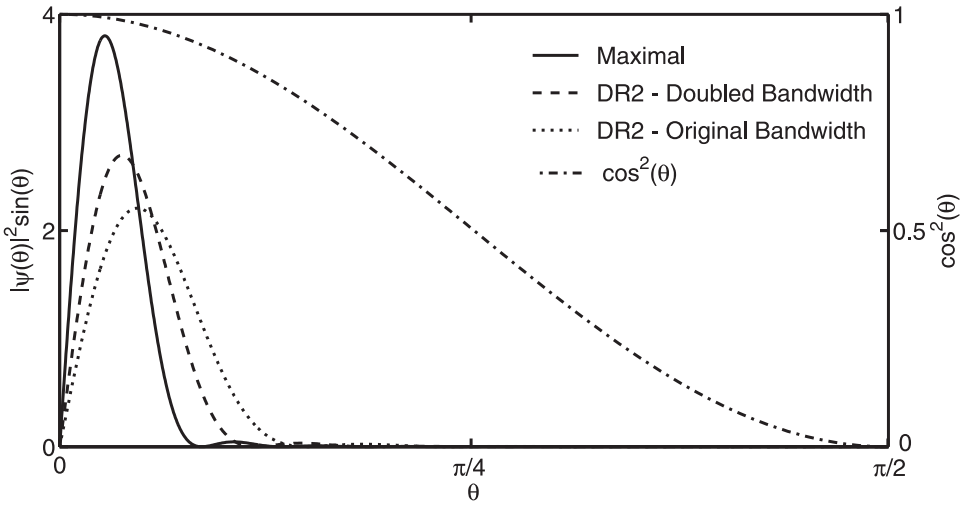

Figure 9.16: Left axis: Normalized angular probability distribution function for the maximal case $|\psi_{max}(\theta)|^2 \sin(\theta)$, and the optimized control function $|\psi_{opt}(\theta)|^2 \sin(\theta)$. Right axis: The value of $\cos^2(\theta)$. The constraints prohibit the evolutionary algorithm from attaining the absolute maximal angular probability distribution function; However, the expectation value of the observable $\langle \cos^2(\theta) \rangle_{opt} = 0.9621$ when using the original bandwidth corresponding to a 100fs Fourier-limited pulse is within 0.025 of the maximum attainable value $\langle \cos^2(\theta) \rangle_{max} = 0.9863$. When doubling the bandwidth (i.e., basing the shaped laser pulse on a 50fs Fourier-limited pulse) $\langle \cos^2(\theta) \rangle_{opt}$ increases to 0.975, which is only 0.0113 away from the maximum attainable value.

such that the cosine-squared alignment is maximized. Practically, it uses a subroutine from the general alignment code for the evaluation, and applies the CMA algorithm for the tuning of the 10 relevant phases. Note that a single function evaluation has the duration of ≈ 0.5s.

We considered 50 different cases of high-quality solutions to the alignment problem (all solutions have cosine-squared-alignment values in the regime of 0.95) - for each test case 100 independent optimizations were run, aiming to tune the phases.

The experimental results are clear and sharp. They are presented at two levels:

1. In all 100 runs for all 50 test-cases - the best solution has always **synchronized phases**. There are different phase values per run, but it does not make a difference for the cosine-squared alignment, as long as the populated levels hold that same phase value. Explicitly, the Sigma-RMS of the phases was calculated:

$$\Delta\varphi^{optimal} = 0.0117$$

2. The 50 test-cases, as originally obtained by the original optimization prior to this optimization procedure, held phases which were not far from being synchronized,

$$\Delta\varphi^{DR2} = 0.0566,$$

and indeed, the optimizations did not improve the cosine-squared alignment dramatically: Always less than 1% improvement was recorded.

We consider this a very strong result - the evolutionary optimization routine managed to tackle the fine-tuning of the quantum control problem, behind the complex transformations and the so-called Schrödinger black-box.

To summarize, while we cannot establish whether the optimization has distributed the population in the best possible way, we do observe that the algorithm has properly phased-up all component wavefunctions with respect to each other. This type of coherent alignment of phases was also observed to be optimal in the mechanistic analysis of another state-to-state control application [163].

9.4 Evolution of Pulses under Dynamic Intensity

Our observation so far regarding the alignment problem, and in particular concerning its zero-Kelvin variant in the previous section, provides us with the motivation to investigate optimized pulse structures that obtain high alignment yield at different laser intensities, and especially their evolution subject to a *slowly-varying laser intensity*. This section is a direct experimental continuation to Section 9.3, considering solely the zero-Kelvin alignment variant with two specific algorithmic approaches that were employed for its optimization: the DR2-plain and CMA-Hermite procedures.

9.4.1 Evolutionary Algorithms in Dynamic Environments

From the algorithmic perspective, the optimization framework becomes now an evolutionary search subject to a dynamic environment [71].

Evolutionary Algorithms are natural candidates for optimization in dynamic environments, due to the straightforward analogy with *organic evolution*, which occurs in a continuously varying environment. Typical approaches for dynamic environments include the promotion of diversity, the use of multi-populations, the introduction of memory-based components, or the assignment of so-called *scouts* that maintain information about the search space.

Evolution Strategies are a particularly good choice, for their built-in mutative self-adaptation mechanism. The standard-ES has been demonstrated to perform well under a dynamic environment of a **time-varying sphere model** ("a landscape with *catastrophes*"), using a comma strategy and with

no recombination (see, e.g., [164]). The mutative self-adaptation mechanism played a crucial role, in allowing a rapid adjustment of the evolving individuals to the time-dependent location of the global maximum: The optimal mutation strategy parameters were learned successfully, without exogenous control. Other empirical studies extended this model to continuously moving peaks, and reported on satisfying adaptation of the standard-ES [165]. Arnold and Beyer considered specific derandomized Evolution Strategies, and showed theoretically that the step-size adaptation mechanism works perfectly well on a moving-sphere problem [166]. In light of these findings, we find our candidate derandomized ES variants perfectly suited for the current optimization task.

9.4.2 Dynamic Intensity Environment: Procedure

In order to observe, and possibly understand how the optimal laser pulse shape evolves from the simple pulse train obtained for $\Omega_{ge} = 40 \times 10^{12}\mathrm{s}^{-1}$ (Figure 9.14 (a)), into a much more complicated pulse-shape for $\Omega_{ge} = 160 \times 10^{12}\mathrm{s}^{-1}$ (Figure 9.14 (b)), a series of calculations were conducted where Ω_{ge} was **increased linearly as a function of the generation number**. In these calculations, the molecule was initially exposed to a shaped laser field with $\Omega_{ge} = 40 \times 10^{12}\mathrm{s}^{-1}$, and over $10,000$ generations this value linearly increased to $\Omega_{ge} = 180 \times 10^{12}\mathrm{s}^{-1}$. This was immediately followed by a linear decrease of the intensity over additional $10,000$ generations, back to the initial value of $\Omega_{ge} = 40 \times 10^{12}\mathrm{s}^{-1}$. Note that a generation involves 10 or 15 function evaluations, for the DR2-plain or CMA-Hermite procedures, respectively. Furthermore, we consider two control resolutions for the plain-parameterization, $n_1 = 80$ versus $n_2 = 160$, in order to test the algorithmic performance in these two search space dimensions.

The analysis of the dynamic intensity environment is discussed next at several levels.

Intensity Milestones: Dynamic vs. Static Optimization

Figure 9.17 presents the best evolution runs of the DR2-plain optimization procedure for $n_1 = 80$ and $n_2 = 160$ pixels, respectively. It contains four curves, which correspond to the evolution progress in the ramped-up and ramped-down laser intensity environments of the two different runs. Note that the ramped-down curves of the two runs merge. The ramped-up curves differ significantly in the initial learning periods, due to the different search space dimensionality, as expected.

Following the initial learning period of the optimization procedure, a smooth increase is observed in the alignment yield $\langle \cos^2(\theta) \rangle$, as a function of the laser intensity. The best $\langle \cos^2(\theta) \rangle$ value, as reported in the static high intensity case (Table 9.5), is successfully recovered: A $\langle \cos^2(\theta) \rangle$ value

of 0.962 was obtained at $\Omega_{ge} = 160 \times 10^{12}\mathrm{s}^{-1}$. Thus, the dynamic environment does not hamper the optimization performance given a desired target intensity, as long as the initial learning period is passed.

Figure 9.18 presents a comparison between the pulse-shape attained by the DR2 during a dynamic-intensity run at the milestone of $\Omega_{ge} = 160 \times 10^{12}\mathrm{s}^{-1}$, to the equivalent optimized pulse-shape attained in the static optimization procedure at the same Rabi frequency milestone, previously shown in Figure 9.14. Several conclusions may be drawn from this comparison. While the $\langle \cos^2(\theta) \rangle$ yield value is similar for both calculations (as well as in further calculations using this approach), the pulse shapes are dramatically different. Evidently, the pulse shape that the algorithm finds is heavily influenced by the way that the adaptation of the pulse intensity steered the calculations through the search landscape. This behavior is consistent with *theoretical* analysis of Quantum Control landscapes and their *level sets* [131, 134].

Evolution of Pulses

We devote this section to the exploration of the pulse shapes obtained in the dynamic intensity environments. Our experimental procedure has essentially an asymmetric nature due to its two stages: The first stage of ramping the intensity from low-to-high requires a learning phase (see Figure 9.17), whereas when reversing the process and bringing the intensity back down the optimization starts from a converged result. Thus, highly optimized solutions can be maintained throughout the latter excursion, and the transition from high-to-low intensity can be continuously observed. This process is illustrated both in Figure 9.19 and in Figure 9.20. In the latter, a sequence of pulses are shown, starting from pulses at low intensity (top-left corner), where the learning process takes place, moving along the snapshot gallery in a matrix-indexing-order fashion, to the center of the plot where the intensity is in its maximal regime, before reducing to a lower intensity again for the pulses shown in the lower-right part of the plot. These latter pulse-shapes are very simple pulse trains, with a pulse separation of $1/(3B_{rot}c) = 2.2\mathrm{ps}$. Such a pulse train is very different from the pulse train obtained for the static problem (Figure 9.14), where a pulse separation of 1.1ps was observed in the static calculation at $\Omega_{ge} = 40 \times 10^{12}\mathrm{s}^{-1}$. Nevertheless, the alignment observed at the end of the optimization of Figure 9.20 reaches a value of $\langle \cos^2(\theta) \rangle = 0.548$, which compares rather well with the value of 0.550 obtained in Figure 9.14. At these low intensities, as previously observed at high intensity, vastly different pulse shapes are able to produce similar optimized values of $\langle \cos^2(\theta) \rangle$. These solutions are on a *level set*, but the present calculations do not reveal if these solutions are on connected (i.e., continuously morphable from one level set to another), or disconnected components of the level set. At low intensity, the $1/(6B_{rot}c) = 1.1\mathrm{ps}$ pe-

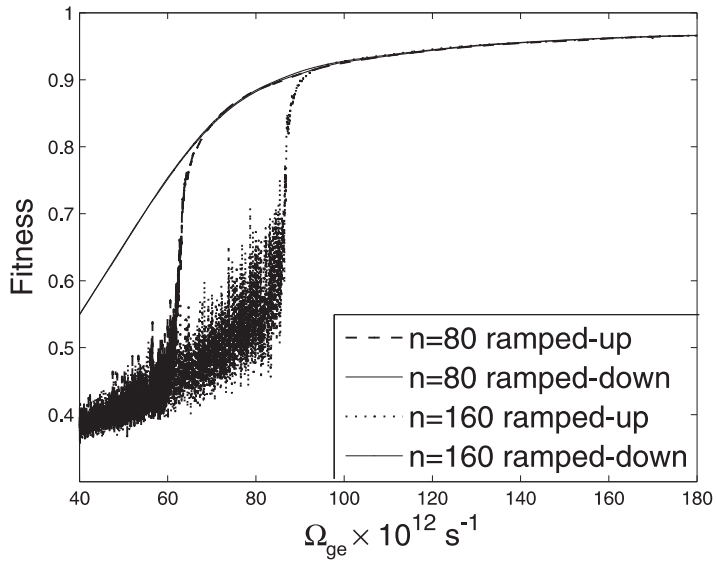

Figure 9.17: Evolution course of the best DR2-plain runs for phase resolutions of $n_1 = 80$ and $n_2 = 160$ pixels, on the **ramped-up** intensity (dashed or dotted, respectively) versus the ramped-down intensity (reversed direction, solid curves that merge for both runs). Each direction corresponds to 10^5 generations (10^6 function evaluations).

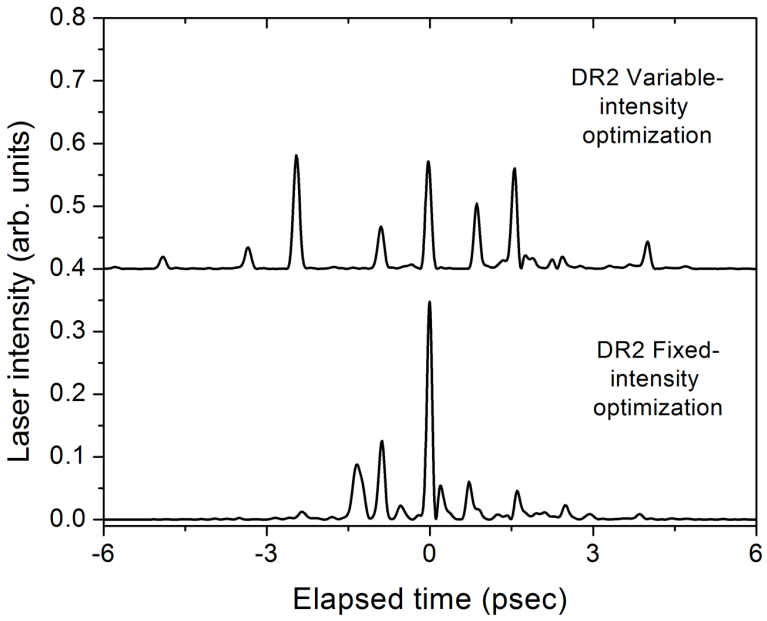

Figure 9.18: Comparison of pulse shapes that were obtained in optimizations employing the DR2-plain procedure, when using a fixed $\Omega_{ge} = 160 \times 10^{12}\mathrm{s}^{-1}$ (bottom, and see Figure 9.14), or – at this same value of $\Omega_{ge} = 160 \times 10^{12}\mathrm{s}^{-1}$ – in the course of an optimization where Ω_{ge} was linearly varied from $40 \times 10^{12}\mathrm{s}^{-1}$ to $180 \times 10^{12}\mathrm{s}^{-1}$.

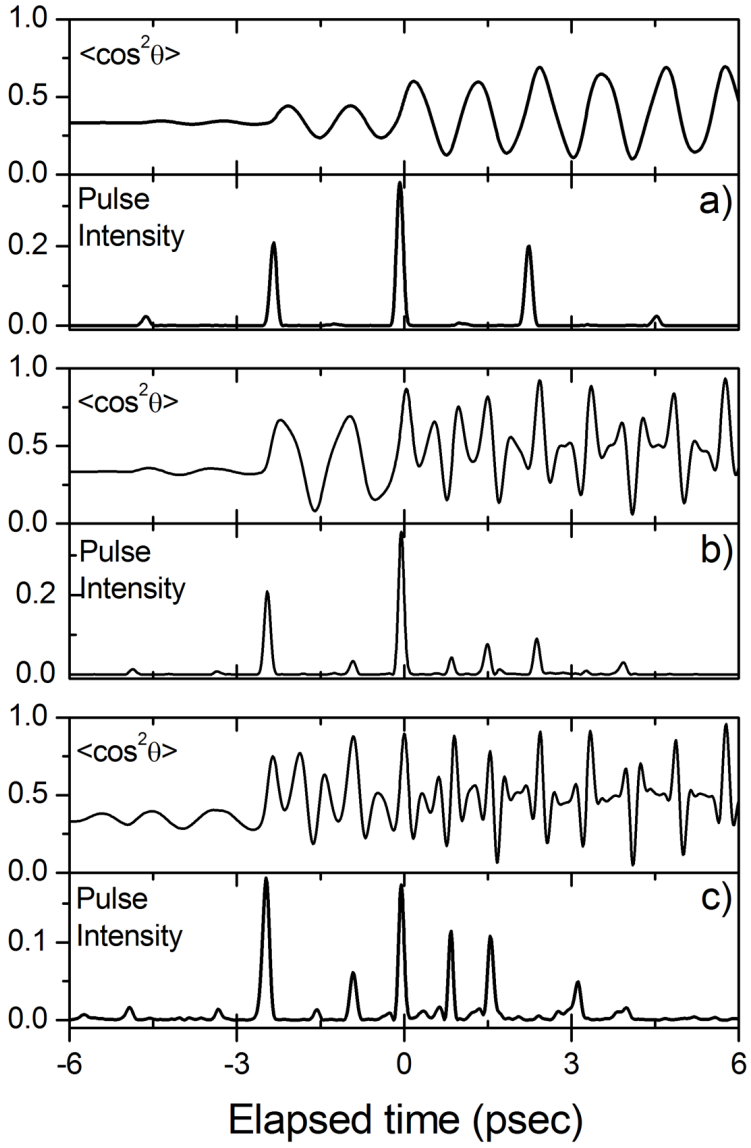

Figure 9.19: Intensity dependence of the alignment $\langle \cos^2(\theta) \rangle$ and the laser pulse shape from the **ramped-up** dynamic intensity environment, subject to a linear increase: $\Omega_{ge} := 40 \times 10^{12}\text{s}^{-1} \rightarrow 180 \times 10^{12}\text{s}^{-1}$. Snapshots are taken at – **(a)** $54 \times 10^{12}\text{s}^{-1}$, **(b)** $110 \times 10^{12}\text{s}^{-1}$, **(c)** $166 \times 10^{12}\text{s}^{-1}$ – and analyzed respectively.

riod observed in Figure 9.14 and the $1/(3B_{rot}c) = 2.2$ps period observed in Figure 9.20 correspond to a laser interaction that occurs once per period $T_{rev_{02}} = 1/(6B_{rot}c) = 1.1$ps of the $J = (0, 2)$ coherent superposition state (Figure 9.14), or every second period (Figure 9.20). This can easily be observed in Figure 9.19, where the temporal behavior is shown for the laser pulse shape and the induced dynamic alignment for $\Omega_{ge} = 54 \times 10^{12}\text{s}^{-1}$, $\Omega_{ge} = 110 \times 10^{12}\text{s}^{-1}$, and $\Omega_{ge} = 166 \times 10^{12}\text{s}^{-1}$. As the intensity is increased, higher rotational states begin to contribute to the rotational wavepacket and the $T_{rev} = 1/(2B_{rot}c) = 3.3$ps rotational period begins to assert itself. This is a consequence of the energy differences between rotational levels J_0 and $J_0 + 2$, being multiples of $2B_{rot}$ for all values of J_0. In the latter half of the pulse ($t > 0$), additional narrowly spaced pulses come into play, being spaced by $T_{rev}/4 = 1/(8B_{rot}c) = 0.8$ps. The occurrence of these new peaks comes at the expense of the peak at 2.2ps, which is considerably weakened in the calculation at $\Omega_{ge} = 110 \times 10^{12}\text{s}^{-1}$ (Figure 9.19(b)), and is completely absent in the calculation at $\Omega_{ge} = 166 \times 10^{12}\text{s}^{-1}$ (Figure 9.19(c)). In the latter calculation a new peak has appeared at a delay of 3.3ps, corresponding to the full revival of the rotational wavepacket formed.

We thus conclude that the optimal pulses observed in the simulations arise as a result of an interplay between the temporal structure that is required to optimize the transfer from $J = 0$ to $J = 2$, leading to peak separations that are a multiple of $1/(6B_{rot}c)$, and the temporal structure that is required to optimize the transfer from there to higher rotational levels, which leads to peak separations that are multiples of $1/(8B_{rot}c)$.

Step-Size and Phase Trajectories

Figure 9.21 presents the calculation of the Euclidean distance between evolving control phase functions that are determined sequentially as optimal every 100 generations (i.e., between following best-individuals), as well as the global step-size of the mutation operator in those time stamps. Dramatic changes between control phases are observed in the initial learning period, as expected. This is followed by a trend of mild changes, with several bursts of $\approx 2\pi$ variations. We propose the so-called *wrapping effect* as an explanation for these $\approx 2\pi$-jumps: The control phase function is subject to $[0, 2\pi]$-periodic boundary conditions, that are enforced by *wrapping* a phase value. Upon examination of the phase space, it is indeed confirmed that these bursts are caused by a boundary wrapping of a phase function value (its index varies). We thus conclude that the variations in the phase space are consistently mild subject to the dynamic laser intensity. This is consistent with the step-size behavior (presented in $\log_{10}$ scale), which stays in the order of 10^{-2} after the learning period, with expected fluctuations.

Interestingly, following the initial learning period, the algorithm "stays in the neighborhood", which seems to be sufficient for determining optimal

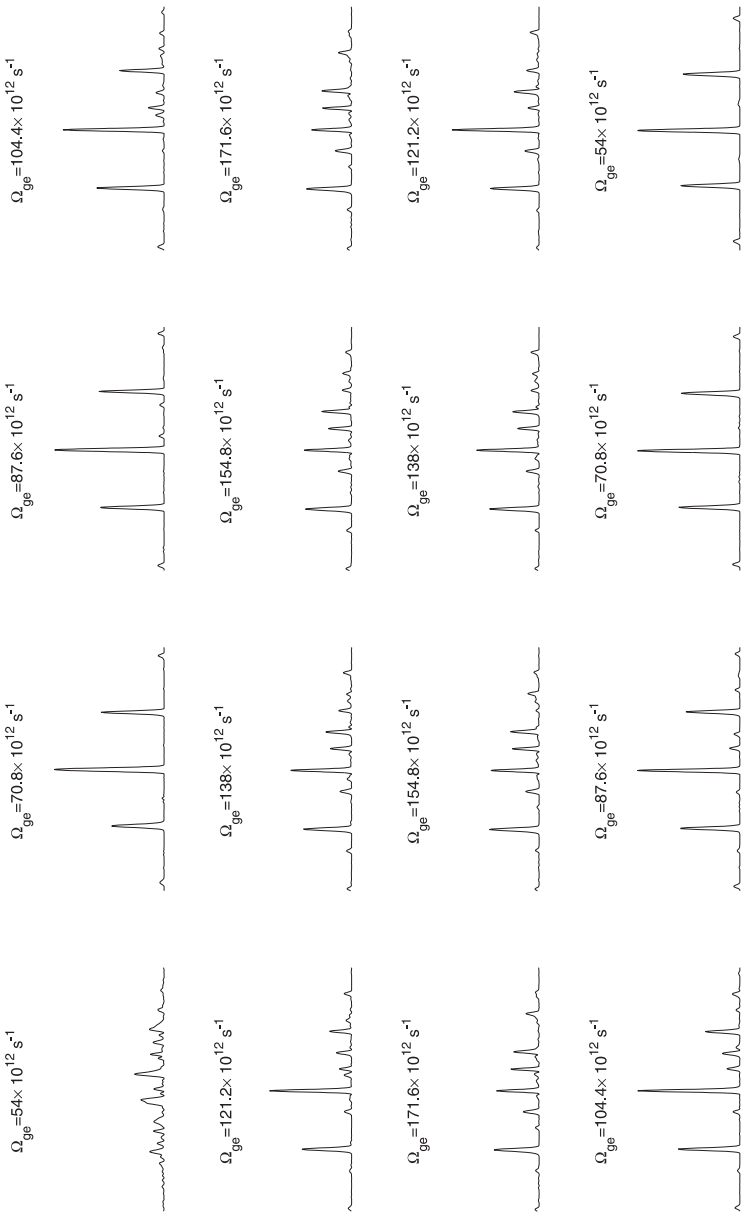

Figure 9.20: Evolution of laser pulses subject to linearly increased followed by linearly decreased laser intensity, $\Omega_{ge} := 40 \times 10^{12}\mathrm{s}^{-1} \rightarrow 180 \times 10^{12}\mathrm{s}^{-1} \rightarrow 40 \times 10^{12}\mathrm{s}^{-1}$, presented as snapshots of optimized pulse shapes at specific intensity milestones. The order follows a matrix-indexing fashion. The pulse-shapes obtained in the end of the process, i.e., after the ramping-down to the regime of low-intensity (bottom right) are a simple pulse train with pulse separation of $1/(3B_{rot}c) = 2.2\mathrm{ps}$.

