UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Matemática
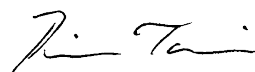
Lucas Resende

Differential Evolution: Theoretical Results on Convergence and Parameter Choosing

Belo Horizonte

2020

Lucas Resende

Differential Evolution: Theoretical Results on Convergence and Parameter Choosing

Final Version

Dissertation submitted to the examination board assigned by the Graduate Program in Mathematics of the Institute of Exact Sciences - ICEX of the Federal University of Minas Gerais, as a partial requirement to obtain a Master Degree in Mathematics.

Advisor: PhD. Ricardo Hiroshi Caldeira Takahashi

Belo Horizonte
2020

ATA DA DEFESA DE DISSERTAÇÃO DO ALUNO LUCAS RESENDE, REGULARMENTE MATRICULADO NO PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA, DO INSTITUTO DE CIÊNCIAS EXATAS, DA UNIVERSIDADE FEDERAL DE MINAS GERAIS, REALIZADA NO DIA 21 DE JULHO DE 2020.

Aos vinte e um dias do mês de julho de 2020, às 10h00, em reunião pública virtual Zoom pelo link:

https://us02web.zoom.us/j/87956948164?pwd=TDNrMWI0bS8ySzVvVXBHalFDbmVpdz09

(conforme mensagem eletrônica da Pró-Reitoria de Pós-Graduação de 26/03/2020, com orientações para a atividade de defesa de dissertação durante a vigência da Portaria nº 1819), reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pelo Colegiado do Programa de Pós-Graduação em Matemática, para julgar a defesa de dissertação do aluno **Lucas Resende**, intitulada: "*Differential Evolution: Theoretical Results on Convergence and Parameter Choosing*", requisito final para obtenção do Grau de mestre em Matemática. Abrindo a sessão, o Senhor Presidente da Comissão, Prof. Ricardo Hiroshi Caldeira Takahashi, após dar conhecimento aos presentes do teor das normas regulamentares do trabalho final, passou a palavra ao aluno para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do aluno. Após a defesa, os membros da banca examinadora reuniram-se reservadamente sem a presença do aluno e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: o aluno foi considerado aprovado sem ressalvas e por unanimidade. O resultado final foi comunicado publicamente ao aluno pelo Senhor Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 21 de julho de 2020.
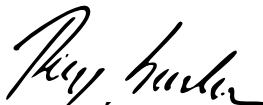
PROF. DR. RICARDO HIROSHI CALDEIRA TAKAHASHI
Orientador (UFMG)

PROF. DR. CARLOS MANUEL MIRA DA FONSECA
Examinador (Engenharia Informática - Universidade de Coimbra/Portugal)

PROFA. DRA. ELIZABETH FIALHO WANNER
Examinadora (Computação - CEFET/MG)

PROF. DR. RÉMY DE PAIVA SANCHIS
Examinador (UFMG)

Acknowledgments

I would also like to thank, in chronological order, the various advisors I have had so far: Michel, for introducing me to academic life; Sokol, for sparking my interest in probability; Denise and Pablo, for their attention and advice; and Ricardo, for allowing me to develop this work with autonomy and freedom.

I also remember my friends P.O. and Luiz, and my ex-girlfriend Gabriela, for their support and patience. And, most importantly, I express my gratitude to my mother, my godmother, and my neighbor for their continuous dedication to my education.

"Eu tô te explicando
Prá te confundir
Eu tô te confundindo
Prá te esclarecer
Tô iluminado
Prá poder cegar
Tô ficando cego
Prá poder guiar"
(Tom Zé)

Resumo

O *Differential Evolution* é um algoritmo evolutivo clássico para otimização. Possui apenas três parâmetros, é fácil de implementar e poderoso. Embora seja amplamente utilizado, com muitos artigos propondo variantes e resultados experimentais, apenas alguns artigos buscam resultados analíticos. Este trabalho tem dois objetivos: um teórico e um prático. O objetivo teórico é ampliar a compreensão da dinâmica do algoritmo, para isso, começamos expondo alguns resultados da literatura e em seguida alguns resultados originais. O objetivo prático é criar um método adequado para escolher os parâmetros do algoritmo. Esse método é derivado dos resultados analíticos e testado com sucesso em relação a funções de teste típicas.