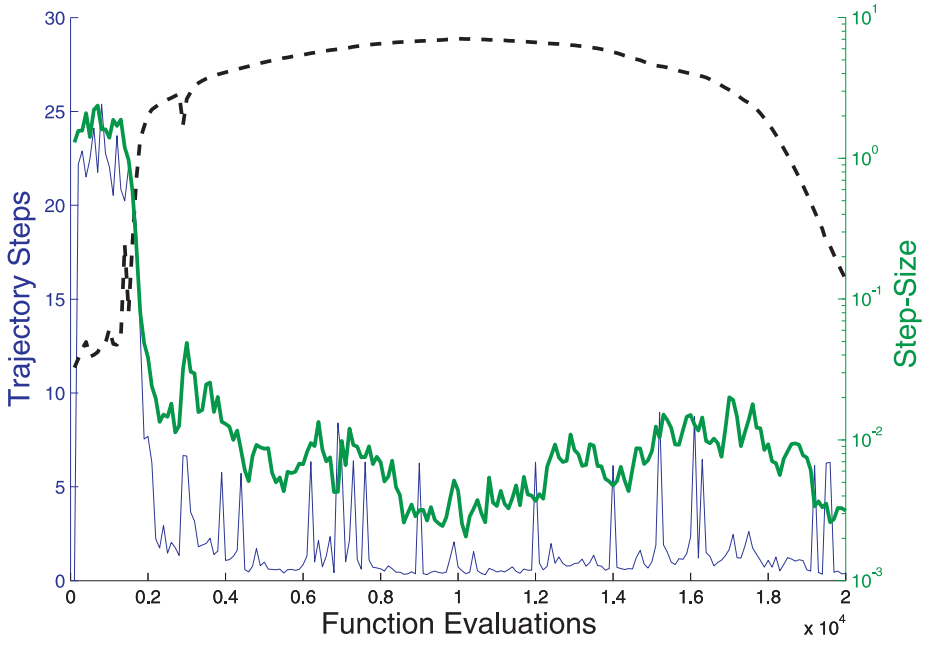

Figure 9.21: The evolution course of the DR2-plain on $n_1 = 80$ pixels subject to the ramping up and down laser intensity environment. Dashed line – unscaled fitness evolution; Thin solid line – the Euclidean distance between evolving control phase functions [scaled on the left axis]; Thick line - global step-size of the mutation operator [log-scaled on the right axis]. Dramatic Euclidean trajectories in the control phase function are observed during the initial learning period, as well as at specific bursts of $\approx 2\pi$ variations, corresponding to the so-called *wrapping effect*.

controls for the continuously changing laser intensity. This means that high alignment yield at different laser intensities corresponds to a neighborhood of the control space.

9.5 Scalability: Control Discretization

In this section we aim at exploring the scalability of the alignment problem with respect to the control resolution. So far, the latter has been fixed in our calculations to $n = 80$. In particular, we would like to study the trade-off between the control resolution, which allows fine-tuning of the electric field, to the success-rate of the evolutionary learning process, subject to a fixed number of function evaluations. Due to computational considerations, we choose to conduct the scalability calculations on the zero-Kelvin variant of the alignment problem. Also, we select the DR2 subject to the plain parameterization as our optimization kernel for this investigation.

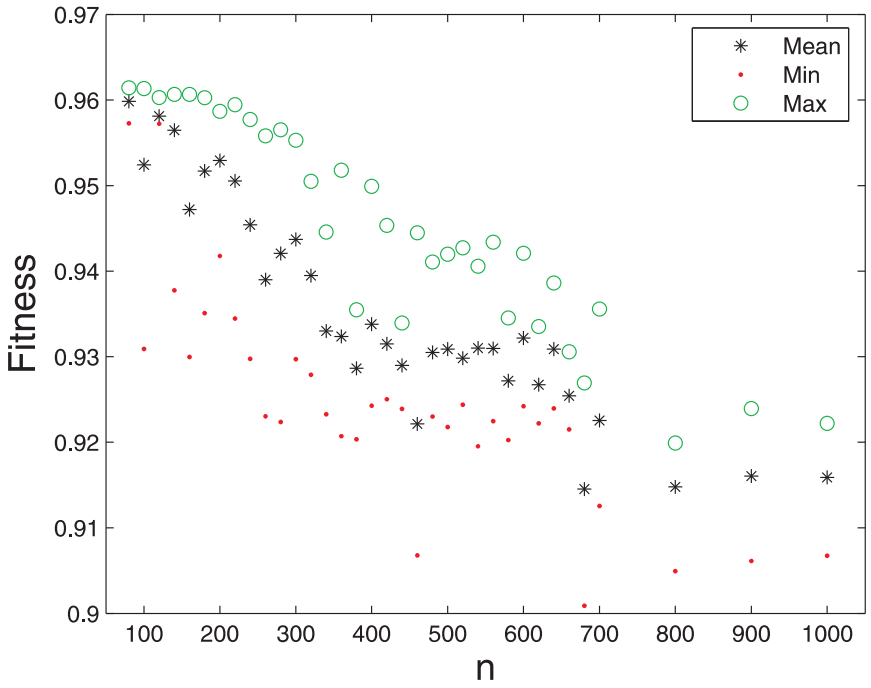

Figure 9.22: Best, mean and worst cosine-squared alignment values obtained by the DR2 for each parameterization, over 10 runs of 20,000 function evaluations each (see legend).

9.5.1 Numerical Observation

We apply the DR2 algorithm to the optimization task in the following manner: 10 runs per control discretization, with $n = \{80, 100, 120, \ldots, 680, 700\}$, and additionally with $n = \{800, 900, 1000\}$. Each run is limited to 20,000 function evaluations.

Figure 9.22 presents the numerical results of these calculations. The best, mean and worst fitness values obtained by the DR2, after 20,000 function evaluations, for each discretization, are presented. As can be observed, the best fitness value is attained for $n = \{80, 100\}$; As the dimension n increases, there seems to be a weak trend of fitness values decrease, but the DR2 still manages to obtain high quality solutions in the regime of 0.94 even for $n = 400$.

A typical evolution run for $n = 100$ is given in Figure 9.23. As can be observed from this plot, a successful learning is obtained after $\approx 5,000$ function evaluations. In higher dimensions, i.e., $n \geq 500$, the DR2 does not succeed in tackling the problem within the limited number of function evaluations. A typical run for $n = 700$ is presented in Figure 9.24.

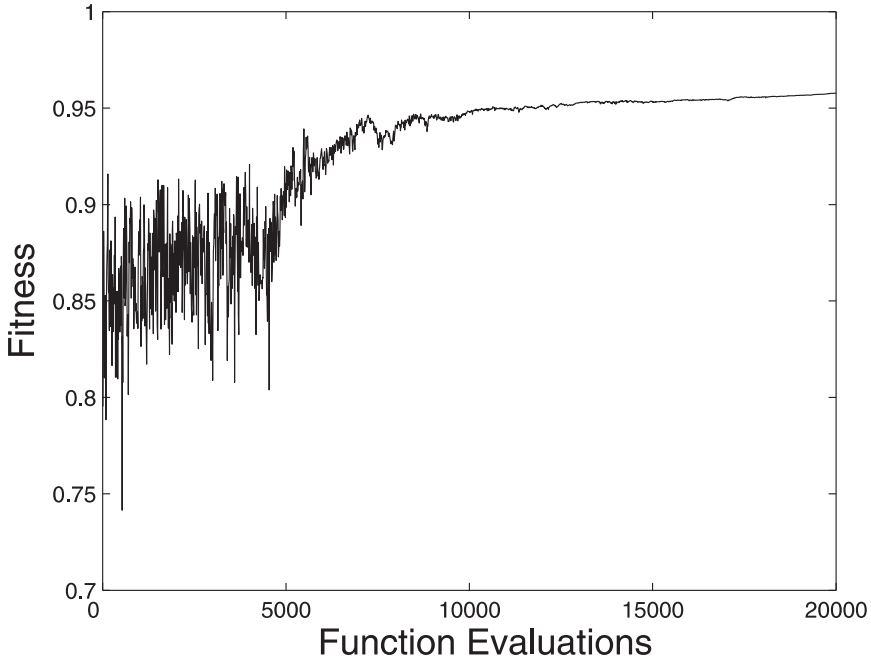

Figure 9.23: A typical DR2 evolution run for $n = 100$, with $20,000$ function evaluations. Successful learning is observed after $\approx 5,000$ evaluations.

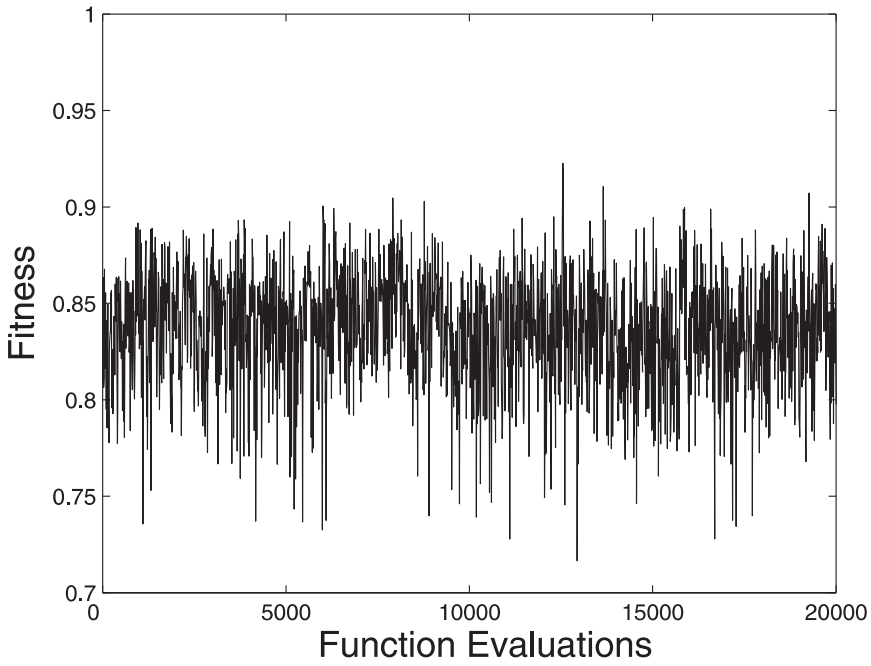

Figure 9.24: A typical DR2 evolution run for $n = 700$, with $20,000$ function evaluations. No successful learning is observed.

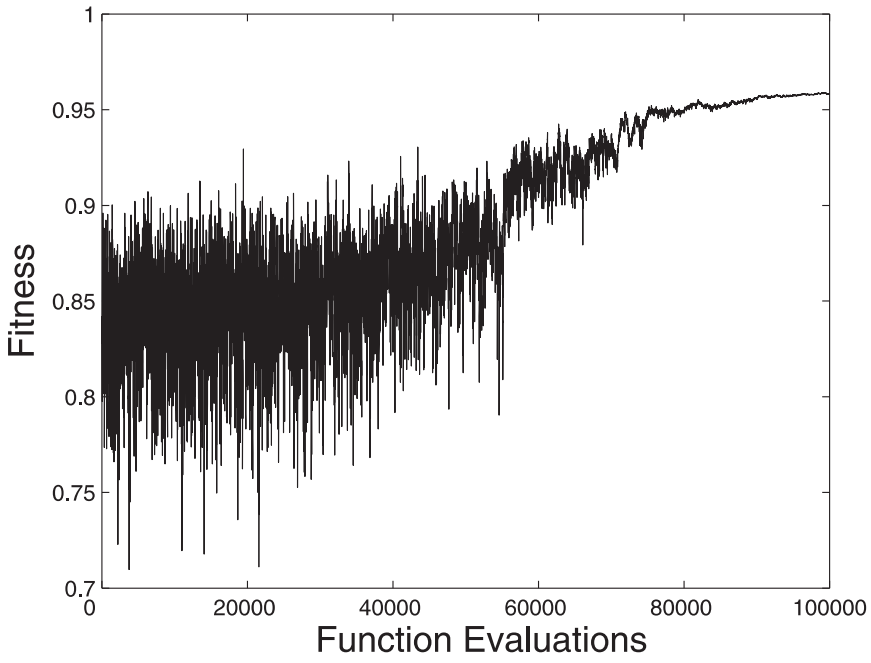

Figure 9.25: DR2 evolution run, for $n = 1000$, with $100,000$ function evaluations. The best cosine-squared alignment value found was $f^* = 0.9583$.

Granting Additional Function Evaluations

Given the numerical results of the previous section, we were interested in the question whether the fixed number of function evaluations posed a limitation on the search and did not allow a successful learning of the decision parameters and convergence into a good solution.

We have conducted another series of runs, limited now to $100,000$ function evaluations, for the extreme case of $n = 1000$. We were surprised to find out that some of the runs did succeed in converging successfully into fine solutions of high yield values. In particular, we would like to point out a run which attained a solution with cosine-squared alignment value of $f^* = 0.9583$, a value which is close to the highest value known to us for this variant of the problem. The plot of that specific evolution run is given in Figure 9.25. A rough observation reveals that the DR2 'takes-off' into a convergence pathway only after $\approx 50,000$ function evaluations, and then it needs additional $30,000$ function evaluations to reach saturation. This numerical observation indicates that the learning task of the decision parameters in this problem is still feasible in higher dimensions of the control function, as long as the granted number of function evaluations is sufficiently large. From the algorithmic perspective, the employed DES variant, the DR2 algorithm, tackled successfully this 1000-dimensional problem. **However, from**

the *physics* perspective, such a high-resolution parameterization does not seem to pay-off, as far as the cosine-squared observable is concerned, and there seems to be no justification to employ discretization of the control phase function with more than n=80 pixels.

9.6 Intermediate Discussion

Our calculations so far, especially in Sections 9.3 and 9.4, show that it is possible to encounter high diversity of optimal solutions in constrained numerical simulations of Quantum Control, and moreover, that the examination of such rich sets of solutions can become an important aspect of the control experiments. The diversity of successful controls likely contains useful dynamical information, and may also provide the *decision maker* with a list of choices to consider for weighing in other ancillary control criteria, e.g., multi-criterion decision making. The present calculations optimizing dynamic molecular alignment in a diatomic molecule exposed to an intense, shaped laser field, provide compelling evidence that the absolute value of the quantity that is being optimized (i.e., the fitness) is the true measure of success, and that the same value of the fitness may be achievable by widely differing laser pulse shapes that share only a limited number of common features. Each of these solutions has the potential of carrying valuable information about the underlying physics, where some of the solutions provided key information on the dynamics of the alignment process. Viewed in this sense, the uniqueness of the fitness value, and the diversity of the solutions that can lead to accomplishment is a blessing in disguise.

We also showed that the optimized alignment yield attained a value which was very close to the maximal possible yield in the current framework, even when the constraints on the optimization translated into a significant distortion of the resultant wavepacket. By relaxing specific constraints, we showed that it was possible to enhance the observable alignment further toward the maximal attainable alignment possible for the rotational basis set used. This outcome leads to the optimistic conclusion that high yields may be obtained, even when *a priori* it seems that the system is subject to severe constraints for constructing the wavepacket. As discussed, the origin of this behavior can be understood in terms of the *variational principle*, as well as the physical observable involving an integration over the wavefunction which hides some of its discrepancies.

As a direct implementation of these conclusions, we would like to complete our work on the optimization of dynamic molecular alignment by means of two additional aspects - multi-objective optimization, as well as niching.

9.7 Multi-Objective Optimization

As further investigation of the alignment problem, we would like to extend our single-criterion optimization approach to a Pareto Optimization approach. As previously introduced in Chapter 5, *Pareto Optimization* aims at attaining the efficient set for a given multi-objective optimization problem and its corresponding Pareto front. In particular, we are interested in removing the penalty approach to high-intensity pulses, and rather consider the fluence of the pulse as an independent objective, subject to *minimization*. Thus, the observable's yield remains as an objective, while we choose to define the total-SHG signal of the electric field as the secondary objective subject to minimization.

Formally, we aim at finding the Pareto front for the following bi-criteria problem:

$$\begin{aligned} f_1 &= \max_{\mathrm{E(t)}} \left\langle \cos^2(\theta) \right\rangle \longrightarrow \max \\ f_2 &= \int_{-\infty}^{\infty} |E(t)|^4 dt \longrightarrow \min \end{aligned} \tag{9.10}$$

In order to select an appropriate optimization method, the following characteristics of the objective functions in the application problem are of importance: Based on our accumulated experience with the problem in its single-criterion form, we assume that the functions f_1 and f_2 are continuous in most points, highly nonlinear and multimodal. Nothing is known yet about the shape of the Pareto front for the application problem. Analytical techniques and methods based on differential calculus are likely to fail in this problem, because of the complexity of the integral equations.

9.7.1 Choice of Methods

We choose to apply the NSGA-II, as presented earlier (Section 5.1.2), to the current task. Due to the duration of the simulator evaluation, we would like to consider a specific metamodel that may allow for the acceleration of the calculations.

Metamodel-Assisted NSGA-II In order to accelerate stochastic optimization algorithms in the presence of time consuming function, metamodels have been frequently proposed (see, e.g., [167, 168, 169]). A metamodel is an approximation of an objective function that is learned from a set of evaluations.

More explicitly, given a set of points $\vec{x}^{(1)}, \ldots, \vec{x}^{(k)} \in \mathbb{R}^n$, and the corresponding evaluations of the objective functions at these points, $\vec{f}^{(1)} = f\left(\vec{x}^{(1)}\right), \ldots, \vec{f}^{(k)} = f\left(\vec{x}^{(k)}\right)$, the metamodel can be used to compute an approximation, denoted by $\hat{f}(\vec{x}) \approx f(\vec{x})$, for any point $\vec{x} \in \mathbb{R}^n$, in a duration which is considerably shorter than the precise evaluation. As expected,

metamodels tend to be more precise near the training points.

Kriging[1], also referred to as *Gaussian random field models*, is a particular type of interpolation model that has been frequently applied for metamodeling [167, 168, 169]. The statistical motivation for this method is that the deterministic objective functions are considered to be realizations of a Gaussian random field $\mathcal{G}$. This assumption makes it possible to compute a measure for the uncertainty of predictions, i.e., each prediction value is associated with a standard deviation that can be used for computing two-sided *confidence intervals*.

It is typically assumed that these random variables $\mathcal{G}_{\vec{x}}$ are correlated by means of a spatial correlation function,

$$c : \mathbb{R}^n \times \mathbb{R}^n \to [-1, 1],$$

i.e., a correlation function that depends only on the positions of the random variables in space. In our study we shall use a correlation function of the form:

$$c(\vec{x}, \vec{x}') = \exp\left(-\theta \left|\vec{x} - \vec{x}'\right|^2\right)$$

The correlation function of the Gaussian random field is estimated from the given data, or given *a-priori*. In this study we apply leave-one-out cross-validation to determine an appropriate value of θ, as suggested in [170]. After the correlation function is estimated, the prediction is made. For this purpose, the conditional Gaussian distribution at the given input vector $\vec{x} \in \mathbb{R}^n$ is computed.

A practical implementation of Kriging has been described by Emmerich [101], and it was successfully employed in engineering design optimization [100, 169, 171]. Multi-objective problems were typically approached by learning metamodels for each objective function separately, in an implementation known as *local Kriging*. We omit here its derivations, and refer the reader to [101].

In the metamodel-assisted NSGA-II [171], Kriging metamodels are used to pre-evaluate the set of offspring solutions and select favorable variants among it for precise evaluation. The uncertainty information can be used to facilitate search in less explored regimes of the landscape.

Algorithm 9 outlines the general Metamodel-Assisted Evolutionary Algorithm (MA-EA), as described by Emmerich [101]. The difference to the generic Evolutionary Algorithm can be summarized as follows:

- All *precisely evaluated points* are stored in a database, denoted by D_t (cf. lines 4 and 9).

[1] Kriging originates from *geostatistics*, and is named after the mining-engineer Krige.

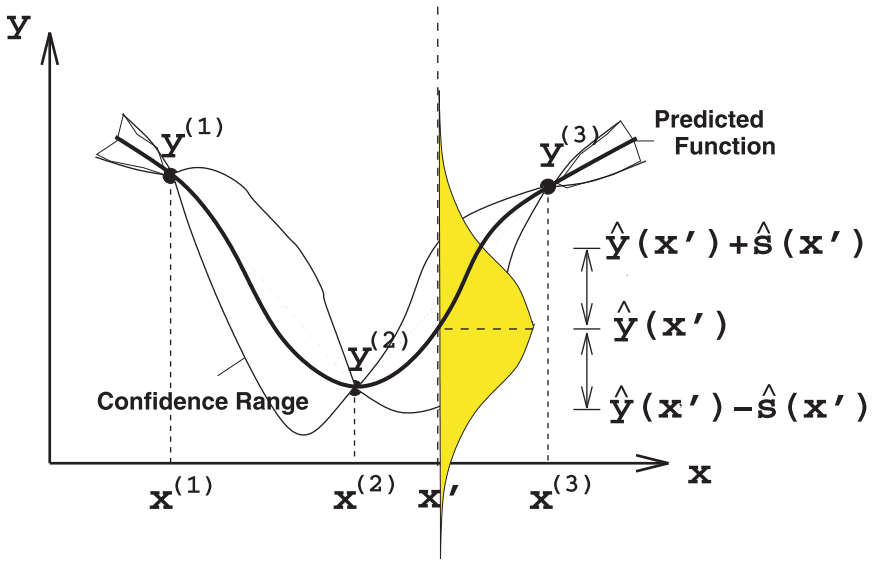

Figure 9.26: Outputs of Gaussian Random Field Metamodels using a $\mathbb{R} \to \mathbb{R}$ mapping example. Three points, $\vec{x}^{(1)}$, $\vec{x}^{(2)}$, and $\vec{x}^{(3)}$ have been evaluated here. The result of each approximated evaluation at a point $\vec{x}'$ is represented by the mean value, $\hat{y}$, and by the standard deviation, $\hat{s}$, of a $1D$ Gaussian distribution. Figure courtesy of Michael Emmerich [101].

- The algorithm filters out less promising solutions (cf. line 8) and thereby reduces the offspring population size. The remaining solutions are then precisely evaluated and considered in the subsequent selection.

There are many possibilities to design filters for that purpose. In this study we restrict ourselves to constant output size filters. The size of the resulting filtered set will be denoted by ν and the corresponding MA-EA will be termed a $(\mu + \nu < \lambda)$-EA. All filters will be rank-based, i.e. they sort the offspring population with respect to some criterion, a so-called *filter criterion*.

We offer a $3D$ visualization in Figure 9.27 in order to gain some intuition into the different concepts of *filters* in the bi-criteria case. In the latter, the Pareto-front approximation of the current population is depicted, as well as three offspring individuals, namely $\vec{x}_1, \vec{x}_2$ and $\vec{x}_3$. The offspring individuals have been evaluated with the Kriging metamodel, and thus their precise values are not yet known, but rather the defining parameters of $2D$ Gaussian random variables, $\mathcal{G}_{\vec{x}_i}$. The distributions of the random variables $\mathcal{G}_{\vec{x}_1}, \mathcal{G}_{\vec{x}_2}$, and $\mathcal{G}_{\vec{x}_3}$ are also visualized in the diagram by means of their *probability density functions*.

Four different criteria have been discussed by Emmerich [101] for assigning a *yield value* to a search point $\vec{x}$, which is based on the prediction provided

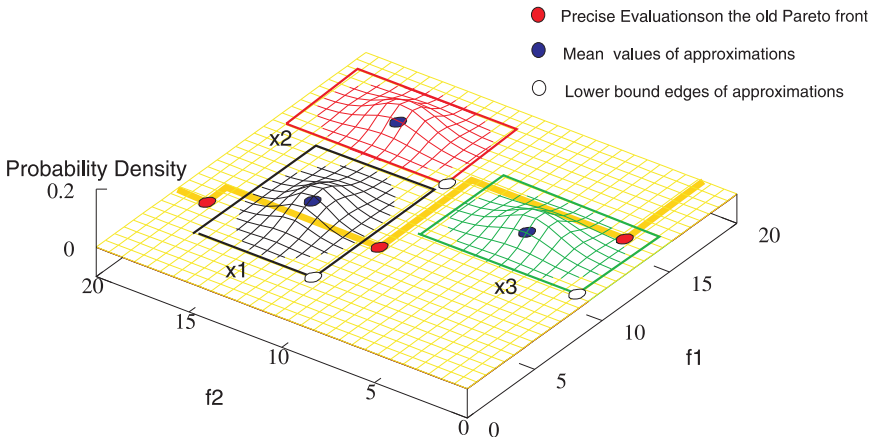

Figure 9.27: Interval boxes for approximations in a solution space with two objectives. Figure courtesy of Michael Emmerich [101].

by the defining parameters of the Gaussian predictor $\mathcal{G}_{\vec{x}}$:

- **Mean Value** Non-dominated / crowding distance sorting, based on the expected value for $\mathcal{G}_{\vec{x}}$ given by $\hat{f}(\vec{x})$.
- **Lower Confidence Bound (LCB)** Non-dominated / crowding distance sorting on the lower bound edge of the confidence interval of $\mathcal{G}_{\vec{x}}$.
- **Probability of improvement (PoI)**: The probability that the realization of $\mathcal{G}_{\vec{x}}$ is non-dominated. It can be computed via integration over the non-dominated set.
- **Expected Improvement (ExI)** The expected increase in the dominated hypervolume for $\mathcal{G}_{\vec{x}}$ is measured.

Modus Operandi

We applied the following algorithmic kernels to the Dynamic Molecular Alignment:

- NSGA-II: The classical variant by Deb [99, 172].
- Metamodel-Assisted EA with Probability of Improvement (PoI-EMOA).
- Metamodel-Assisted EA with Expected Improvement (ExI-EMOA).

The parameterization of these methods is $\mu = 50$, $\nu = 0.2 \cdot \lambda$, with two different settings for λ: $\lambda = 250$ and $\lambda = 50$. The parameters of the mutation operator and recombination operator have been chosen as described by Deb

Algorithm 9 $(\mu + \lambda)$-**MA-EA**

1: $t \leftarrow 0$
2: $P_t \leftarrow \text{init}()$ $\{P_t \in \mathcal{S}^\mu$: Set of solutions$\}$
3: Evaluate(P_t)
4: $D_t \leftarrow P_t$
5: **while** $t < t_{\max}$ **do**
6: $G_t \leftarrow \text{Generate}(P_t)$ {Generate λ variations}
7: Metamodel_evaluate(G_t) {Metamodel is derived from D_t}
8: $Q_t = \text{Filter}(G_t)$
9: $D_{t+1} \leftarrow D_t \cup Q_t$
10: $P_{t+1} \leftarrow \text{Select}(Q_t \cup P_t)$ {Rank and select μ best}
11: $t \leftarrow t + 1$
12: **end while**

[99]. Due to implementation considerations, in practice both objectives were minimized, and therefore we assign:

$$f_1 \rightarrow \max \Longrightarrow -f_1 \rightarrow \min$$

9.7.2 Numerical Observation

Figures A.14, A.15 and A.16 present the results of our calculations, where the 20%, 50% (median), and 80% attainment surfaces are plotted. Each one of them refers to 5 runs with 20,000 evaluations per run. In order to make the curves easier to be distinguished, we zoomed-in a box around the *knee point* of the Pareto front approximations.

Discussion

The results clearly indicate that there is a conflict between the two objectives, as suspected. Thus, Pareto optimization is an appropriate tool for solving this problem. The fact that a *convex* Pareto front has been observed suggests that good compromise solutions are likely to be found. We observe a sharp increasing flank at both ends of the approximated Pareto front. Regions of fair trade-offs range from about -0.6 to -0.4 in the $(-f_1)$ coordinate.

There are significant differences in the behavior of the multi-objective EA variants. The best coverage of the Pareto front has been achieved by the ExI-EMOA. This variant is the only variant that found solutions for f_1 above 0.58. The highest value found was 0.6184. The PoI-EMOA resulted in approximations with lower spread. However, the precision of this EMOA variant was better in the regions covered. This result is consistent with some theoretical findings reported in [171], as well as with their numerical assessment on artificial problems reported there. The expected improvement measure puts emphasis on exploring unknown regions, while the probability

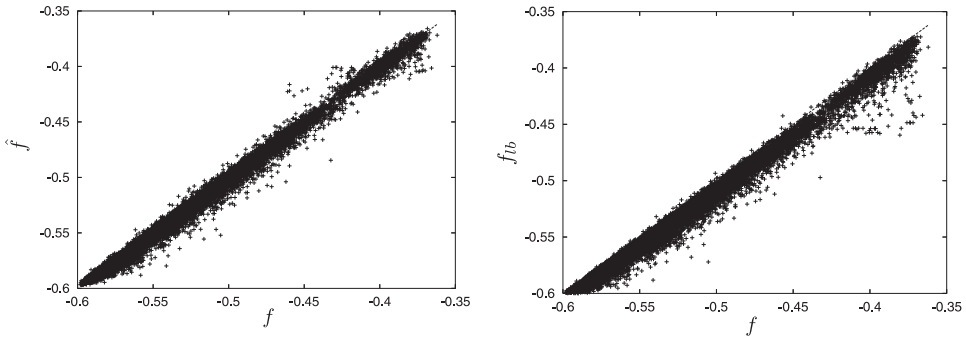

Figure 9.28: Left: $f - \hat{f}$ - plot for $(-f_1)$; Right: $f - f_{lb}$ - plot for $(-f_1)$.

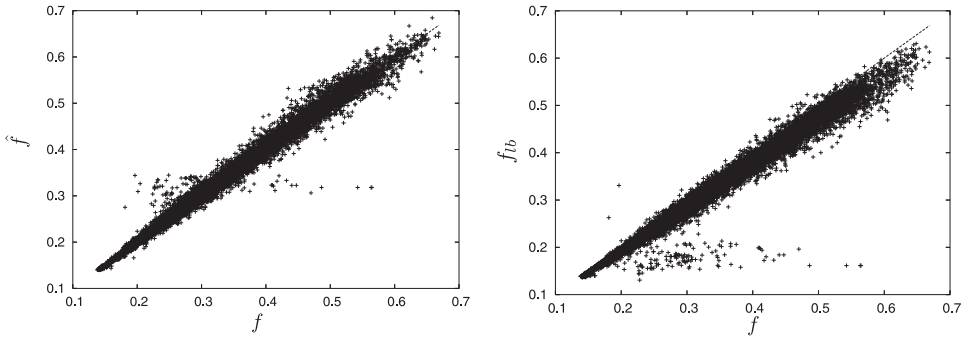

Figure 9.29: Left: $f - \hat{f}$ - plot for f_2; Right: $f - f_{lb}$ - plot for f_2.

of improvement have the tendency to carry out better exploitation of visited regions. Overall, the metamodel-assistance seems to be a valuable ingredient for this problem, as can be seen by comparing the results of the NSGA-II with those of the metamodel-assisted EMOA.

A more detailed analysis of the metamodel-based approximations was performed, in order to assess whether the metamodeling worked as expected from theory. The results are displayed in Figures 9.28 and 9.29, for one of the runs with the ExI-EMOA ($\lambda = 250$). The $f - \hat{f}$ plots indicate that in the whole range of function values the results obtained with the metamodel were strongly correlated with the true function values. The error bandwidth for f_1 is about 10% of its range versus 15% for f_2 with respect to its range. These results correspond to results in similar studies in metamodel-assisted optimization [171]. Moreover, the lower confidence bounds, denoted by f_{lb}, have been compared to the outcome of the precise evaluations, f. Here, the 95.45%-lower confidence bounds, as computed by the Kriging method, have been assessed for their validity (see Figures 9.28 and 9.29). The results are in conformity with theory for f_1. However, some outliers for f_2 in the region

of f_{lb} from $0.15 - 0.2$ should be reported. However, these outliers did not seem to hamper the algorithmic performance.

From the physics point of view the obtained result is interesting, since it shows the nature of the trade-off between the alignment's observable and the intensity of the electric field, expressed here by means of the second harmonic generation signal. The importance of the intensity criterion is likely to govern the decision of the expert on the trade-off surface, which is to look for solutions with relatively good f_1 values in the region of fair trade-offs.

9.8 Application of Niching

We shall apply here our DES niching algorithms to the zero-Kelvin variant of the dynamic molecular alignment. Following the application of niching to the population transfer problem in the rotational framework, as described in Section 8.3, we take into consideration the diversity measure issue, and fully adopt the conclusions drawn in Section 8.3.1.

Modus Operandi

We consider here three niching strategies:

1. The $(1, \lambda)$-DR2 – for being the best method to perform on this problem, and also as a representative of first-order strategies.
2. The $(1, \lambda)$-CMA – as a representative of second-order information strategies.
3. The $(1 + \lambda)$-CMA – as a representative of elitist strategies.

We conduct 10 runs per method, searching for $q = 3$ niches, subject to plain parameterization of the control phase at $n = 80$ pixels. Each run was limited to $15,000$ function evaluations per niche.

9.8.1 Numerical Observation

The calculations are discussed at several levels.

Niche-Radius

Following the derivation done for the niche radius in the population transfer problem in Section 8.3.2, we conducted preliminary runs with a niche-radius of $\rho = 110$. However, it performed poorly, in an equivalent way to its initial performance on the population transfer problem: The DR2 as well as the CMA-comma failed, and the CMA-plus obtained good solutions only for the first niche.

Table 9.6: Three niches obtained in 10 runs – averaged yield values (in parentheses - best value attained) – for the three employed niching strategies.

Ranked-Niches	**DR2**	**CMA**	**CMA+**
Best niche	0.9417 (0.9605)	0.8553 (0.9029)	0.9517 (0.9585)
2^{nd}-best niche	0.8477 (0.9552)	0.8229 (0.8561)	0.9493 (0.9525)
3^{rd}-best niche	0.8054 (0.8558)	0.7966 (0.8161)	0.9365 (0.9484)

Table 9.7: Niches correlation for the niches obtained in 10 runs – averaged cross-correlation values, as defined in Eq. 8.16.

Niches Correlation	**DR2**	**CMA**	**CMA+**
$c_{1,2}$	0.6784	0.6952	0.6312
$c_{1,3}$	0.6288	0.6905	0.6062
$c_{2,3}$	0.7593	0.6951	0.6414

We managed to get satisfying results for $\tilde{\rho} = 55$, as will be reported here. Thus, consider all the reported results here as obtained with $\tilde{\rho} = 55$.

Success-Rate

The *cosine-squared alignment* of the three methods, for the three obtained niches, is presented in Table 9.6. We can observe a clear trend - the CMA+ mechanism outperformed the other mechanisms, with consistent location of three good niches on average. However, the DR2 mechanism managed to obtain the top-quality solutions for the best as well as for the 2^{nd}-best niches, in consistency with our previously reported results. The latter typically failed to locate a 3^{rd} good niche. The CMA comma-strategy, on the other hand, simply failed in obtaining satisfying niching results on this landscape.

Niches Cross-Correlation

We calculated the cross-correlation coefficients for the obtained pulse-shapes of the different niches, as defined in Eq. 8.16. The results of these calculations are presented in Table 9.7. We may state that the pulse-shapes of the different niches are weakly correlated to one another. In particular, it is interesting to note the low correlation values of the the CMA+ kernel.

Laser Pulse Designs

Our definition of a distance measure to this problem has been proved to be successful. The obtained pulses in the time-domain had indeed different characteristics, representing different conceptual laser-pulse designs. Three niches, obtained in a typical CMA+ run, are plotted by means of their pulse intensities and revival structures in Figures A.17, A.19 and A.21.

Conceptual Quantum Structures Revisited

We would like to offer an additional analysis for our niching solutions. Figures A.18, A.20 and A.22 provide the SWFT picture for the obtained solutions. It can be observed that these three solutions represent the same conceptual quantum structure of states population. This SWFT observation reinforces our conclusions concerning the correlation between the employed optimization routine in combination with the applied parameterization to specific conceptual quantum structures, as drawn in Section 9.3.1. Therefore, we do not find it surprising that all three obtained pulses share the same 'behind-the-scenes physics', due to the fact that they were all obtained with the same algorithmic kernel (e.g., CMA+), subject to the plain parameterization. This observation does not contradict our primary conclusion that the niching process has been successful in locating three different pulse shapes in the temporal domain, as initially required. It simply reveals an additional, well-hidden, degeneracy among the solutions. In the next section we shall offer a way to remove this second degeneracy completely.

Removing the Second Degeneracy

Given the additional degeneracy which was encountered in the SWFT space, one can further develop a problem-specific diversity measurement. In this case, our idea is to consider the *wavepacket space*, and more explicitly, to evaluate the differences between the population of rotational levels, $\left|a_j^{(t)}\right|^2$, as the measurement of diversity between niches. The implementation itself is straightforward, due to the fact that the vector of population coefficients is given by the alignment-routine. Since the coefficients are normalized, subject to the normalization postulate of Quantum Mechanics, it is fairly simple to estimate the niche radius in this case.

Niche Radius: Wavepacket Space According to the Quantum Mechanics normalization postulate, the wavepacket coefficients in the N-dimensional Hilbert space are normalized:

$$\sum_{j}^{N} \left|a_j^{(t)}\right|^2 = 1$$

In the wavepacket treatment for removing the second degeneracy, these coefficients play the role of the decision parameters, as far as the diversity measurement is concerned.

The calculation of r of Eq. 3.3 simply reads:

$$r = \frac{1}{2}\sqrt{\sum_{j=1}^{N_{rot}} \left|a_j^{(t)}\right|^2} = \frac{1}{2}$$

With $q = 3$ and $N_{rot} = 20$, Eq. 3.5 yields:

$$\rho = \frac{\frac{1}{2}}{3^{\frac{1}{20}}} \approx 0.47 \tag{9.11}$$

Thus, we set it to $\rho = 0.5$. We choose to employ only the CMA+ kernel in this case, subject to plain as well as Hermite parameterizations, aiming to show feasibility of the defined diversity measure.

This newly-defined diversity-measurement for the alignment problem has been observed to be successful. The obtained pulses in the temporal domain had indeed different characteristics, and in particular their shapes differed in a satisfying manner. We consider here the results obtained when the Hermite parameterization was employed. The best niche obtained in every run was typically of the optimal class known to us: Both the cosine-squared alignment yield, as well as the pulse shape and the population profile, were associated with the best solutions reported previously. The second-best niche was a representative of a sub-optimal set of solutions: It had a lower value of cosine-squared alignment yield and a different profile of population. However, note that the third-best located niche was not typically an interesting solution, as it had dramatically lower alignment values in comparison to the first two niches. The temporal pulse-shapes themselves were very weakly correlated.

Typical solutions of best and second-best niches are plotted in Figures A.23 and A.25, with their corresponding SWFT pictures in Figures A.24 and A.26.

Discussion

We would like to summarize our numerical observation of the applied niching algorithms to the dynamic molecular alignment problem. Niching with the CMA+ kernel performed best, while always obtaining three niches of high-quality laser pulses. The DR2 found the best solution, in consistency with our previously reported observations, but did not perform well on the secondary niches. The CMA-comma failed to obtain satisfying niching results.

The original calculation of the niche radius was not successful at the practical level, as reported for the population transfer problem. After introducing a factor of 0.5 to the original value, the niching process was observed

to be successful. The obtained pulse-shapes were typically weakly correlated, as required.

As far as the algorithmic performance is concerned, we adopt the conclusions drawn for the application of niching to the population-transfer problem. We thus ascribe the failure in practice of the originally calculated niche radius, as well as the compromised performance of the comma-strategy kernels on the secondary niches, to the highly constrained nature of the landscape when underposed to a radius-based niching framework.

Furthermore, we have applied a physics numerical assessment, at the quantum rotational picture, with the so-called SWFT technique. The latter has supported previous observations concerning the correlation between optimization routines in combination with parameterizations to conceptual quantum structures. This observation revealed that all three niches of a given run, which differ sufficiently at the laser-pulse design level (temporal domain), typically share the same conceptual quantum structure at the SWFT picture (wavepacket space). We offered another diversity measure, which relies on the physics information, in order to remove this second degeneracy. This approach indeed succeeded in that, and obtained multiple solutions corresponding to different conceptual designs.

While the growing corpus of knowledge could be represented by the diameter of an expanding circle, the horizons of ignorance and open questions would be then represented by the area of that circle.
Chinese proverb

Summary and Outlook

Our journey has gone so far through the realms of *Natural Sciences*, while keeping a guiding torch of *Computing* and *Operations Research*. The journey is coming to its closure, and thus we would like to summarize it.

Our starting point was the field of Evolution Strategies, a computational discipline which stems from Evolutionary Biology. We presented it in Chapter 1, and described in detail a new generation of its algorithms, the so-called Derandomized Evolution Strategies. We suggested to consider these state-of-the-art ES variants as powerful optimization methods with local-search capabilities.

Chapter 2 was the gateway to *niching*, and treated a wide spectrum of related topics. In particular, we deepened furthermore into the world of Biology, exploring the topics of *diversity* and *organic variations*. We turned from there back to the optimization arena, where we considered a definition of the *attraction basin*, the part of the search landscape which is equivalent to the ecological niche. We discussed the important issue of population diversity within Evolution Strategies. Especially, we reviewed previous research conducted on the loss of diversity in ES, due to two main components: Selective pressure (take-over effect), and drift (neutral effects, associated with both recombination and selection). We thus reached the conclusion that an Evolution Strategy which employs a small population will inevitably lose its population diversity.

At the same time, we presented calculations which suggested that ES with small populations are subject to a so-called *mutation drift*. The latter allows for easy translation of populations from one location to another, an effect that has the potential to boost fast speciation. This observation thus provided us with further motivation to apply niching with DES, algorithmic variants which typically employ small populations.

This was followed by a survey of classical niching methods, mainly from the GA field. We concluded this introductory chapter with postulating our mission statement with respect to *niching*. In short, we argued that a niching technique should attain the optimal interplay between the partitioning into stable subpopulations and the exploitation of each niche by means of an efficient optimizer with local-search capabilities.

Armed with this mission statement, and motivated by various results

suggesting that DES would be an attractive choice for algorithmic kernels in a niching framework, we accepted upon ourselves the challenge. Chapter 3, the core of Part I, introduced our proposed framework of niching in derandomized-ES, subject to a fixed niche-radius approach. The framework was inspired by biological concepts and by classical GA niching techniques. In biological terms, the proposed algorithm was associated with a speciation model of individual *alpha-males*. Following a detailed description of our method, we outlined a testbed of artificial multimodal continuous landscapes. Upon the application of the proposed algorithm to the search of minima in these landscapes, we analyzed the numerical observation with the so-called MPR Analysis. The latter allowed us to derive parametric values that typically define the behavior of each DES variant as a niching kernel. Our observation concluded that the CMA plus-strategy, which has the lowest *niching acceleration*, performed better than the other DES variants. Our proposed explanation for that considered the niching problem as a constrained optimization problem, where a plus-strategy is argued to have an advantage for ES.

Chapter 4 was a direct extension of Chapter 3, and it aimed at treating the niche radius problem. By employing the CMA algorithmic kernel, we proposed two different approaches for self-adaptation of niche-radii and niche-shapes, based on *step-size coupling* and the application of the *Mahalanobis distance*, respectively. We tested the various proposed variants on artificial multimodal landscapes, including landscapes with even and uneven spread of optima. The performance was highly satisfying, and was investigated by means of the MPR Analysis.

In Chapter 5 we introduced our niching framework into the multi-objective arena. Our stated mission was to treat multi-global optimization problems. More specifically, our goal was to boost diversity in the decision space, and by doing so to offer more choice in the typically conflicting decision making process. We derived a *multi-parent* niching-CMA variant for that purpose, and showed that the application to a specific set of multi-objective problems required only mild algorithmic adjustments. The observed performance was highly satisfying, and provided us with the desired *proof of concept*.

Chapter 6 was the gateway to Part II, reviewing the main topics of OCT and OCE in the context of optimization. It outlined various important theoretical results, which concluded that controllable unconstrained quantum systems have extrema that correspond to perfect control, or to no control at all. Furthermore, perfect control could be typically obtained with only first-order (*gradient*) information while climbing-up the QC landscape; At the top of the landscape, there is an infinite number of attainable optimal points. Despite the fact that these results are valid for "perfect" theoretical landscapes with no constraints, they play an important role in posing QC

optimization problems, and in suggesting certain remarkable properties that might be instantiated in practice. Some of the work reported here corroborated some of these properties.

Our practical work on Quantum Control systems began in Chapter 7, where we considered two systems of two-photonic processes both in simulations and in the laboratory. Upon analysis of pre-mature convergence of DES variants on these landscapes, due to the unrestricted search employed by them, we introduced the so-called *wrapping* operator into the ES framework. The CMA outperformed the other algorithms on those landscapes, even without using its second-order (*covariance*) information. We found these results to be an experimental corroboration of the OCT landscape analysis discussed in Chapter 6.

The *quantum rotational framework*, which constituted a considerable part of our research, was treated in this study at several levels throughout Chapters 8 and 9. Chapter 8 laid out the Quantum Mechanical foundations of the rotational framework, and posed the so-called *population transfer problem*. The latter was investigated by means of simulations at different laser intensities, which revealed a rich landscape with a wide variety of optimal solutions. Moreover, it was observed that the number of independent solutions critically depends on the difficulty of the problem, determined by the laser intensity. The study of the *rotational population transfer* problem was concluded with the application of our niching algorithms. The latter required the definition of a tailor-made distance metric, due to invariance properties of the control phase function. The numerical simulations obtained good niching results, where the *elitist* CMA kernel performed best. Due to the fact that the original niche-radius calculation for this landscape failed in practice, as well to the fact that the *comma-strategies* did not perform well on *secondary optima*, we speculated that the introduction of a radius-based niching approach to this landscape posed a highly-constrained optimization problem.

Last but not least, the *dynamic molecular alignment* problem was extensively investigated in Chapter 9. We began the chapter by providing the motivation for obtaining molecular alignment, and then formally posed the problem. Following a straightforward application of DES to the problem we further approached it from multiple angles. We developed a parameterization method, that was shown to boost the convergence of DES on the alignment landscape. Moreover, we introduced a simplified variant of the original alignment problem, at zero Kelvin temperature, which allowed an improved investigation from the Physics perspective. The examination of certain DES variants subject to specific parameterizations resulted in a fruitful study of optimality, where two classes of solutions, optimal and sub-optimal, were revealed. This optimality study also concluded that despite the considerable difference between the composition of the optimized wavepacket and the maximally attainable wavepacket, the obtained yield in the optimization was typically fairly close to the maximally attainable yield. This excellent

behavior was explained by means of the *variational principle*.

We proceeded with optimizing the alignment problem subject to a dynamic intensity environment. This resulted in a new perspective on the *evolution of laser pulses*, and confirmed furthermore our understanding of the optimal structures within laser pulses applied to this problem.

This was followed by the employment of a multi-criteria approach to the alignment problem, while considering the minimization of the total *second harmonic generation* signal as a secondary objective with respect to the alignment yield objective. Due to the heavy computational cost of the simulator, we introduced the so-called Kriging Metamodel in order to boost our calculations. This application confirmed our suspicion of the existence of a conflict between the two objectives, which had been treated previously by means of a penalty term.

Finally, we applied our niching algorithms to the alignment problem. By following the tailor-made distance metric introduced in Chapter 8, our first round of calculations obtained satisfying results. All the different niches represented, nevertheless, the same conceptual quantum design, as expected from our previous investigation of optimality. Thus, we carried out a second round of calculations, where the distance between the niches was measured in the wavepacket space. The latter achieved the goal of removing the observed degeneracy. We linked the failure of the originally calculated niche radius to the compromised performance of the comma-strategy kernels on the secondary niches, and ascribed both to the highly constrained nature of the landscape when underposed to a fixed radius-based niching framework.

Upon concluding this study, the message that we would like to post is three-fold. Firstly, we would like to encourage the application of niching methods to high-dimensional real-world hard problems, for providing the *decision makers* with the choice among several optimal or near-optimal solutions. As was demonstrated here, the proposed niching framework with DES kernels was capable of providing satisfying results on the investigated Quantum Control landscapes. Furthermore, we showed that the employment of a domain-specific tailor-made diversity measure is possible, when necessary. Secondly, we believe that the important *multiple optima identification* task has not yet attracted the proper attention of the scientists in the Evolutionary Computation community, and some would even claim that it is often neglected. Therefore, we hope that a corresponding *sub-community* within the EC community will emerge in the near future. Thirdly, we argue that the field of Quantum Control is a highly attractive testbed for optimization methods, as well as a rich arena of challenging open problems. As such, it should enjoy the powerful capabilities of state-of-the-art Evolutionary approaches, at all possible levels: multi-criterion optimization, niching techniques, optimization in environments with uncertainty, etc.

Outlook

We believe that we compiled a genuine interdisciplinary study, with two main contributing components: The first, the introduction of niching with the powerful kernels of Derandomized ES variants to the arena of multimodal function optimization, and the second, the introduction of Quantum Control to state-of-the-art evolutionary approaches. We, nevertheless, believe that there are still various directions for future research.

It would be interesting to further apply our proposed niching framework to additional search landscapes, either artificial or from the real-world applications domain. In addition, the multi-globality goal in multi-objective optimization could be further explored, by means of extended algorithmic developments and by means of an application to practical optimization problems.

Another challenging direction would be the development of additional niching frameworks with DES kernels which do not utilize a niche-radius based approach. As devoted followers of the **No Free Lunch Theorem**, we believe that there is always room for competing methods.

On the Quantum Control front, there are still many open research topics that are related to our study. At the experimental level, it would obviously be exciting to optimize in the laboratory the Dynamic Molecular Alignment. These experiments are approaching count-down at Amolf-Amsterdam, upon the completion of this dissertation.

On that note, Quantum Control Experiments introduce many possible optimization topics for future research. Such topics are the investigation of noise and its effect on algorithmic performance, robustness of obtained controls, the application of niching as well as multi-criteria optimization in the experimental learning-loop, and others.

By introducing these challenges we conclude this study, which hopefully turned out to be an enjoyable natural computing experience for the reader.

We would like to end with the simple call: "*keep it natural!*".

Appendix A

Additional Figures

We present here additional figures in full-color format.

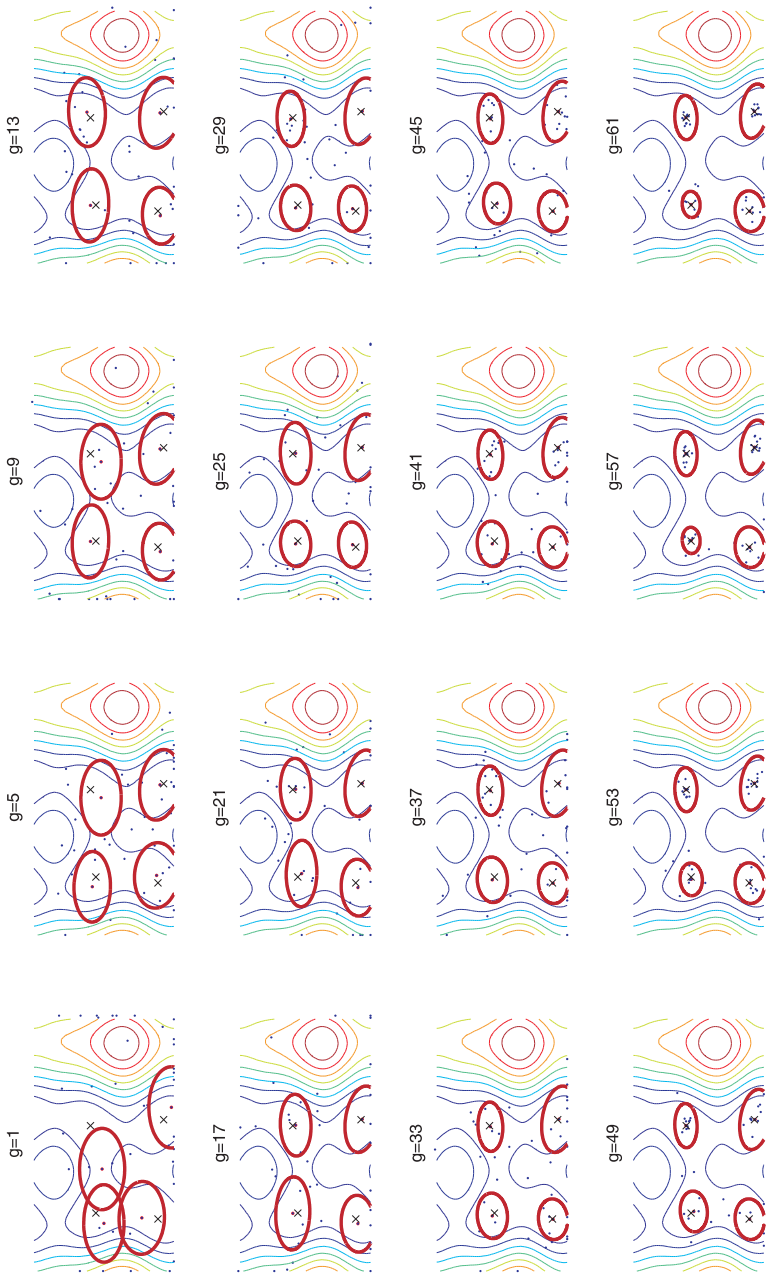

Figure A.1: A snapshot gallery: The adaptation of the *classification-ellipses*, subject to the Mahalanobis metric with the updating covariance matrix, with the CMA+ kernel on the $2D$ Fletcher-Powell problem. Images are taken in the box $[-\pi, \pi]^2$. Contours of the landscape are given as the background, where the X's indicate the real optima, the dots are the evolving individuals, and the ellipses are plotted centered about the peak individual. A snapshot is taken every 4 generations (i.e., every 160 function evaluations), as indicated by the counter.

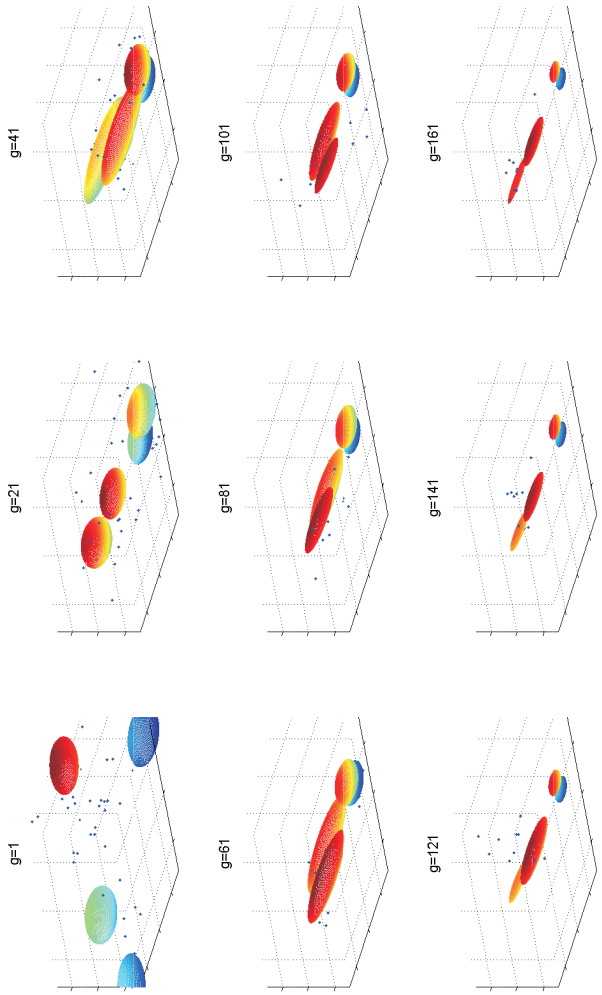

Figure A.2: A $3D$-snapshot gallery: The adaptation of the *classification-ellipses*, subject to the Mahalanobis metric with the updating covariance matrix, with the CMA+ kernel on the $3D$ Fletcher-Powell problem. Images are taken in the box $[-\pi, \pi]^3$. The ellipses are centered about the evolving peak individuals. A snapshot is taken every 20 generations (i.e., every 800 function evaluations), as indicated by the counter.

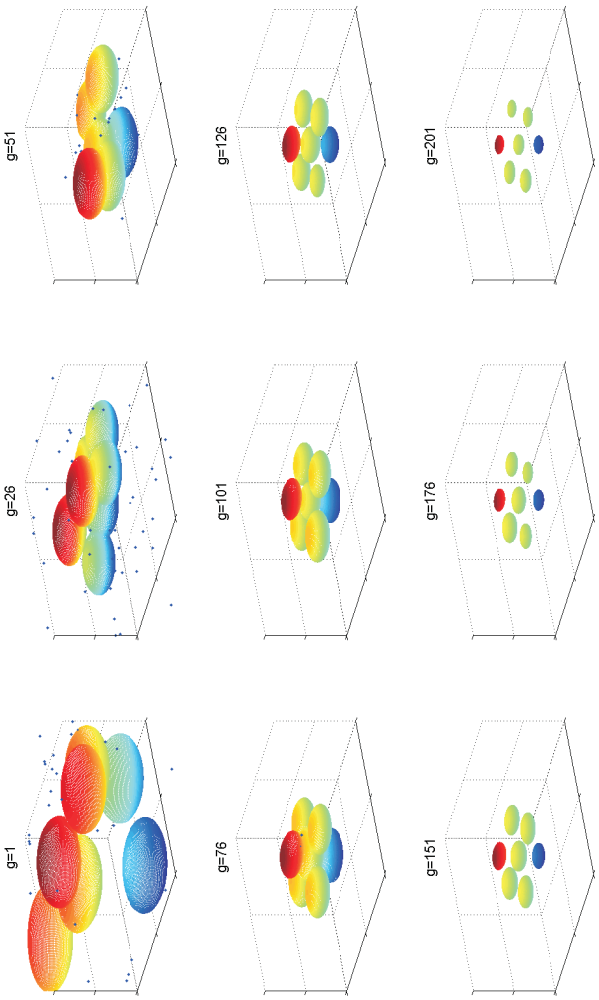

Figure A.3: A $3D$-snapshot gallery: The adaptation of the *classification-ellipses*, subject to the Mahalanobis metric with the updating covariance matrix, with the CMA+ kernel on the $3D$ Ackley problem. Images are taken in the box $[-2, 2]^3$. The ellipses are centered about the evolving peak individuals, and are observed to adapt simultaneously. A snapshot is taken every 25 generations (i.e., every 1750 function evaluations), as indicated by the counter.

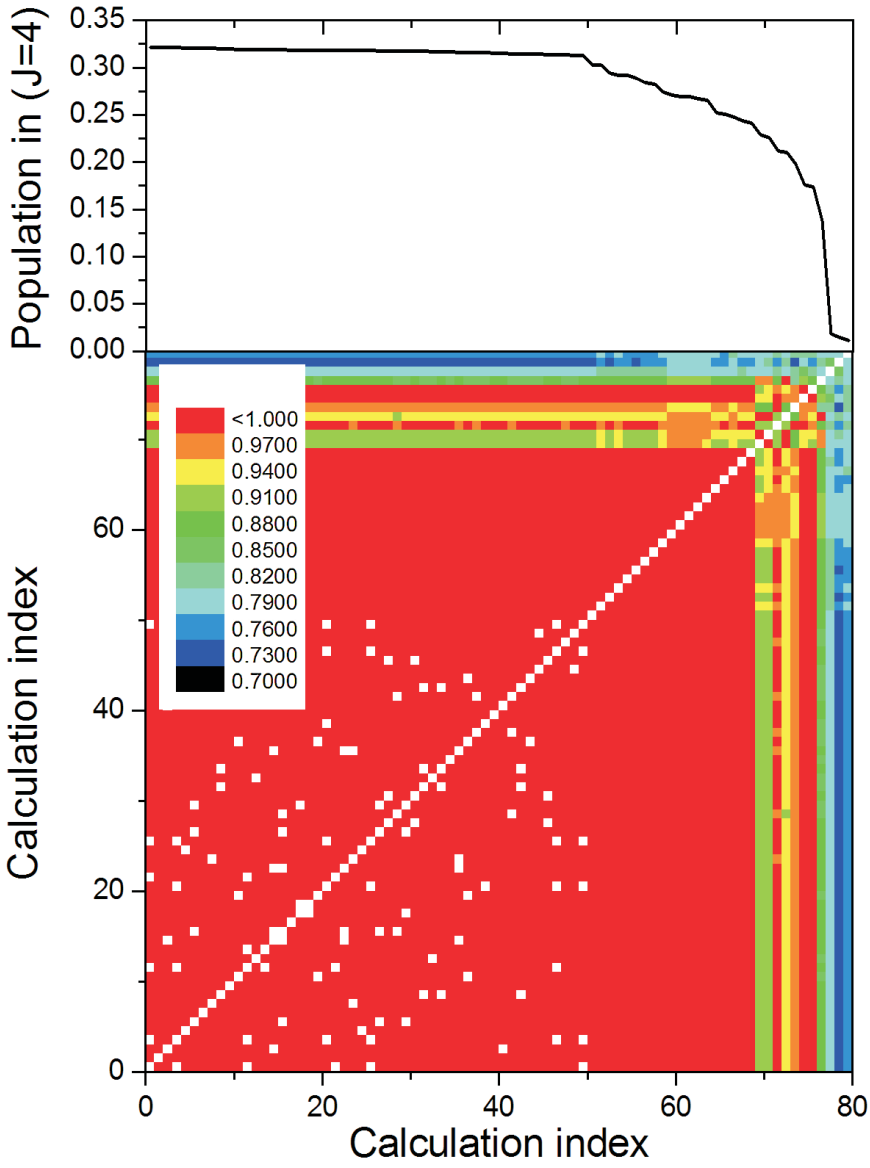

Figure A.4: Population transfers from $J = 0$ to $J = 4$ obtained in 80 runs of the DR2 algorithm with $\Omega_{ge} = 80 \times 10^{12}\mathrm{s}^{-1}$ (top), along with the correlation coefficient between the solutions, as defined in Eq. 8.16 (bottom). The solutions that perform best are highly correlated. Pixels in white correspond to cross-correlation value of 1 (after rounding-off).

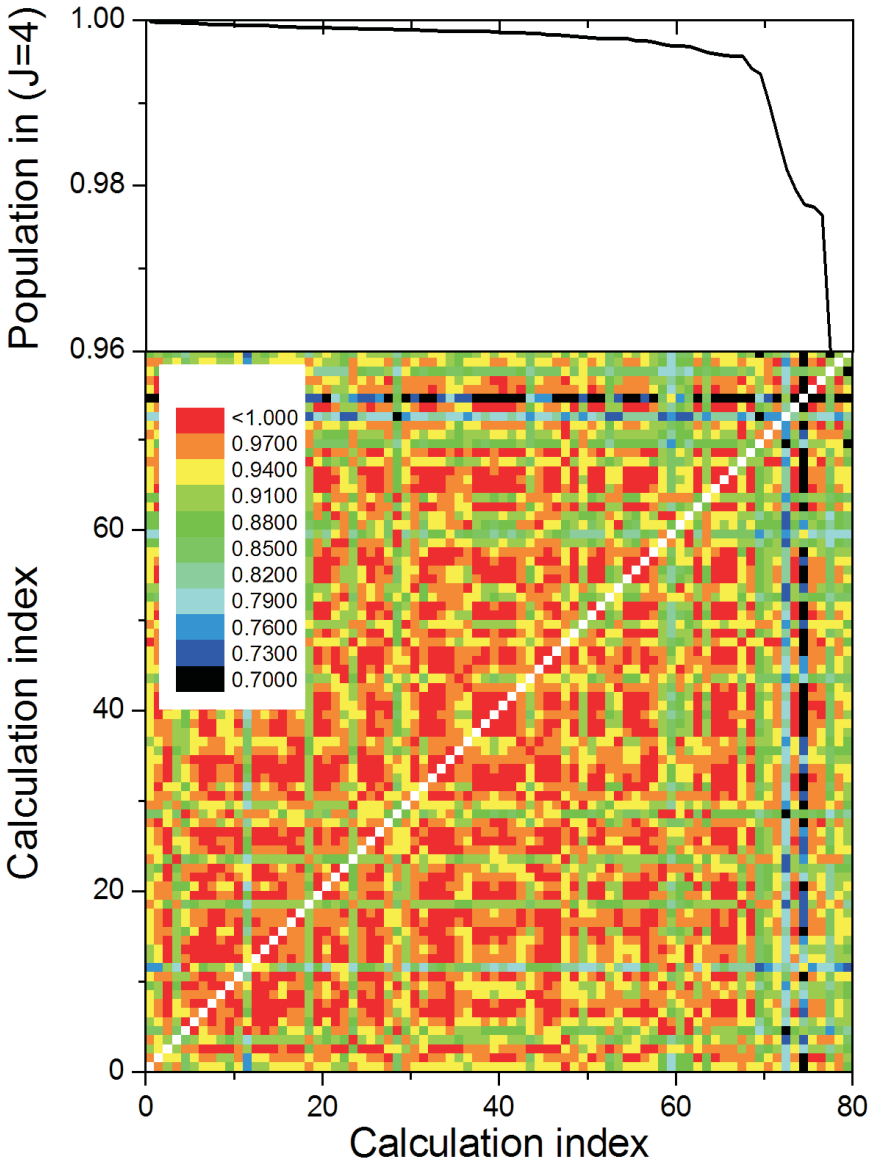

Figure A.5: Population transfers from $J = 0$ to $J = 4$ obtained in 80 runs of the DR2 algorithm with $\Omega_{ge} = 120 \times 10^{12}\mathrm{s}^{-1}$ (top), along with the correlation coefficient between the solutions, as defined in Eq. 8.16 (bottom). The solutions that perform well can be divided into a finite group of solutions that are highly correlated within the group but not with solutions outside the group. Pixels in white correspond to cross-correlation value of 1 (after rounding-off).

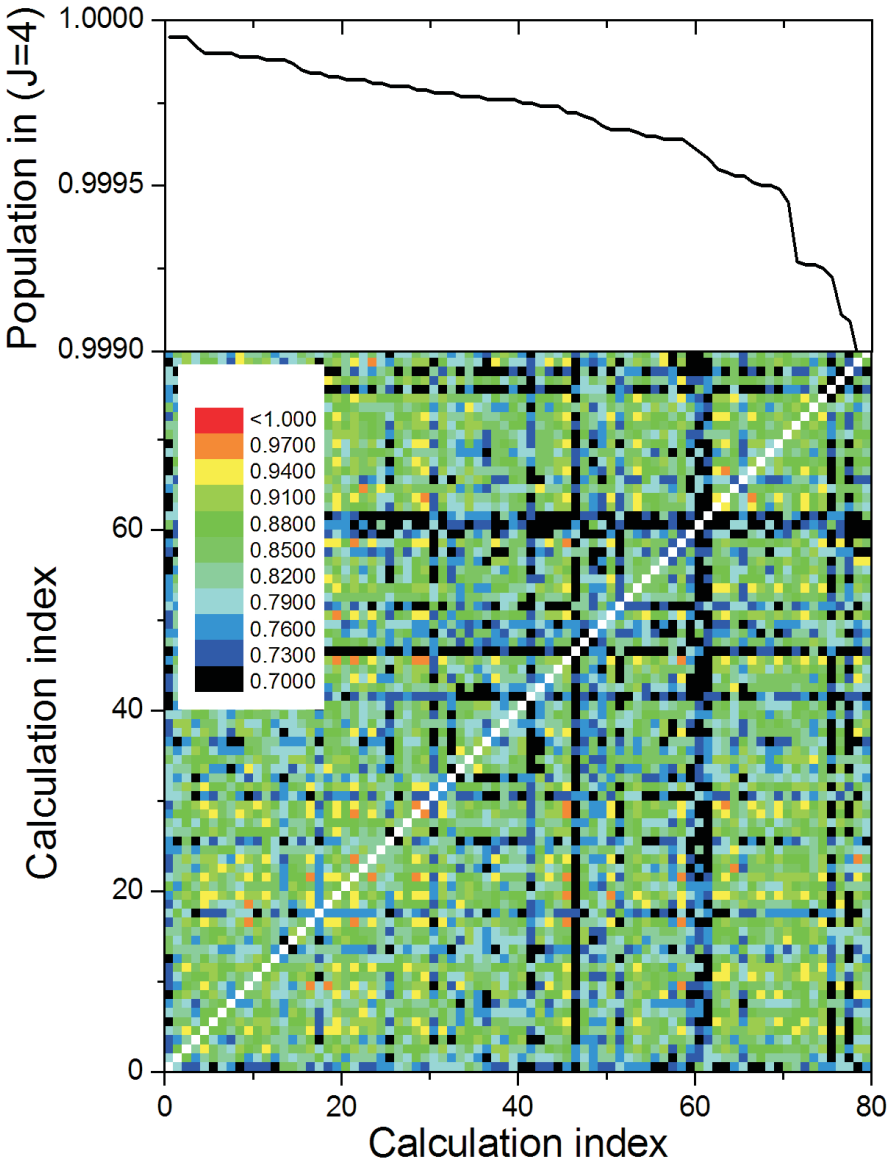

Figure A.6: Population transfers from $J = 0$ to $J = 4$ obtained in 80 runs of the DR2 algorithm with $\Omega_{ge} = 160{\times}10^{12}\mathrm{s}^{-1}$ (top), along with the correlation coefficient between the solutions, as defined in Eq. 8.16 (bottom). Many near-perfect solutions exist that are only weakly correlated to each other. Pixels in white correspond to cross-correlation value of 1 (after rounding-off).

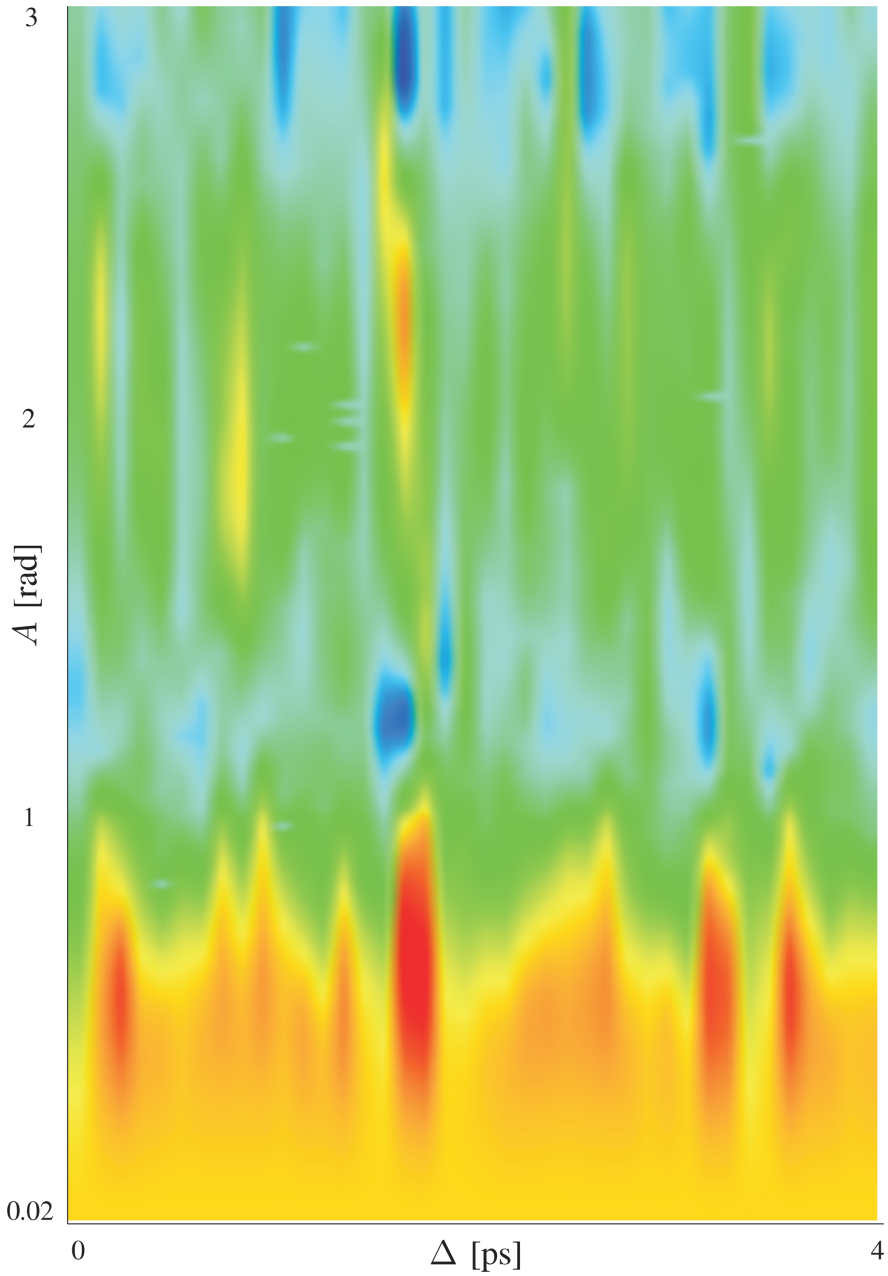

Figure A.7: Contourplot of $\langle \cos^2(\theta) \rangle$ as a function of A and Δ as defined in Eq. 9.6 for a peak Rabi frequency of $\Omega_{ge} = 180 \cdot 10^{12} \mathrm{s}^{-1}$. The color scale ranges from 0.3551 (blue) to 0.688 (red). Figure courtesy of Christian Siedschlag [162].

Alignment and Revival Structure of two obtained solutions. Thin red line: Alignment; Thick black line: Laser pulse intensity.

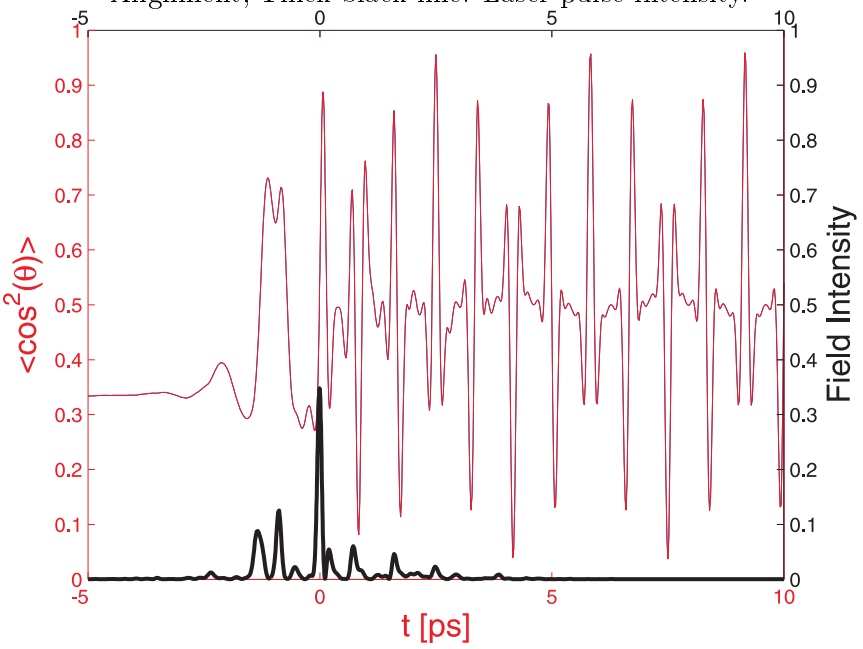

Figure A.8: A typical *optimal* solution, obtained by the DR2-plain; Alignment yield: $< \cos^2(\theta) >= 0.9622$.

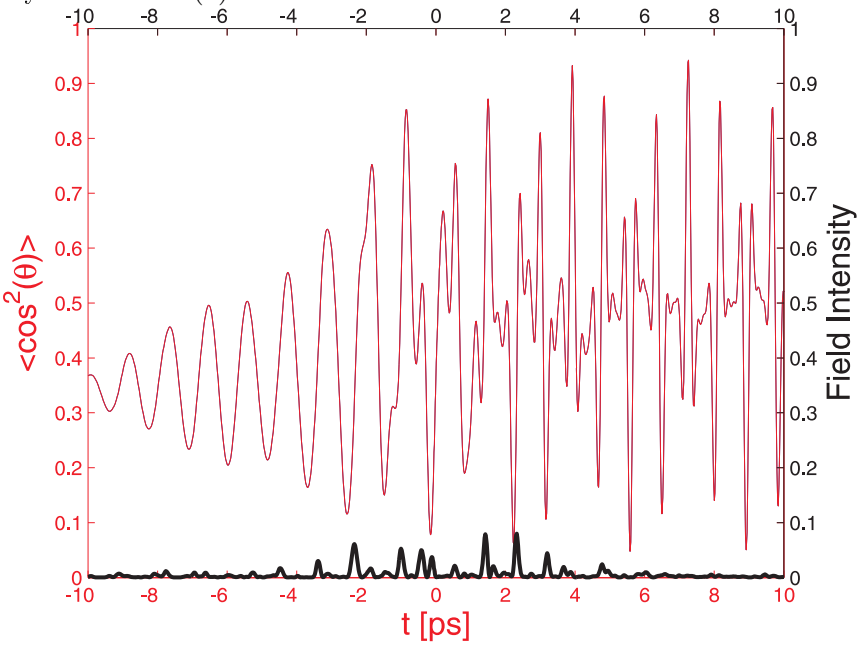

Figure A.9: A typical *sub-optimal* solution, obtained by the CMA-plain: A smooth exponential envelope of the revival structure is observed; Alignment yield: $< \cos^2(\theta) >= 0.9505$.

Sliding Window Fourier Transform applied to the revival structures of obtained solutions (e.g., thin-red alignment curve of Figure A.8). The values are log-scaled, and represent how high the rotational levels of the molecules are populated as a function of the interaction time.

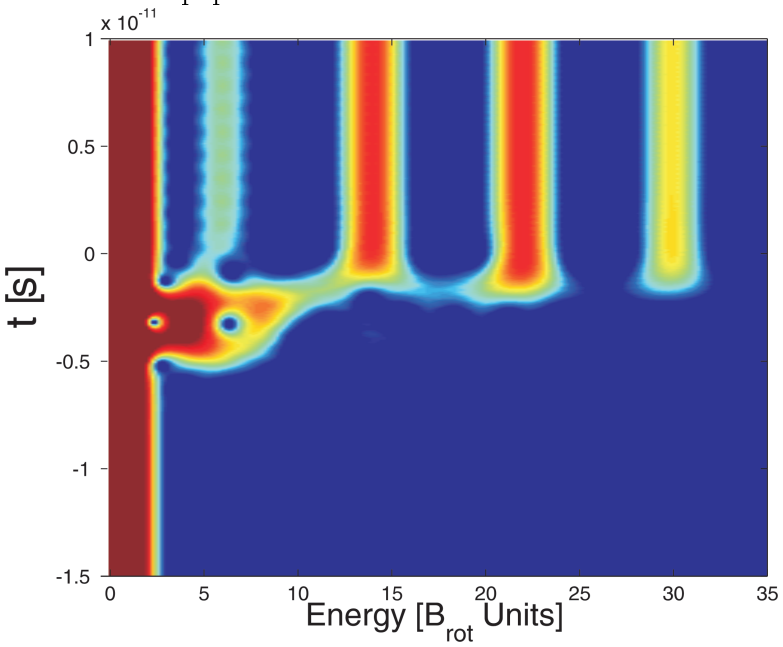

Figure A.10: DR2 with plain-parameterization: The 4^{th} rotational level, corresponding to $J = 6$, is mostly populated after the interaction.

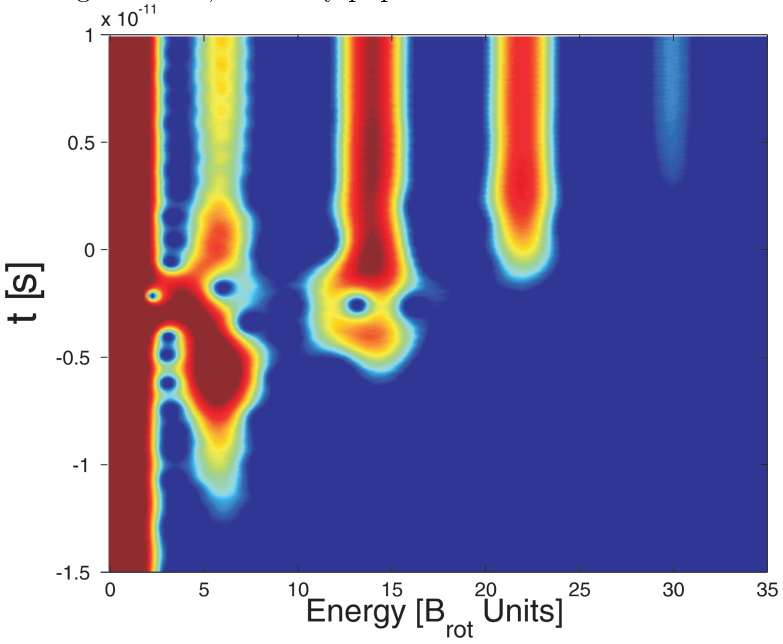

Figure A.11: CMA with plain-parameterization: All five first rotational levels are populated gradually after the interaction.

Sliding Window Fourier Transform applied to the revival structures of obtained solutions: continued.

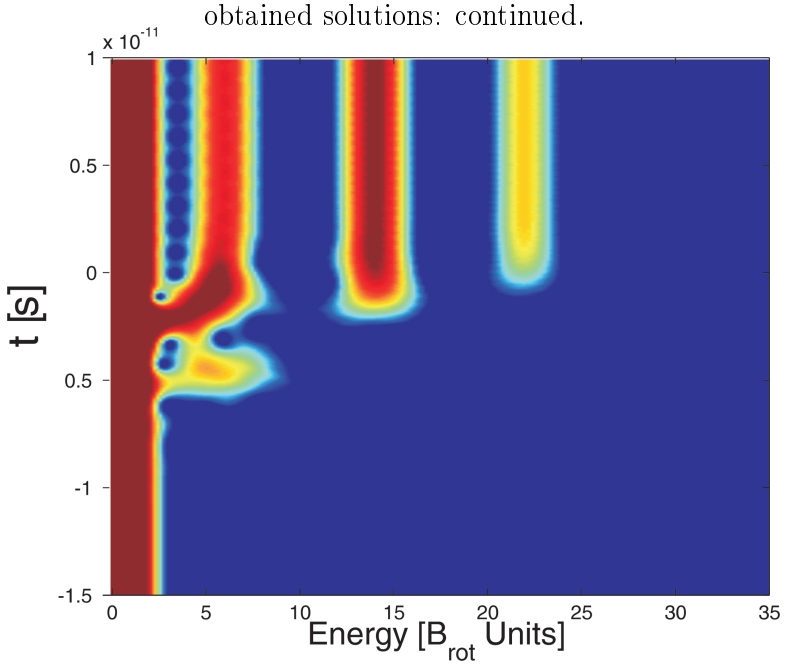

Figure A.12: DR2 with Hermite-parameterization: The four first rotational levels are populated gradually after the interaction.

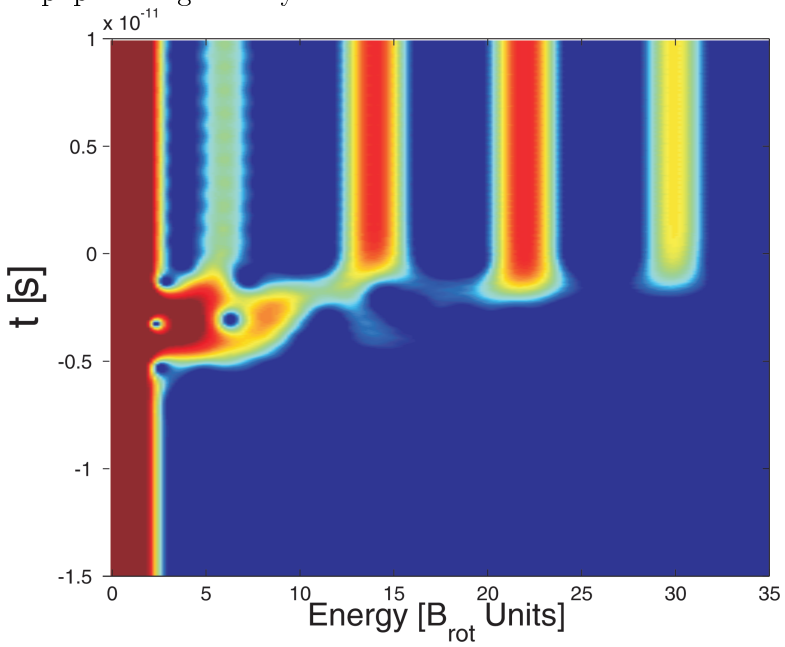

Figure A.13: CMA with Hermite-parameterization: The 4^{th} rotational level, corresponding to $J = 6$, is mostly populated after the interaction.

Attainment surfaces for the bi-criteria optimization of the Dynamic Molecular Alignment problem.

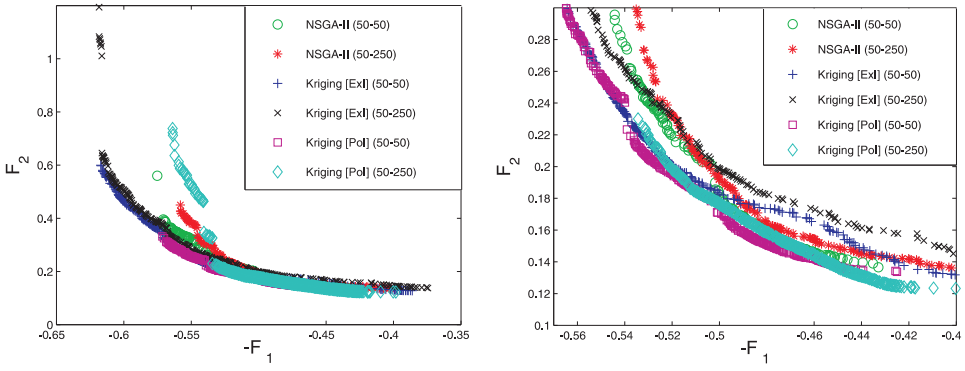

Figure A.14: Left: 20% Attainment Surfaces; Right: zoom-in.

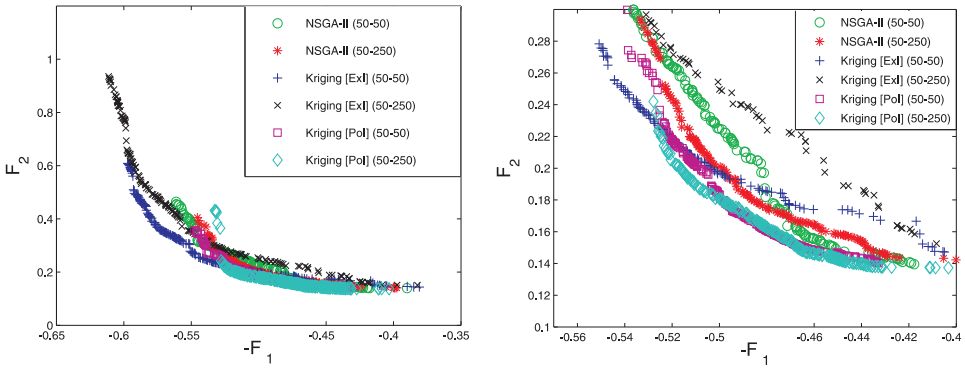

Figure A.15: Left: Median Attainment Surfaces; Right: zoom-in.

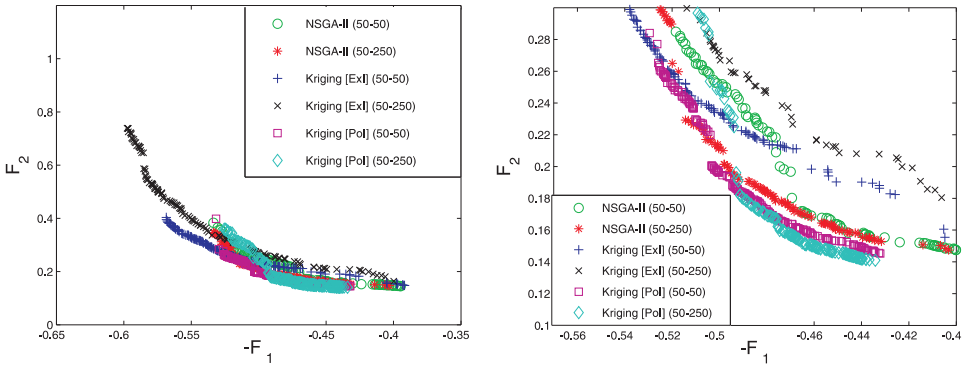

Figure A.16: Left: 80% Attainment Surfaces; Right: zoom-in.

Niching for the Dynamic Molecular Alignment problem; Best-niche results: Revival structure and the corresponding SWFT picture.

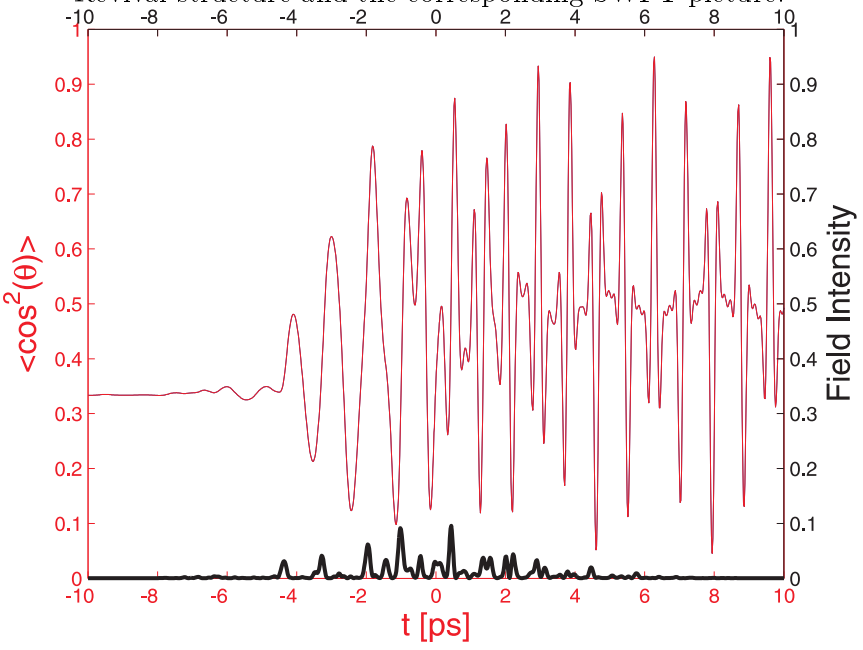

Figure A.17: Best niche: $< \cos^2(\theta) >= 0.9524$.

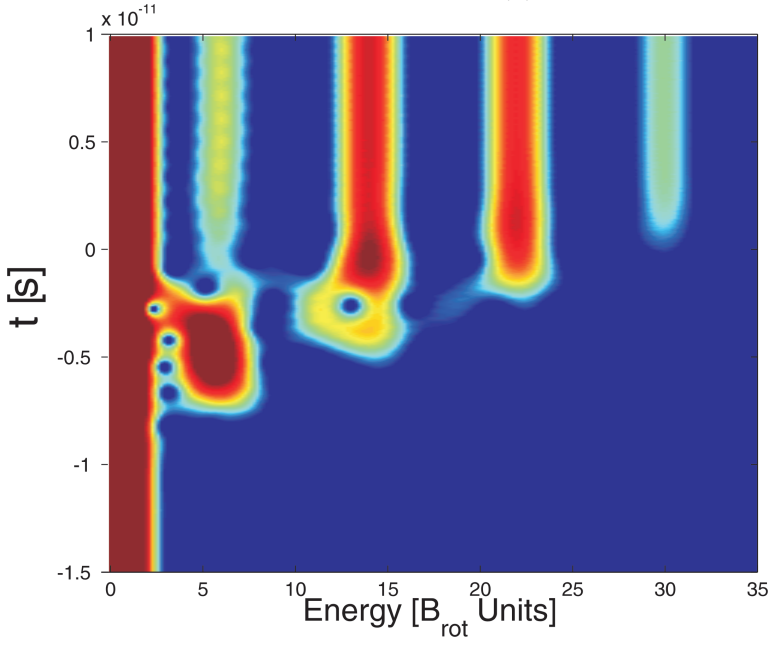

Figure A.18: SWFT picture of the best niche's solution.

Niching for the Dynamic Molecular Alignment problem; 2^{nd}-best niche results: Revival structure and the corresponding SWFT picture.

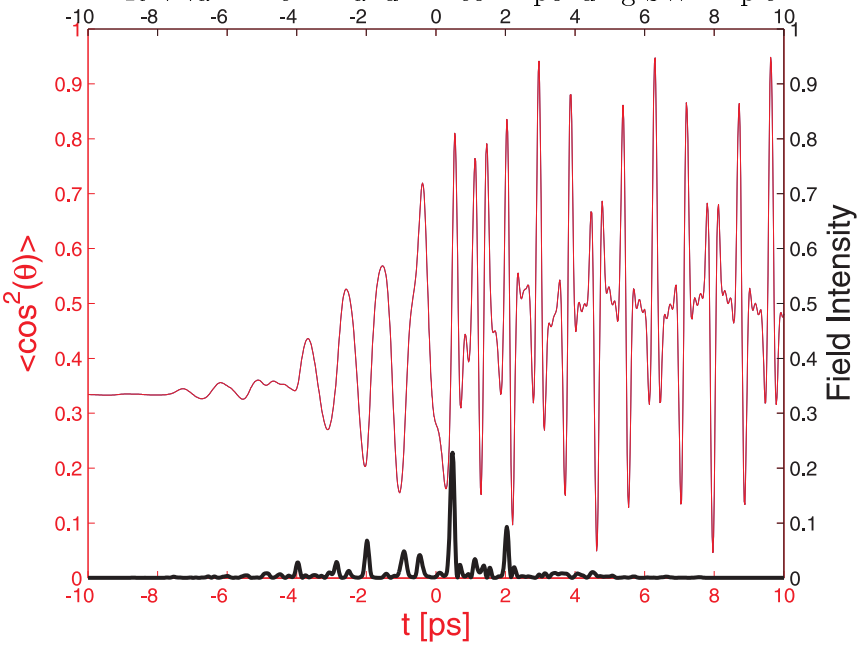

Figure A.19: 2^{nd}-best niche: $< \cos^2(\theta) >= 0.9513$.

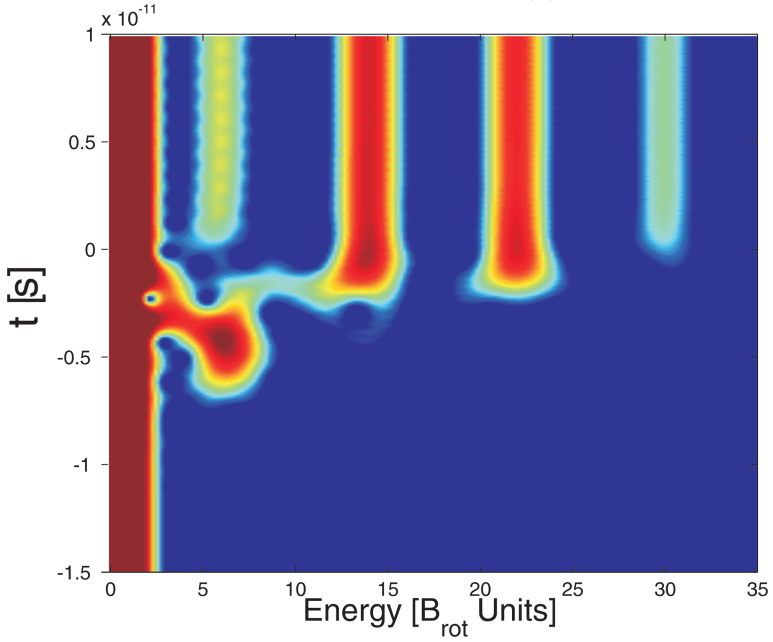

Figure A.20: SWFT picture of the 2^{nd}-best niche's solution.

Niching for the Dynamic Molecular Alignment problem; 3^{rd}-best niche results: Revival structure and the corresponding SWFT picture.

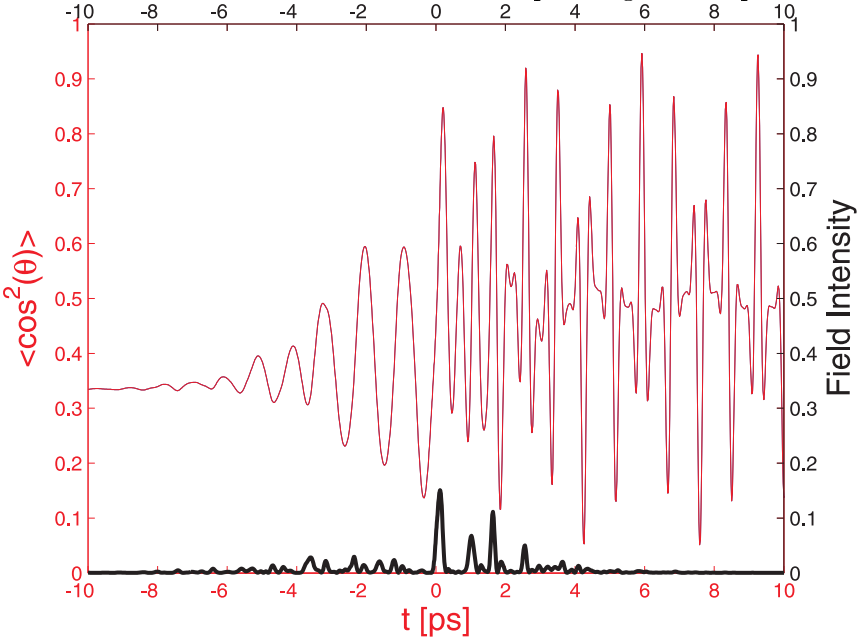

Figure A.21: 3^{rd}-best niche: $< \cos^2(\theta) >= 0.9466$.

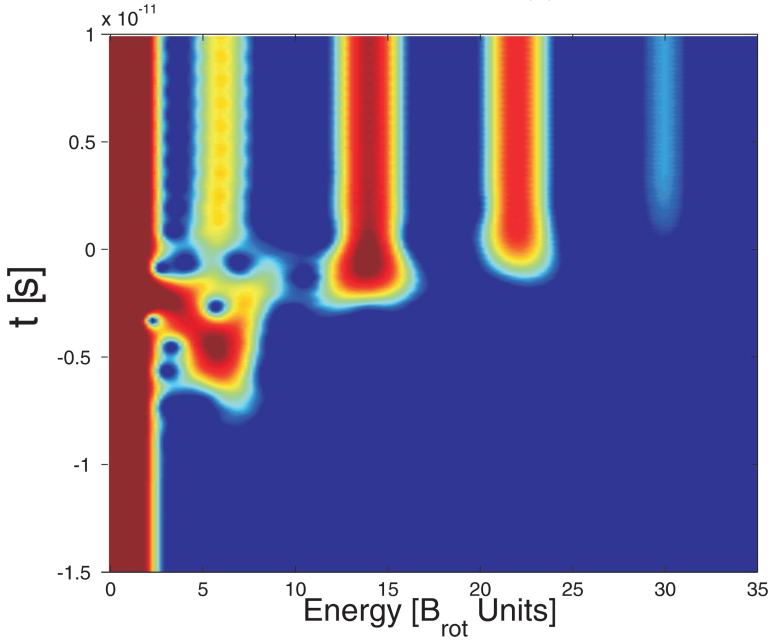

Figure A.22: SWFT picture of the 3^{rd}-best niche's solution.

Niching in the wavepacket space; A typical best-niche: Revival structure and the corresponding SWFT picture.

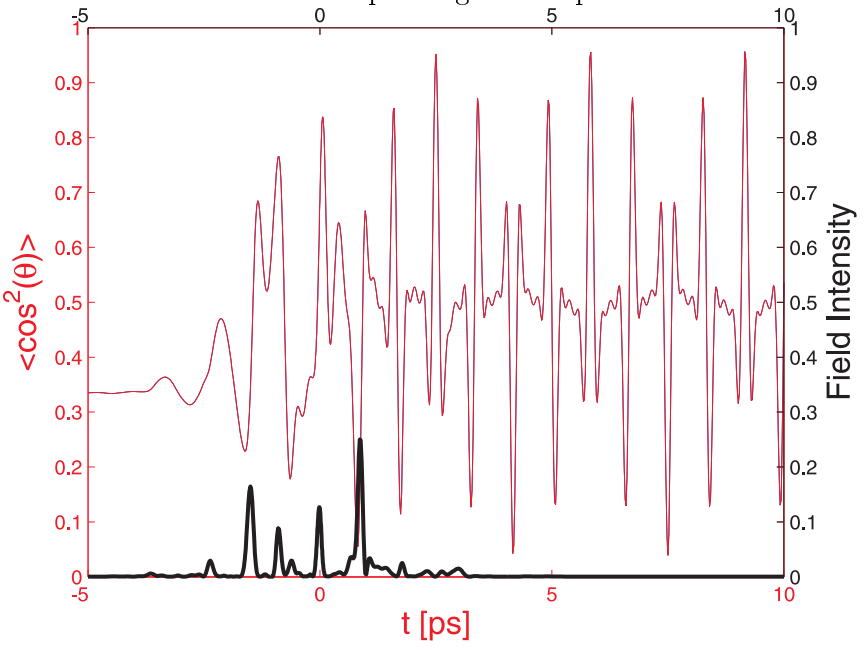

Figure A.23: Optimal niche: $\langle \cos^2(\theta) \rangle = 0.9596$.

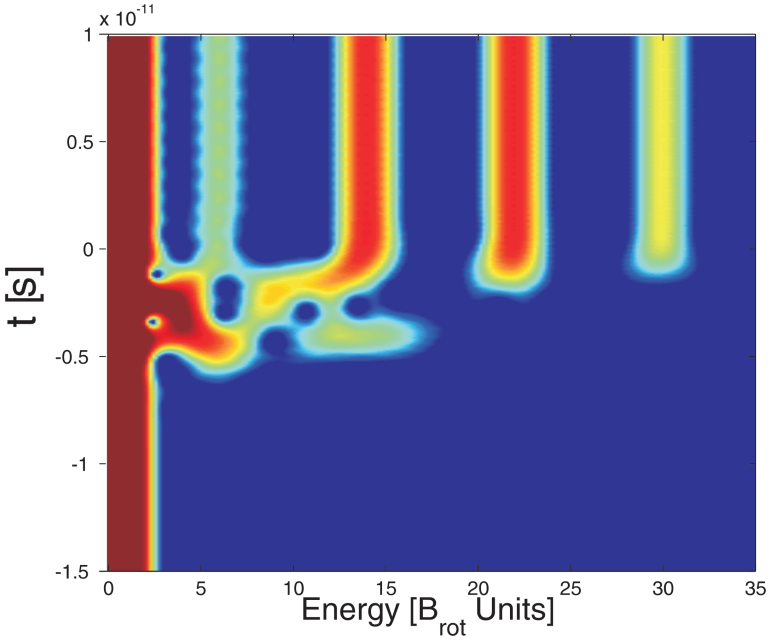

Figure A.24: Optimal niche: 4^{th} rotational level, corresponding to $J = 6$, is mostly populated after the interaction.

Niching in the wavepacket space; A typical 2^{nd}-best niche: Revival structure and the corresponding SWFT picture.

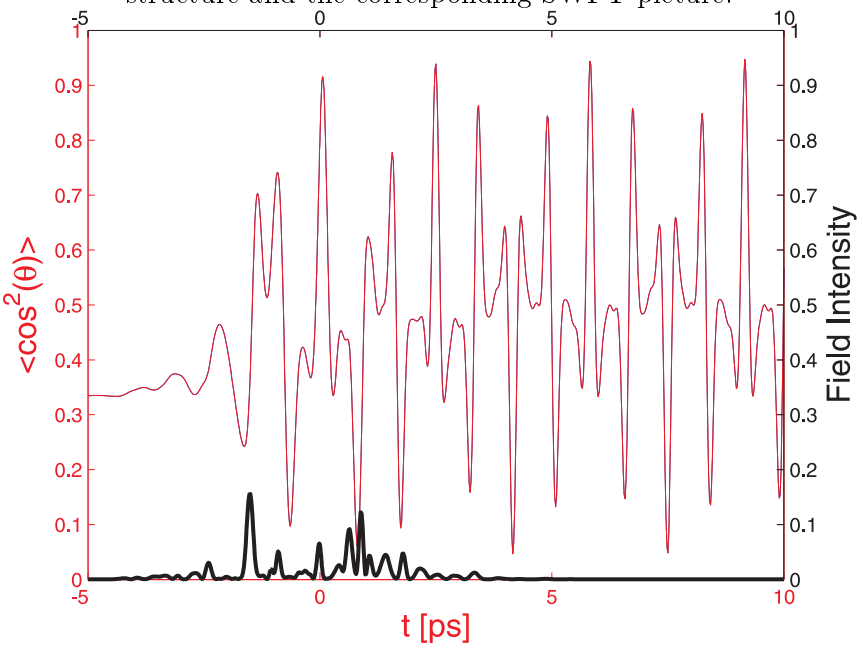

Figure A.25: Sub-optimal niche: $\langle \cos^2(\theta) \rangle = 0.9472$.

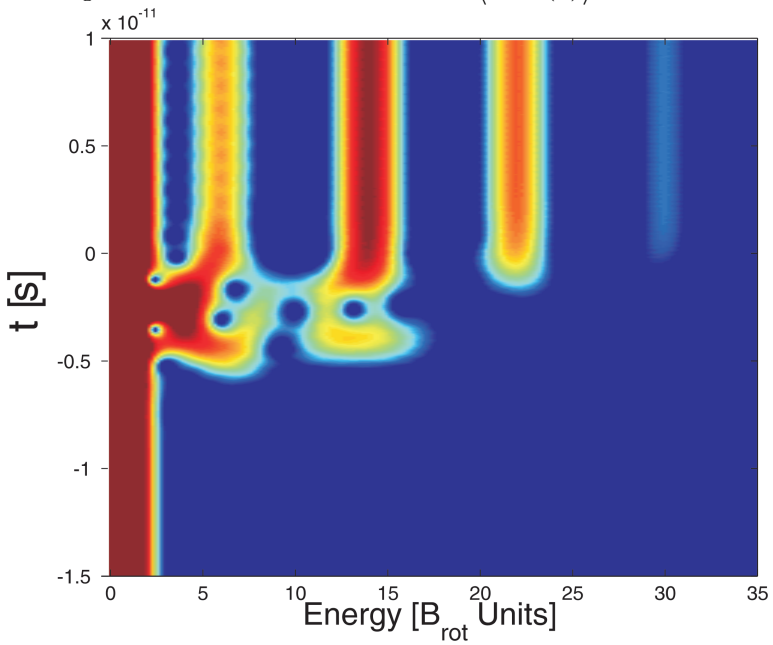

Figure A.26: Sub-optimal niche: 3^{rd} rotational level, corresponding to $J = 4$, is mostly populated after the interaction.

Mathematicians: You cannot work with them,
you cannot work without them.
John Doe

Appendix B

Complete-Basis Functions

Here is a brief summary of the fundamental mathematical concepts behind the Complete-Basis-Functions Parameterization, as presented in Section 9.2.2. This part is mainly based on Abramowitz [173] and Kaplan [174]. Let $f(x)$ be given in the interval $a \leq x \leq b$, and let

$$\xi_1(x), \xi_2(x), \ldots, \xi_k(x), \ldots \tag{B.1}$$

be functions which are all piecewise continuous in this interval.

The set $\{\xi_k(x)\}_{k=1}^{\infty}$ is called *complete* if it can span any piecewise continuous function $f(x)$, e.g.,

$$f(x) = \sum_{k=1}^{\infty} c_k \xi_k(x), \tag{B.2}$$

where the coefficients c_k are given by:

$$c_k = \frac{1}{B_k} \int_a^b f(x)\, \xi_k(x)\, dx, \quad B_k = \int_a^b [\xi_k(x)]^2\, dx \tag{B.3}$$

The convergence is guaranteed by the so-called *completeness theorem*. Explicitly, the series

$$R_m = \int_a^b \left(f(x) - \sum_{k=1}^{m} c_k \xi_k(x) \right)^2 dx = \int_a^b \left(f(x) - S_m(x) \right)^2 dx \tag{B.4}$$

converges to *zero* for sufficiently large m:

$$\lim_{m \to \infty} R_m = 0, \tag{B.5}$$

where we denoted the sequence of *partial sums* as $S_m(x)$:

$$S_m(x) = \sum_{k=1}^{m} c_k \xi_k(x) \tag{B.6}$$

By definition, the convergence of the series of functions is equivalent to the convergence of $S_m(x)$ to $f(x)$:

$$\lim_{m\to\infty} S_m(x) = f(x) \tag{B.7}$$

The Fourier (Trigonometric) Series

A *trigonometric series* is an expansion of a periodic function in terms of a sum of *sines* and *cosines*, making use of the orthogonality property of the harmonic functions. Without loss of generality, let us consider from now on the interval $[0, L]$. Let $f(x)$ be a single-valued function defined on that interval, then its *trigonometric series* or *trigonometric expansion* is given by:

$$\tilde{f}(x) = \frac{1}{2}a_0 + \sum_{k=1}^{\infty} a_k \cos\left(\frac{2\pi k}{L}\cdot x\right) + \sum_{k=1}^{\infty} b_k \sin\left(\frac{2\pi k}{L}\cdot x\right) \tag{B.8}$$

If the coefficients a_k and b_k satisfy certain conditions, then the series is called a *Fourier series*.
If $f(x)$ is periodic with period L, and has continuous first and second derivatives for all x in the interval, it is guaranteed that the trigonometric series of $f(x)$ will converge uniformly to $f(x)$ for all x; This is referred to as satisfying the *Dirichlet* conditions. We shall refer in this study to the *trigonometric* series as the *Fourier* series.

Other Sets of Functions

If one is indeed interested in periodic functions, there is no natural alternative but using the trigonometric series. However, if one is concerned with other representations of a general function over a given interval, a great variety of other sets of functions is available, e.g.:

- ***Legendre polynomials***, $P_k(x)$:

$$\begin{aligned} P_k(x) = \frac{(2k-1)(2k-3)\cdots 1}{k!}\Bigg\{x^k - \frac{k(k-1)}{2(k-1)}x^{k-2} + \\ + \frac{k(k-1)(k-2)(k-3)}{2\cdot 4(2k-1)(2k-3)}x^{k-4} - \cdots\Bigg\} \end{aligned} \tag{B.9}$$

  which can also be defined via Rodrigues' formula:

$$P_0(x) = 1 \qquad P_k(x) = \frac{1}{2^k k!}\frac{d^k}{dx^k}\left(x^2-1\right)^k,\ k = 1, 2, \ldots \tag{B.10}$$

  If $f(x)$ satisfies the *Dirichlet conditions* mentioned earlier, then there will exist a Legendre series expansion for it in the interval $-1 < x < 1$. For illustration, the first 10 *Legendre polynomials* are plotted in Figure B.1.

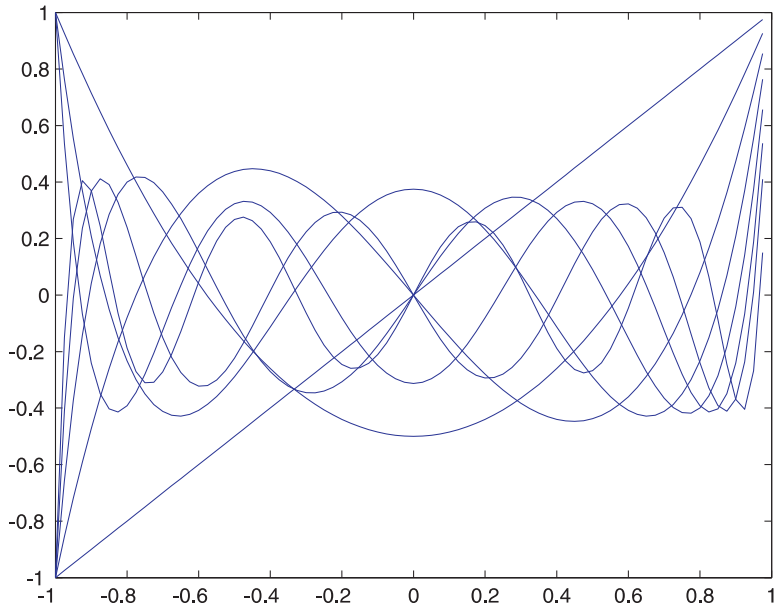

Figure B.1: The First 10 *Legendre* Polynomials.

- ***Bessel Function of the First Kind and of Order*** l, $J_l(x)$:

$$J_l(x) = \sum_{k=0}^{\infty} \frac{(-1)^k \, x^{l+2k}}{2^{l+2k} \cdot k! \cdot \Gamma(l+k+1)} \tag{B.11}$$

with $\Gamma(\alpha)$ as defined in Eq. 1.36. Given a fixed $l \geq 0$, the functions $\{\sqrt{x} J_l(\lambda_{lk} x)\}_{k=1}^{\infty}$ form an orthogonal complete system over the interval $0 \leq x \leq 1$.

- ***Hermite polynomials***, $H_k(x)$:

$$H_k(x) = (-1)^k \exp\{x^2\} \frac{d^k}{dx^k} \left(\exp\{-x^2\}\right), \; k = 0, 1, \ldots \tag{B.12}$$

The *Hermite polynomials* form a complete set of functions over the infinite interval $-\infty < x < \infty$, with respect to the weight function $\exp\left(-\frac{1}{2}x^2\right)$.

- ***Chebyshev polynomials of the First Kind***, $T_k(x)$:

$$T_k(x) = \frac{k}{2} \sum_{r=0}^{\lfloor k/2 \rfloor} \frac{(-1)^r}{k-r} \binom{k-r}{r} (2x)^{k-2r}, \; k = 0, 1, \ldots \tag{B.13}$$

The *Chebyshev polynomials of the First Kind* form a complete set of functions over the interval $[-1, 1]$ with respect to the weight function $\frac{1}{\sqrt{1-x^2}}$.

Higher Dimensions

An expansion by means of a complete set of functions can be generalized for higher dimensions. For illustration, let us consider the two-dimensional case of the *trigonometric series*. The functions $\cos(\frac{2\pi k}{L}\cdot x)\cdot\cos(\frac{2\pi l}{L}\cdot y)$, $\sin(\frac{2\pi k}{L}\cdot x)\cdot\cos(\frac{2\pi l}{L}\cdot y)$, $\cos(\frac{2\pi k}{L}\cdot x)\cdot\sin(\frac{2\pi l}{L}\cdot y)$, and $\sin(\frac{2\pi k}{L}\cdot x)\cdot\sin(\frac{2\pi l}{L}\cdot y)$ form an orthonormal complete system of functions in the box $[(0,0),(0,L),(L,0),(L,L)]$. Given a function in that domain, $f(x,y)$, its expansion can then be written in the form:

$$\begin{aligned} f(x,y) = \sum_{k=0}^{\infty}\sum_{l=0}^{\infty}\lambda_{kl}\cdot\Big\{ & a_{kl}\cos(\frac{2\pi k}{L}x)\cos(\frac{2\pi l}{L}y)+ \\ & +b_{kl}\sin(\frac{2\pi k}{L}x)\cos(\frac{2\pi l}{L}y)+c_{kl}\cos(\frac{2\pi k}{L}x)\sin(\frac{2\pi l}{L}y)+ \\ & +d_{kl}\sin(\frac{2\pi k}{L}x)\sin(\frac{2\pi l}{L}y)\Big\} \end{aligned} \tag{B.14}$$

Corollary

An infinite series of complete basis functions converges to any "reasonably well behaving" function. Hence, it is straightforward to approximate a given function with a finite series of those functions, i.e., by cutting its tail from a certain point. In principle, the sum $S_m(x)$ (Eq. B.6) can always be found to a desired degree of accuracy by adding up enough terms of the series. For practical applications, the corollary is that every function can be approximated using a series of complete basis functions, to whatever desired or practical accuracy. Moreover, this corollary can be easily generalized to any desired dimension.

Bibliography

[1] T. Bäck, *Evolutionary Algorithms in Theory and Practice*. New York, NY, USA: Oxford University Press, 1996.

[2] L. J. Fogel, *Artificial Intelligence through Simulated Evolution*. New York, NY, USA: John Wiley, 1966.

[3] J. H. Holland, "Outline for a Logical Theory of Adaptive Systems," *Journal of the ACM (JACM)*, vol. 9, no. 3, pp. 297–314, 1962.

[4] ——, *Adaptation in Natural and Artificial Systems*. Ann Arbor: The University of Michigan Press, 1975.

[5] H.-P. Schwefel, *Evolution and Optimum Seeking*. New York, NY, USA: John Wiley & Sons, Inc., 1995.

[6] I. Rechenberg, *Evolutionsstrategies: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Stuttgart, Germany: Frommann-Holzboog Verlag, 1973.

[7] H.-G. Beyer and H.-P. Schwefel, "Evolution Strategies a Comprehensive Introduction," *Natural Computing: An International Journal*, vol. 1, no. 1, pp. 3–52, 2002.

[8] A. Törn and A. Zilinskas, *Global Optimization*, ser. Lecture Notes in Computer Science. Springer, 1987, vol. 350.

[9] I. Zang and M. Avriel, "On Functions whose Local Minima are Global," *JOTA*, vol. 16, pp. 183–190, 1975.

[10] ——, "A Note on Functions whose Local Minima are Global," *JOTA*, vol. 18, pp. 556–559, 1976.

[11] D. Whitley, K. E. Mathias, S. B. Rana, and J. Dzubera, "Evaluating Evolutionary Algorithms," *Artificial Intelligence*, vol. 85, no. 1-2, pp. 245–276, 1996.

[12] T. Bäck, G. Rudolph, and H.-P. Schwefel, "Evolutionary Programming and Evolution Strategies: Similarities and Differences," in *Proceedings*

of the second Annual Conference on Evolutionary Programming. La Jolla, CA, USA: Evolutionary Programming Society, 1993, pp. 11–22.

[13] H.-P. Schwefel, "Kybernetische Evolution als Strategie der experimentellen Forschung in der Strömungstechnik," Master's thesis, Technical University of Berlin, 1965.

[14] I. Rechenberg, "Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution," Ph.D. dissertation, Technical University of Berlin, 1971.

[15] H.-G. Beyer, *The Theory of Evolution Strategies.* Heidelberg: Springer, 2001.

[16] N. Hansen and A. Ostermeier, "Completely Derandomized Self-Adaptation in Evolution Strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001.

[17] C. Igel, T. Suttorp, and N. Hansen, "A Computational Efficient Covariance Matrix Update and a (1+1)-CMA for Evolution Strategies," in *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2006.* New York, NY, USA: ACM Press, 2006, pp. 453–460.

[18] W. Gottschalk, *Allgemeine Genetik.* Stuttgart: Georg Thieme Verlag, 1989.

[19] H.-P. Schwefel, "Collective Phenomena in Evolutionary Systems," in *Problems of Constancy and Change – The Complementarity of Systems Approaches to Complexity, Proc. 31st Annual Meeting*, P. Checkland and I. Kiss, Eds., vol. 2. Budapest: Int'l Soc. for General System Research, 1987, pp. 1025–1033.

[20] T. Bäck, U. Hammel, and H.-P. Schwefel, "Evolutionary Computation: Comments on the History and Current State," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 3–17, 1997.

[21] H.-P. Schwefel, "Evolutionsstrategie und numerische Optimierung," Dr.-Ing. Thesis, Technical University of Berlin, Department of Process Engineering, 1975.

[22] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning.* Reading, MA: Addison Wesley, 1989.

[23] H.-G. Beyer, "An Alternative Explanation for the Manner in which Genetic Algorithms Operate," *BioSystems*, vol. 41, no. 1, pp. 1–15, 1997.

[24] G. Rudolph, "On Correlated Mutations in Evolution Strategies," in *Parallel Problem Solving from Nature - PPSN II.* Amsterdam: Elsevier, 1992, pp. 105–114.

[25] A. Ostermeier, A. Gawelczyk, and N. Hansen, "A Derandomized Approach to Self Adaptation of Evolution Strategies," *Evolutionary Computation*, vol. 2, no. 4, pp. 369–380, 1994.

[26] ——, "A Derandomized Approach to Self Adaptation of Evolution Strategies," TU Berlin, Tech. Rep. TR-93-003, 1993.

[27] ——, "Step-Size Adaptation Based on Non-Local Use of Selection Information," in *Parallel Problem Solving from Nature - PPSN III*, ser. Lecture Notes in Computer Science, vol. 866. Springer, 1994, pp. 189–198.

[28] N. Hansen, A. Ostermeier, and A. Gawelczyk, "On the Adaptation of Arbitrary Normal Mutation Distributions in Evolution Strategies: The Generating Set Adaptation," in *Proceedings of the Sixth International Conference on Genetic Algorithms (ICGA6).* San Francisco, CA: Morgan Kaufmann, 1995, pp. 57–64.

[29] N. Hansen and A. Ostermeier, "Adapting Arbitrary Normal Mutation Distributions in Evolution Strategies: the Covariance Matrix Adaptation," in *Proceedings of the 1996 IEEE International Conference on Evolutionary Computation.* Piscataway, NJ: IEEE, 1996, pp. 312–317.

[30] D. Lindley, *Introduction to Probability and Statistics from a Bayesian Viewpoint: Inference.* London, UK: Cambridge University Press, 1965, vol. 2.

[31] N. Hansen and S. Kern, "Evaluating the CMA Evolution Strategy on Multimodal Test Functions," in *Parallel Problem Solving from Nature - PPSN V*, ser. Lecture Notes in Computer Science, vol. 1498. Amsterdam: Springer, 1998, pp. 282–291.

[32] H.-G. Beyer, "Toward a Theory of Evolution Strategies: On the Benefit of Sex - the $(\mu/mu_I, \lambda)$ Theory," *Evolutionary Computation*, vol. 3, no. 1, pp. 81–110, 1995.

[33] C. Igel, N. Hansen, and S. Roth, "Covariance Matrix Adaptation for Multi-objective Optimization," *Evolutionary Computation*, vol. 15, no. 1, pp. 1–28, 2007.

[34] C. A. Coello Coello, "A Survey of Constraint Handling Techniques used with Evolutionary Algorithms," Laboratorio Nacional de Informática Avanzada, Xalapa, Veracruz, México, Tech. Rep. Lania-RI-99-04, 1999.

[35] A. Auger and N. Hansen, "Performance Evaluation of an Advanced Local Search Evolutionary Algorithm," in *Proceedings of the 2005 Congress on Evolutionary Computation CEC-2005*. Piscataway, NJ, USA: IEEE Press, 2005, pp. 1777–1784.

[36] T. Bäck, "Selective Pressure in Evolutionary Algorithms: A Characterization of Selection Mechanisms," in *Proceedings of the First IEEE Conference on Evolutionary Computation (ICEC'94), Orlando FL*, Z. Michalewicz, J. D. Schaffer, H.-P. Schwefel, D. B. Fogel, and H. Kitano, Eds. Piscataway, NJ, USA: IEEE Press, 1994, pp. 57–62.

[37] S. Mahfoud, "Niching Methods for Genetic Algorithms," Ph.D. dissertation, University of Illinois at Urbana Champaign, 1995.

[38] G. Avigad, A. Moshaiov, and N. Brauner, "Concept-Based Interactive Brainstorming in Engineering Design," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 8, no. 5, pp. 454–459, 2004.

[39] ——, "Interactive Concept-based Search using MOEA: The Hierarchical Preferences Case," *International Journal of Computational Intelligence*, vol. 2, no. 3, pp. 182–191, 2005.

[40] J. J. Cristiano, C. C. White, and J. K. Liker, "Application of Multiattribute Decision Analysis to Quality Function Deployment for Target Setting," *IEEE Transactions on Systems, Man, and Cybernetics: Part C*, vol. 31, no. 3, pp. 366–382, 2001.

[41] S. Freeman and J. C. Herron, *Evolutionary Analysis*. Redwood City, CA, USA: Benjamin Cummings, 3rd Edition, 2003.

[42] C. R. Darwin, *The Origin of Species: By Means of Natural Selection or The Preservation of Favoured Races in the Struggle for Life*. New York, NY, USA: Bantam Classics, 1999.

[43] R. A. Fisher, "Darwinian Evolution of Mutations," *Eugenics Review*, vol. 14, pp. 31–34, 1922.

[44] S. Wright, "Evolution in Mendelian Populations," *Genetics*, vol. 16, pp. 97–159, 1931.

[45] M. Kimura, *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press, 1983.

[46] S. M. Scheiner and C. J. Goodnight, "The Comparison of Phenotypic Plasticity and Genetic Variation in Populations of the Grass Danthonia Spicata," *Evolution*, vol. 38, no. 4, pp. 845–855, 1984.

[47] A. Bradshaw, "Evolutionary Significance of Phenotypic Plasticity in Plants," *Advanced Genetics*, vol. 13, pp. 115–155, 1965.

[48] B. A. McPheron, D. C. Smith, and S. H. Berlocher, "Genetic Differences between Host Races of Rhagoletis Pomonella," *Nature*, vol. 336, pp. 64–66, 1988.

[49] K. Tsui, "An Overview of Taguchi Method and Newly Developed Statistical Methods for Robust Design," *IIE Transactions*, vol. 24, pp. 44–57, 1992.

[50] M. Lunacek and D. Whitley, "The Dispersion Metric and the CMA Evolution Strategy," in *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO-2006*. New York, NY, USA: ACM, 2006, pp. 477–484.

[51] J. Doye, R. Leary, M. Locatelli, and F. Schoen, "Global Optimization of Morse Clusters by Potential Energy Transformations," *INFORMS, Journal On Computing*, vol. 16, no. 4, pp. 371–379, 2004.

[52] J. N. Kapur and H. K. Kesavan, *Entropy Optimization Principles with Applications*. Harcourt Brace Jovanovich, 1992.

[53] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 79–86, 1951.

[54] K. Deb and D. E. Goldberg, "An Investigation of Niche and Species Formation in Genetic Function Optimization," in *Proceedings of the third international conference on Genetic Algorithms*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989, pp. 42–50.

[55] H.-G. Beyer, "On the Dynamics of GAs without Selection," in *Foundations of Genetic Algorithms 5*, W. Banzhaf and C. Reeves, Eds. San Francisco, CA: Morgan Kaufmann, 1999, pp. 5–26.

[56] L. Schönemann, M. Emmerich, and M. Preuss, "On the Extiction of Sub-Populations on Multimodal Landscapes," in *Proc. of the Int'l Conf. on Bioinspired optimization Methods and their Applications, BIOMA 2004*. Jožef Stefan Institute, Slovenia, 2004, pp. 31–40.

[57] M. Preuss, L. Schönemann, and M. Emmerich, "Counteracting Genetic Drift and Disruptive Recombination in $(\mu + /, \lambda)$-EA on Multimodal Fitness Landscapes," in *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2005*. New York, NY, USA: ACM Press, 2005, pp. 865–872.

[58] H.-G. Beyer, E. Brucherseifer, W. Jakob, H. Pohlheim, B. Sendhoff, and T. B. To, "Evolutionary Algorithms - Terms and Definitions," http://ls11-www.cs.uni-dortmund.de/people/beyer/EA-glossary/, 2002.

[59] M. Preuss, "Niching Prospects," in *Proc. of the Int'l Conf. on Bioinspired optimization Methods and their Applications, BIOMA 2006*. Jožef Stefan Institute, Slovenia, 2006, pp. 25–34.

[60] G. Singh and K. Deb, "Comparison of Multi-Modal Optimization Algorithms based on Evolutionary Algorithms," in *Proceedings of the 2006 annual conference on Genetic and evolutionary computation, GECCO 2006*. New York, NY, USA: ACM Press, 2006, pp. 1305–1312.

[61] D. E. Goldberg and J. Richardson, "Genetic algorithms with sharing for multimodal function optimization," in *Proceedings of the Second International Conference on Genetic Algorithms and Their Application*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc., 1987, pp. 41–49.

[62] X. Yin and N. Germany, "A Fast Genetic Algorithm with Sharing using Cluster Analysis Methods in Multimodal Function Optimization," in *Proceedings of the International Conference on Articial Neural Nets and Genetic Algorithms, Innsbruck, Austria, 1993*. Springer, 1993, pp. 450–457.

[63] M. Jelasity, "UEGO, an Abstract Niching Technique for Global Optimization," in *Parallel Problem Solving from Nature - PPSN V*, ser. Lecture Notes in Computer Science, vol. 1498. Amsterdam: Springer, 1998, pp. 378–387.

[64] B. Miller and M. Shaw, "Genetic Algorithms with Dynamic Niche Sharing for Multimodal Function Optimization," in *Proceedings of the 1996 IEEE International Conference on Evolutionary Computation (ICEC'96)*, New York, NY, USA, 1996, pp. 786–791.

[65] A. Petrowski, "A Clearing Procedure as a Niching Method for Genetic Algorithms," in *Proceedings of the 1996 IEEE International Conference on Evolutionary Computation (ICEC'96)*, New York, NY, USA, 1996, pp. 798–803.

[66] A. D. Cioppa, C. D. Stefano, and A. Marcelli, "On the Role of Population Size and Niche Radius in Fitness Sharing," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 6, pp. 580–592, 2004.

[67] K. A. de Jong, "An Analysis of the Behavior of a Class of Genetic Adaptive Systems," Ph.D. dissertation, University of Michigan, Ann Arbor, 1975.

[68] K. Deb and S. Tiwari, "Omni-optimizer: A Procedure for Single and Multi-objective Optimization," in *Evolutionary Multi-Criterion Optimization, Third International Conference, EMO 2005*, ser. Lecture Notes in Computer Science, vol. 3410. Springer, 2005, pp. 47–61.

[69] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. NJ, USA: Prentice Hall, 1999.

[70] V. Hanagandi and M. Nikolaou, "A Hybrid Approach to Global Optimization using a Clustering Algorithm in a Genetic Search Framework," *Computers and Chemical Engineering*, vol. 22, no. 12, pp. 1913–1925, 1998.

[71] J. Branke, *Evolutionary Optimization in Dynamic Environments*. Norwell, MA, USA: Kluwer Academic Publishers, 2001.

[72] J. Gan and K. Warwick, "Dynamic Niche Clustering: A Fuzzy Variable Radius Niching Technique for Multimodal Optimisation in GAs," in *Proceedings of the 2001 Congress on Evolutionary Computation CEC2001*. COEX, World Trade Center, 159 Samseong-dong, Gangnam-gu, Seoul, Korea: IEEE Press, 2001, pp. 215–222.

[73] O. Aichholzer, F. Aurenhammer, B. Brandtstätter, T. Ebner, H. Krammer, and C. Magele, "Niching Evolution Strategy with Cluster Algorithms," in *Proceedings of the 9th Biennial IEEE Conference on Electromagnetic Field Computations*. IEEE Press, 2000, p. 137.

[74] F. Streichert, G. Stein, H. Ulmer, and A. Zell, "A Clustering Based Niching EA for Multimodal Search Spaces," in *Proceedings of the International Conference Evolution Artificielle*, ser. Lecture Notes in Computer Science, vol. 2936. Springer, 2003, pp. 293–304.

[75] S. Ando, J. Sakuma, and S. Kobayashi, "Adaptive Isolation Model using Data Clustering for Multimodal Function Optimization," in *Proceedings of the 2005 conference on Genetic and evolutionary computation, GECCO 2005*. New York, NY, USA: ACM Press, 2005, pp. 1417–1424.

[76] H. Ramalhinho-Lourenco, O. C. Martin, and T. Stützle, "Iterated Local Search," Department of Economics and Business, Universitat Pompeu Fabra, Economics Working Papers 513, Nov. 2000.

[77] A. Auger and N. Hansen, "A Restart CMA Evolution Strategy With Increasing Population Size," in *Proceedings of the 2005 Congress on Evolutionary Computation CEC-2005*. Piscataway, NJ, USA: IEEE Press, 2005, pp. 1769–1776.

[78] D. Beasley, D. R. Bull, and R. R. Martin, "A Sequential Niche Technique for Multimodal Function Optimization," *Evolutionary Computation*, vol. 1, no. 2, pp. 101–125, 1993.

[79] P. B. Grosso, "Computer Simulations of Genetic Adaptation: Parallel Subcomponent Interaction in a Multilocus Model," Ph.D. dissertation, University of Michigan, Ann Arbor, MI, USA, 1985.

[80] P. Adamidis, "Review of Parallel Genetic Algorithms Bibliography," Automation and Robotics Lab., Dept. of Electrical and Computer Eng., Aristotle University of Thessaloniki, Greece, Tech. Rep., 1994.

[81] W. Martin, J. Lienig, and J. Cohoon, "Island (Migration) Models: Evolutionary Algorithms based on Punctuated Equilibria," in *Handbook of Evolutionary Computation*, T. Bäck, D. B. Fogel, and Z. Michalewicz, Eds. Oxford University Press, New York, and Institute of Physics Publ., Bristol, 1997, pp. C6.3:1–16.

[82] W. M. Spears, "Simple Subpopulation Schemes," in *Proceedings of the 3rd Annual Conference on Evolutionary Programming*. World Scientific, 1994, pp. 296–307.

[83] K. Deb and W. M. Spears, "Speciation Methods," in *The Handbook of Evolutionary Computation*, T. Bäck, D. Fogel, and Z. Michalewicz, Eds. IOP Publishing and Oxford University Press, 1997.

[84] R. K. Ursem, "Multinational Evolutionary Algorithms," in *Proceedings of the 1999 Congress on Evolutionary Computation (CEC 1999)*. Piscataway NJ: IEEE Press, 1999, pp. 1633–1640.

[85] C. Stoean, M. Preuss, R. Gorunescu, and D. Dumitrescu, "Elitist Generational Genetic Chromodynamics - a New Radii-Based Evolutionary Algorithm for Multimodal Optimization," in *Proceedings of the 2005 Congress on Evolutionary Computation (CEC'05)*. Piscataway NJ: IEEE Press, 2005, pp. 1839–1846.

[86] R. E. Smith and C. Bonacina, "Mating Restriction and Niching Pressure: Results from Agents and Implications for General EC," in *Proceedings of the 2003 Conference on Genetic and Evolutionary Computation, GECCO 2003*, ser. Lecture Notes on Computer Science, vol. 2724. Chicago: Springer-Verlag, 2003, pp. 1382–1393.

[87] O. Kramer and H.-P. Schwefel, "On Three New Approaches to Handle Constraints within Evolution Strategies," *Natural Computing: An International Journal*, vol. 5, no. 4, pp. 363–385, 2006.

[88] O. M. Shir and T. Bäck, "Niching in Evolution Strategies," LIACS, Leiden University, Tech. Rep. TR-2005-01, 2005.

[89] ——, "Niching in Evolution Strategies," in *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO-2005*. New York, NY, USA: ACM Press, 2005, pp. 915–916.

[90] ——, "Dynamic Niching in Evolution Strategies with Covariance Matrix Adaptation," in *Proceedings of the 2005 Congress on Evolutionary Computation CEC-2005*. Piscataway, NJ, USA: IEEE Press, 2005, pp. 2584–2591.

[91] S. W. Mahfoud, "A Comparison of Parallel and Sequential Niching Methods," in *Proceedings of the Sixth International Conference on Genetic Algorithms*, L. Eshelman, Ed. San Francisco, CA: Morgan Kaufmann, 1995, pp. 136–143.

[92] N. Hansen, A. Gawelczyk, and A. Ostermeier, "Sizing the Population with respect to the Local Progress in $(1, \lambda)$-Evolution Strategies - A Theoretical Analysis," in *Proceedings of the 1995 IEEE International Conference on Evolutionary Computation*. New York, NY, USA: IEEE, 1995, pp. 312–317.

[93] P. N. Suganthan, N. Hansen, J. J. Liang, K. Deb, Y. P. Chen, A. Auger, and S. Tiwari, "Problem Definitions and Evaluation Criteria for the CEC 2005 Special Session on Real-Parameter Optimization," `http://www.ntu.edu.sg/home/EPNSugan/`, Nanyang Technological University, Singapore, Tech. Rep., 2005.

[94] O. M. Shir, "Niching in Evolution Strategies," Master's thesis, Leiden University, 2005.

[95] C. Stoean, M. Preuss, R. Stoean, and D. Dumitrescu, "Disburdening the Species Conservation Evolutionary Algorithm of Arguing with Radii," in *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2007*. New York, NY, USA: ACM Press, 2007, pp. 1420–1427.

[96] P. Giorgi, C.-P. Jeannerod, and G. Villard, "On the Complexity of Polynomial Matrix Computations," in *ISSAC '03: Proceedings of the 2003 international symposium on Symbolic and algebraic computation*. New York, NY, USA: ACM Press, 2003, pp. 135–142.

[97] M. Ehrgott, *Multicriteria Optimization*, 2nd ed. Berlin: Springer, 2005.

[98] E. Zitzler, "Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications," Ph.D. dissertation, ETH Zurich, Switzerland, 1999.

[99] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. New York: Wiley, 2001.

[100] M. Emmerich, N. Beume, and B. Naujoks, "An EMO Algorithm using the Hypervolume Measure as Selection Criterion," in *Proc. Evolutionary Multi-Criterion Optimization: Third Int'l Conference (EMO 2005)*, ser. Lecture Notes in Computer Science, vol. 3410. Berlin: Springer, 2005, pp. 62–76.

[101] M. Emmerich, "Single- and Multi-objective Evolutionary Design Optimization Assisted by Gaussian Random Field Metamodels," Ph.D. dissertation, University of Dortmund, Germany, 2005.

[102] G. Rudolph, B. Naujoks, and M. Preuss, "Capabilities of EMOA to Detect and Preserve Equivalent Pareto Subsets," in *Proc. Evolutionary Multi-Criterion Optimization: Fourth Int'l Conference (EMO 2007)*, ser. Lecture Notes in Computer Science, vol. 4403. Berlin: Springer, 2007, pp. 36–50.

[103] J. Horn, N. Nafpliotis, and D. E. Goldberg, "A Niched Pareto Genetic Algorithm for Multiobjective Optimization," in *Proceedings of the First IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence*. Piscataway, New Jersey: IEEE Service Center, 1994, pp. 82–87.

[104] A. Toffolo and E. Benini, "Genetic Diversity as an Objective in Multi-Objective Evolutionary Algorithms," *Evolutionary Computation*, vol. 11, no. 2, pp. 151–167, 2003.

[105] E. Zitzler, K. Deb, and L. Thiele, "Comparison of Multiobjective Evolutionary Algorithms: Empirical Results," *Evolutionary Computation*, vol. 8, no. 2, pp. 173–195, 2000.

[106] T. Bäck, "Self-Adaptation," in *Handbook of Evolutionary Computation*, T. Bäck, D. B. Fogel, and Z. Michalewicz, Eds. Oxford University Press, New York, and Institute of Physics Publ., Bristol, 1997, pp. C7.1:1–15.

[107] D. Büche, S. D. Müller, and P. Koumoutsakos, "Self-Adaptation for Multi-objective Evolutionary Algorithms," in *EMO*, ser. Lecture Notes in Computer Science, vol. 2632. Springer, 2003, pp. 267–281.

[108] J.-W. Klinkenberg, M. Emmerich, A. Deutz, O. M. Shir, and T. Bäck, "Accelerating SMS-EMOA for Problems with Time-Expensive Evaluations using Kriging, Self-Adaptation, and MPI," in *Multiple Criteria Decision Making for Sustainable Energy and Transportation Systems: Proceedings of MCDM 2008, The 19th International Conference on Multiple Criteria Decision Making*, ser. Lecture Notes in Economics and Mathematical Systems, vol. 634. Heidelberg, Germany: Springer Physica-Verlag, 2010, pp. 301–312.

[109] M. Preuss, B. Naujoks, and G. Rudolph, "Pareto Set and EMOA Behavior for Simple Multimodal Multiobjective Functions," in *Parallel Problem Solving from Nature, PPSN IX*, ser. Lecture Notes in Computer Science, vol. 4193. Springer, 2006, pp. 513–522.

[110] M. Emmerich and A. Deutz, "Test Problems Based on Lamé Superspheres," in *EMO-2007*, ser. Lecture Notes in Computer Science, vol. 4403. Springer, 2007, pp. 922–936.

[111] D. Tannor and S. Rice, "Control of Selectivity of Chemical Reaction via Control of Wave Packet Evolution," *Chem. Phys.*, vol. 83, 1985.

[112] P. Brumer and M. Shapiro, "Control of Unimolecular Reactions using Coherent Light," *Chem. Phys. Lett.*, vol. 126, no. 6, 1986.

[113] R. S. Judson and H. Rabitz, "Teaching Lasers to Control Molecules," *Phys. Rev. Lett.*, vol. 68, no. 10, pp. 1500–1503, Mar 1992.

[114] W. S. Warren, H. Rabitz, and M. Dahleh, "Coherent Control of Quantum Dynamics: The Dream Is Alive," *Science*, vol. 259, pp. 1581–1589, Mar 1993.

[115] H. Rabitz, R. de Vivie-Riedle, M. Motzkus, and K. Kompa, "Whither the Future of Controlling Quantum Phenomena?" *Science*, vol. 288, pp. 824–828, May 2000.

[116] P. Nuernberger, G. Vogt, T. Brixner, and G. Gerber, "Femtosecond Quantum Control of Molecular Dynamics in the Condensed Phase," *Phys Chem Chem Phys.*, vol. 9, no. 20, pp. 2470–2497, 2007.

[117] C. J. Bardeen, V. V. Yakovlev, K. R. Wilson, S. D. Carpenter, P. M. Weber, and W. S. Warren, "Feedback Quantum Control of Molecular Electronic Population Transfer," *Chem. Phys. Lett.*, vol. 280, no. 1-2, 1997.

[118] T. Weinacht, J. Ahn, and P. Bucksbaum, "Controlling the Shape of a Quantum Wavefunction," *Nature*, vol. 397, no. 233, 1999.

[119] D. Zeidler, S. Frey, W. Wohlleben, M. Motzkus, F. Busch, T. Chen, W. Kiefer, and A. Materny, "Optimal Control of Ground-State Dynamics in Polymers," *Chem. Phys.*, vol. 116, no. 12, Mar 2002.

[120] R. Levis, G. M. Menkir, and H. Rabitz, "Selective Bond Dissociation and Rearrangement with Optimally Tailored, Strong-Field Laser Pulses," *Science*, vol. 292, pp. 709–713, Apr 2001.

[121] T. Brixner and G. Gerber, "Femtosecond Polarization Pulse Shaping," *Opt. Lett.*, vol. 26, no. 8, pp. 557–559, 2001.

[122] J. Kunde, B. Baumann, S. Arlt, F. Morier-Genoud, U. Siegner, and U. Keller, "Adaptive Feedback Control of Ultrafast Semiconductor Nonlinearities," *Appl. Phys. Lett.*, vol. 77, no. 7, 2000.

[123] J. L. Herek, W. Wohlleben, R. J. Cogdell, D. Zeidler, and M. Motzkus, "Quantum Control of Energy Flow in Light Harvesting," *Nature*, vol. 417, no. 533, 2002.

[124] A. P. Peirce, M. A. Dahleh, and H. Rabitz, "Optimal Control of Quantum-Mechanical Systems: Existence, Numerical Approximation, and Applications," *Phys. Rev. A*, vol. 37, no. 12, Jun 1988.

[125] S. Shi and H. Rabitz, "Quantum Mechanical Optimal Control of Physical Observables in Microsystems," *Chem. Phys.*, vol. 92, no. 364, Jan 1990.

[126] F. Schwabl, *Quantum Mechanics*. Berlin: Springer, 2002.

[127] T.-S. Ho and H. Rabitz, "Why do Effective Quantum Controls Appear Easy to Find?" *Journal of Photochemistry and Photobiology A: Chemistry*, vol. 180, no. 3, Jun 2006.

[128] R. Chakrabarti and H. Rabitz, "Quantum Control Landscapes," *International Reviews in Physical Chemistry*, vol. 26, no. 4, pp. 671–735, 2007.

[129] H. Rabitz, M. M. Hsieh, and C. M. Rosenthal, "Quantum Optimally Controlled Transition Landscapes," *Science*, vol. 303, pp. 1998–2001, Mar 2004.

[130] H. Rabitz, T.-S. Ho, M. Hsieh, R. Kosut, and M. Demiralp, "The Topology of Optimally Controlled Quantum Mechanical Transition Probability Landscapes," *Phys. Rev. A*, vol. 74, no. 1, p. 012721, July 2006.

[131] M. Demiralp and H. Rabitz, "Optimally Controlled Quantum Molecular Dynamics: A Perturbation Formulation and the Existence of Multiple Solutions," *Phys. Rev. A*, vol. 47, no. 2, pp. 809–816, Feb 1993.

[132] ——, "Optimally Controlled Quantum Molecular Dynamics: The Effect of Nonlinearities on the Magnitude and Multiplicity of Control-Field Solutions," *Phys. Rev. A*, vol. 47, no. 2, pp. 831–837, Feb 1993.

[133] A. Rothman, T.-S. Ho, and H. Rabitz, "Observable-Preserving Control of Quantum Dynamics over a Family of Related Systems," *Phys. Rev. A*, vol. 72, no. 2, p. 023416, 2005.

[134] ——, "Exploring the Level Sets of Quantum Control Landscapes," *Phys. Rev. A*, vol. 73, no. 5, p. 053401, 2006.

[135] R. Chakrabarti, R. Wu, and H. Rabitz, "Computational Complexity of Quantum Optimal Control Landscapes," 2007, to be submitted.

[136] H. T. Siegelmann, A. Ben-Hur, and S. Fishman, "Computational Complexity for Continuous Time Dynamics," *Phys. Rev. Lett.*, vol. 83, no. 7, pp. 1463–1466, Aug 1999.

[137] T. C. Weinacht and P. H. Bucksbaum, "Using Feedback for Coherent Control of Quantum Systems," *Journal of Optics B*, vol. 4, no. 3, 2002.

[138] M. Shapiro and P. Brumer, "Coherent Control of Atomic, Molecular, and Electronic Processes," *Advances in Atomic, Molecular, and Optical Physics*, vol. 42, no. 287, 2000.

[139] R. Kosloff, S. Rice, P. Gaspard, S. Tersigni, and D. Tannor, "Wavepacket Dancing: Achieving Chemical Selectivity by Shaping Light Pulses," *Chem. Phys.*, vol. 139, 1989.

[140] M. Roth, "Optimal Dynamic Discrimination in the Laboratory," Ph.D. dissertation, Princeton University, 2007.

[141] J. Vaughan, T. Feurer, K. Stone, and K. Nelson, "Analysis of Replica Pulses in Femtosecond Pulse Shaping with Pixelated Devices," *Optics Express*, vol. 14, no. 3, pp. 1314–1328, 2006.

[142] R. Bracewell, *The Fourier Transform and Its Applications*. McGraw-Hill Book Company, 1965.

[143] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*, 2nd ed. Cambridge, UK: Cambridge University Press, 1992.

[144] D. Meshulach and Y. Silberberg, "Coherent Quantum Control of Two-Photon Transitions by a Femtosecond Laser Pulse," *Nature*, vol. 396, no. 239, 1998.

[145] ——, "Coherent Quantum Control of Multiphoton Transitions by Shaped Ultrashort Optical Pulses," *Phys. Rev. A*, vol. 60, no. 2, 1999.

[146] J. Roslund, M. Roth, and H. Rabitz, "Laboratory Observation of Quantum Control Level Sets," *Phys. Rev. A*, vol. 74, no. 4, p. 043414, 2006.

[147] O. M. Shir and T. Bäck, "The Second Harmonic Generation Case Study as a Gateway for ES to Quantum Control Problems," in *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO-2007*. New York, NY, USA: ACM Press, 2007, pp. 713–721.

[148] O. M. Shir, J. N. Kok, M. J. Vrakking, and T. Bäck, "Gaining Insight into Laser Pulse Shaping by Evolution Strategies," in *Proceedings of IWINAC-2007*, ser. Lecture Notes in Computer Science, vol. 4527. Springer, 2007, pp. 467–477.

[149] D. V. Arnold and H.-G. Beyer, "Local Performance of the $(\mu/\mu_I, \lambda)$-ES in a Noisy Environment," in *Foundations of Genetic Algorithms, 6*, W. Martin and W. Spears, Eds. San Francisco, CA: Morgan Kaufmann, 2001, pp. 127–141.

[150] D. Zeidler, S. Frey, K.-L. Kompa, and M. Motzkus, "Evolutionary Algorithms and their Application to Optimal Control Studies," *Phys. Rev. A*, vol. 64, no. 2, p. 023420, Jul 2001.

[151] R. Fanciulli, L. Willmes, J. Savolainen, P. van der Walle, T. Bäck, and J. L. Herek, "Evolution Strategies for Laser Pulse Compression," in *Proceedings of the International Conference Evolution Artificielle*, ser. Lecture Notes in Computer Science, vol. 4926. Springer, 2008, pp. 219–230.

[152] V. Beltrani, "Frequency Shaping and the Alignment Problem," Princeton University, Princeton NJ, USA, Tech. Rep., 2008.

[153] F. Rosca-Pruna and M. J. Vrakking, "Revival Structures in Picosecond Laser-Induced Alignment of I2 Molecules," *Journal of Chemical Physics*, vol. 116, no. 15, pp. 6579–6588, 2002.

[154] H. Stapelfeldt and T. Seideman, "Colloquium: Aligning Molecules with Strong Laser Pulses," *Rev. Mod. Phys.*, vol. 75, no. 2, pp. 543–557, Apr 2003.

[155] B. Friedrich and D. Herschbach, "Steric Proficiency of Polar 2-Sigma Molecules in Congruent Electric and Magnetic Fields," *Phys. Chem. Chem. Phys.*, vol. 2, pp. 419–428, 2000.

[156] N. Hay, R. Velotta, M. Lein, R. de Nalda, E. Heesel, M. Castillejo, and J. P. Marangos, "High-Order Harmonic Generation in Laser-Aligned Molecules," *Phys. Rev. A*, vol. 65, no. 5, p. 053805, Apr 2002.

[157] T. Seideman, "Revival Structure of Aligned Rotational Wave Packets," *Phys. Rev. Lett.*, vol. 83, no. 24, pp. 4971–4974, Dec 1999.

[158] M. Leibscher, I. S. Averbukh, and H. Rabitz, "Molecular Alignment by Trains of Short Laser Pulses," *Phys. Rev. Lett.*, vol. 90, no. 21, p. 213001, May 2003.

[159] ——, "Enhanced Molecular Alignment by Short Laser Pulses," *Phys. Rev. A*, vol. 69, no. 1, p. 013402, 2004.

[160] C. M. Dion, A. Keller, and O. Atabek, "Optimally Controlled Field-Free Orientation of the Kicked Molecule," *Phys. Rev. A*, vol. 72, no. 2, p. 023402, 2005.

[161] K.-S. Leung and Y. Liang, "Evolution Strategies with a Fourier Series Auxiliary Function for Difficult Function Optimization," in *IDEAL*, ser. Lecture Notes in Computer Science, vol. 2690. Springer, 2003, pp. 303–312.

[162] C. Siedschlag, O. M. Shir, T. Bäck, and M. J. J. Vrakking, "Evolutionary Algorithms in the Optimization of Dynamic Molecular Alignment," *Optics Communications*, vol. 264, pp. 511–518, Aug 2006.

[163] A. Mitra and H. Rabitz, "Mechanistic Analysis of Optimal Dynamic Discrimination of Similar Quantum Systems," *J. Phys. Chem. A.*, vol. 108, p. 4778, 2004.

[164] T. Bäck, "On the Behavior of Evolutionary Algorithms in Dynamic Environments," in *Proceedings of the 1998 International Conference on Evolutionary Computation (ICEC'98)*. Piscataway, NJ, USA: IEEE Press, 1998, pp. 446–451.

[165] L. Schönemann, "Optimal Number of Evolution Strategies Mutation Step Sizes in Dynamic Environments," in *Proceedings of the 2005 conference on Genetic and evolutionary computation, GECCO 2005*. New York, NY, USA: ACM Press, 2005, pp. 923–924.

[166] D. V. Arnold and H.-G. Beyer, "Random Dynamics Optimum Tracking with Evolution Strategies," in *Parallel Problem Solving from Nature - PPSN VII*, ser. Lecture Notes in Computer Science, vol. 2439. Berlin: Springer, 2002, pp. 3–12.

[167] A. Ratle, "Accelerating the Convergence of Evolutionary Algorithms by Fitness Landscape Approximations," in *Parallel Problem Solving by Nature - PPSN V*, ser. Lecture Notes in Computer Science. Berlin: Springer-Verlag, 1998, pp. 87–96.

[168] M. El-Beltagy, P. Nair, and A. Keane, "Metamodelling Techniques for Evolutionary Optimisation of Computationally Expensive Problems: Promises and Limitations," in *Proceedings of the Genetic and Evolutionary Computation Conference, GECCO-1999*, W. Banzhaf, J. Daida, A. Eiben, M. Garzon, V. Honavar, M. Jakiela, and R. Smith, Eds. Morgan Kaufman, 1999, pp. 196–203.

[169] M. Emmerich, A. Giotis, M. Özdemir, T. Bäck, and K. Giannakoglou, "Metamodel-Assisted Evolution Strategies," in *Parallel Problem Solving from Nature - PPSN VII*, ser. Lecture Notes in Computer Science, vol. 2439. Berlin: Springer, 2002, pp. 361–370.

[170] M. Emmerich, "A Rigorous Analysis of Two Bi-Criteria Problem Families with Scalable Curvature of the Pareto Fronts," Leiden University, Tech. Rep., 2005.

[171] M. Emmerich, K. Giannakoglou, and B. Naujoks, "Single- and Multiobjective Evolutionary Optimization Assisted by Gaussian Random Field Metamodels," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 4, pp. 421–440, August 2006.

[172] K. Deb, M. Mohan, and S. Mishra, "A Fast Multiobjective Evolutionary Algorithm for Finding Well-Spread Pareto-Optimal Solutions," KanGAL, Kanpur, India, Tech. Rep. 2003002, 2003.

[173] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. National Bureau of Standards, Applied Math. Series 55, Dover Publications, 1965.

[174] W. Kaplan, *Advanced Calculus*, 3rd ed. Reading, MA, USA: Addison-Wesley, 1983.

Index

Samenvatting (Dutch)

Op alle niveaus van het dagelijks leven word je regelmatig geconfronteerd met systemen die in hun natuurlijke omgeving functioneren en daarbij een zekere mate van optimaal gedrag vertonen. Zulk optimaal gedrag vormt hierdoor een belangrijke inspiratiebron voor allerlei gebieden. Binnen het vakgebied Natural Computing is het de bedoeling berekeningstechnieken te ontwikkelen die zo goed mogelijk gebundelde verschijnselen uit de natuur benaderen, op basis waarvan deze technieken op hun beurt vaak heel goed presteren in informatieverwerkingsprocessen. Uit een lange lijst van natural-computing-deelgebieden zijn we in het bijzonder geïnteresseerd geraakt in het uitermate boeiende gebied van Organic Evolution - Organische Evolutie - en in zijn rekentegenhanger, het zogenoemde gebied van de Evolutionaire Algorithmen (EA). Door een optimalisatieprobleem naar een kunstmatig-biologische omgeving om te zetten, benaderen EA inderdaad bepaalde stukjes uit de Darwinistisch dynamica en streven die EA er daarbij naar, goed passende oplossingen te bereiken in termen van de probleemsituatie. Daarbij is een populatie van mogelijke oplossingen onderhevig aan kunstmatige, dat wil zeggen gesimuleerde variatie. Vervolgens overleven zulke mogelijke oplossingen een dergelijke simulatie op basis van concrete criteria voortvloeiend uit het gekozen selectiemechanisme.

De oorspronkelijke bedoeling van ons onderzoek was om bepaalde varianten van EA, Evolutionaire Strategieën geheten (ES), uit te breiden naar deelpopulaties van pilot-oplossingen die parallel toegroeien naar verschillende oplossingen van het probleem. Dit idee is gebaseerd op een begrip uit de evolutietheorie, organic speciation, de organisch-evolutionaire ontwikkeling per soort. Waar het hier op neer komt is, dat de manier van denken binnen Natural Computing dieper dient in te gaan op theorieën uit de Evolutionaire Biologie en in het licht van de gewenste evolutionaire soortontwikkeling creatieve oplossingen dient te vinden voor de kunstmatige populatie. De zogenoemde Niche-technieken vormen de uitbreiding van EA naar deelpopulaties met ieder hun eigen evolutionaire ontwikkeling. Zij zijn al bestudeerd vanaf het begin van de EA en wel voornamelijk binnen de populaire variant van de Genetisch Algorithmen (GA). Naast de theoretische uitdaging om zulke technieken te ontwerpen, daarbij krachtig ondersteund door biologie-geinspireerde motivatie, zijn er ook goede gronden vanuit de praktijk om

dit te proberen. Met name vanuit het vakgebied van de besliskunde, dat al rechtstreeks baat heeft bij de opkomst van het gebied van globale optimalizering, wordt duidelijk dat er dringend behoefte is aan meervoudigheid van verschillende optimale oplossingen. In een ideaal geval zullen deze meervoudige oplossingen, zoals verkregen uit de optimalisatie-aanpak, onderling een hoge mate van diversiteit vertonen en zullen zij verschillende conceptuele ontwerpen voor oplossingen vertegenwoordigen.

Terwijl we de bedoeling hadden dit onderzoek voornamelijk te richten op niche-technieken in ES, waren we er tevens vanaf het begin op uit de algorithmes waar we op uit zouden komen, te gebruiken voor praktische toepassingen in het pas ontsloten gebied van Quantum Control (QC). Dit laatste biedt een enorme verscheidenheid aan veel-dimensionale continue optimalisatieproblemen, zowel op theoretisch als op experimenteel niveau. In dit opzicht heeft QC de potentie een ideale testomgeving te zijn voor evolutionaire optimalisatie, in het bijzonder voor niche-aanpakken. Dit komt door enkele opmerkelijke karakteristieken van zogeheten QC-landschappen. Typerend voor zulke landschappen is, zoals bewezen in QC-theorie, dat ze oneindig veel optimale oplossingen hebben. Door dit alles is de combinatie van niche-onderzoek en zijn toepassingen op QC-landschappen voor ons heel intrigerend. Toen we deze overweldigende, ideale rijkdom aan oplossingen binnen QC-landschappen dan ook eenmaal hadden opgemerkt, hebben we besloten een op zichzelf staand deel van dit proefschrift te wijden aan Quantum Control. In symbolische zin vormt deze interdisciplinaire studie daarmee een gesloten natural-computing-cirkel, waarin biologisch-georiënteerd onderzoek van organische evolutie, met name die binnen een soort, bijdraagt aan de ontwikkeling van rekenmethoden om toepassingen binnen de natuurkunde als geheel op te lossen en in het bijzonder binnen Quantum Control. Naar ons idee wordt deze symbolische zienswijze nog verder versterkt door het stochastische karakter van EA. Aldus, biologisch geïnspireerd door Evolutionaire Biologie in het algemeen en door organic speciation in het bijzonder en tevens op scherp door de drijfveer meervoudig optimale oplossingen te willen vinden voor het beter nemen van beslissingen in praktijsituaties, doen we in deze studie verslag van onze reis, vertrokken vanuit diversiteit in de natuur, beland bij conceptuele ontwerpen in Quantum Control.

Dit proefschrift bestaat uit twee delen: Deel I introduceert een niche-framework voor een klasse van state-of-the-art ES-algorithmen, namelijk de Derandomized Evolution Strategies (DES), en gaat in op het uitproberen van de voorgestelde algorithmen in kunstmatige landschappen. Deel II geeft een overzicht van de voornaamste aspecten van Quantum Control binnen de algemene context van globale functie-optimalisatie. Vervolgens worden de experimentele waarnemingen van de DES algorithmen gepresenteerd en tevens die van de voorgestelde niche-algorithmen zoals toegepast op verschillende QC-systemen, zowel in laboratoriumsituaties als op verschillende niveaus van numerieke simulatie.

There is only one way to win 100m free-style swimming:
Start at your maximum speed, and slowly accelerate.
Alon 'Krembo' Sagiv, **MIVTZA SAVTA**; Dror Shaul

Acknowledgments

Let me begin by thanking Thomas Bäck, my Professor, who threw me in to deep water very early, supported me all along at every possible personal level, and has been constantly an inspiring raw model. *Thomas*, you have become a true friend, and I will always cherish this period of working with you.

I would like to thank Joost Kok, for voluntarily playing a dedicated role of my Dutch sponsor, and for going with me through every step in this journey. *Joost*, thank you for everything along the way, I really appreciate your personal support.

I thank Michael Emmerich, who supervised my research with devotion that research proposals cannot describe. *Michael*, the long fruitful discussions, the pedantic yet extremely helpful iterations ("necessary evil", as you used to call them), and the endless willingness to assist were a crucial contribution to this work, and I am grateful for that.

I thank Marc Vrakking, the *Physics godfather* of this work, for the assistance at every level of work, and for his remarkable patience in the joint work with Computer Scientists.

I would like to thank Christian Siedschlag, with whom I had the pleasure to collaborate in the first half of my PhD period. Christian had an important contribution to the early foundations of this work as well as to the personal nature of my PhD candidacy, and I thank him for that.

Although my period as a Visiting Scholar Research Collaborator at Princeton University spanned only four months out of a total research period of three and a half years, it was yet one of the most significant phases of my PhD. I would like to devote special thanks to Herschel Rabitz, who hosted me in his group, dedicated hours of his time for priceless discussions, and contributed an enormous deal to my learning experience at Princeton.
Hersch, thank you very much indeed for everything during the amazing period of Fall 2007. I would also like to thank Tak-San Ho, Jason Dominy, Vinny Beltrani and Jon Roslund, of the Rabitz group, for their contribution to this work.

The Dutch Kingdom generously hosted me, allowing me to carry out my research in ideal conditions, and I am sincerely grateful for that.
I am indebted to FOM, the Dutch Foundation for Research on Fundamental Matter, and its personnel at Utrecht, for their constant support. Especially, I would like to thank them for financing my research period at Princeton University. I also thank Wim Aspers and Marloes van der Nat of the LIACS.

I would like to thank Luuk Groenewegen for the fascinating extra curricular discussions, especially in the fields of linguistics and literature. Also, I thank him for his assistance in compiling the Dutch Samenvatting.

I will always associate the beginning of my PhD period with my work at LUSM. I would like to dedicate special thanks to the wonderful family of LUSM, for their love and support, and especially to Judith de Wilde and Hans Borgman.

Last, but certainly not least, I would like to thank my beloved family in Jerusalem. *IMA*, *ABA*, *Shai*, and *Yael*, I could not have done this without your endless love and support, which you kept speeding-up to me at every hour of every single day, no matter where I was around the globe.

Nota bene: I thank Dr. Evil and Austin Powers for providing me with the inspiration in my work on the laser.