Palavras-chave: evolução diferencial; otimização; escolha de parâmetros; algoritimos estocásticos; algoritimos genéticos.

Abstract

The Differential Evolution is a long-established evolutionary algorithm for optimization. It has only three parameters, it is easy to implement and powerful. While being widely used, with many papers proposing variants and experimental results, only a few articles seek analytical results. This work has two objectives: a theoretical one and a practical one. The theoretical objective is to enlarge the understanding of the dynamics of the algorithm, for this, we begin exposing some results from the literature and then some original results. The practical objective is to create a suitable method to choose the parameters of the algorithm. That method is derived from the analytical results and successfully tested against typical test functions.

Keywords: differential evolution; optimization; parameter choosing; stochastic algorithms; genetic algorithms.

Contents

Chapter 1

The algorithm

1.1 Basic algorithm and variants

Proposed by Price and Storn in a series of papers [11, 15, 16], the Differential Evolution is a long-established algorithm [10] that aims to optimize functions $f : X \to P$ where X is a continuous domain, called search space, and P is a partially ordered set (in this text we study the case $X = \mathbb{R}^n$ and $P = \mathbb{R}$). In this entire monograph the objective function is $f : \mathbb{R}^n \to \mathbb{R}$ and the dimension of the search space is n.

An execution of the algorithm starts by uniformly selecting a set

$$X^0 = \{x_1^0, \cdots, x_N^0\} \subset \mathbb{R}^n$$

with N points, known as the initial population, from a domain region where one expects to find the global minimum. After the initialization the algorithm iterates producing a sequence of populations $X^0 \to X^1 \to X^2 \to \cdots \to X^{T_f}$ until a stop criterion is reached. The population $X^t = \{x_1^t, \cdots, x_N^t\} \subset X$ is the t-th generation and each element $x_i^t \in X^t$ is an individual.

To build the $(t+1)$-th generation the algorithm processes the t-th generation using three operations (as in Algorithm 1): *reproduction and mutation*, *crossover* and *selection*. The way each one of these operations is performed defines each variant of the Differential Evolution. We use the notation DE/X/Y/Z to denote these variants, X stands for the reproduction, Y for the mutation and Z for the crossover. The selection is always elitist, e.g., given two individuals x and y with $f(x) < f(y)$ we will always select x.

1.1.1 Reproduction and mutation

Algorithm 1 relies on two operations: mutation (that outputs an individual u_i) and crossover (that outputs an individual o_i). The three main configurations for the

---

Algorithm 1: General Differential Evolution algorithm.

Input: N, n, F, C_r, $f(x)$
/* N: number of individuals in population   */
/* n: decision variable dimension   */
/* F: scaling factor   */
/* C_r: crossover rate   */
/* $f(x)$:   objective function   */
Output: x^*
/* Generate the initial population (N individuals in $\mathbb{R}^n$):   */
1 $X^0 \leftarrow$ **initial_population**(N, n)
/* Let $X^0 = \{x_1, \cdots, x_N\}$ where $x_i \in \mathbb{R}^n$ for all $i = 1, \cdots, n$ */
2 $t \leftarrow 0$
3 **while** **_not_** _stopping condition_ **do**
4   **for** $i \leftarrow 1 : N$ **do**
       /* Perform mutation: */
5        $u_i \leftarrow$ **mutation**(F, i, X^t)
       /* Perform crossover: */
6        $o_i \leftarrow$ **crossover**(C_r, u_i, i, X^t)
       /* Perform elitist selection: */
7   **if** $f(o_i) < f(x_i)$ **then**
          /* Change x_i: */
8           $x_i \leftarrow o_i$
9   **else**
          /* Keep x_i: */
10          $x_i \leftarrow x_i$
11     $t \leftarrow t + 1$
     /* Store the population for the next iteration: */
12     $X^t \leftarrow \{x_1, \ldots, x_N\}$
/* Get the best point that was visited: */
13 $x^* \leftarrow \arg\min_{i \in \{1, \ldots, N\}} f(x_i)$

---

reproduction and mutation operator are:

1. DE/rand/k/Z, where u_i is generated by uniformly choosing $2k+1$ distinct individuals $\{x^t_{r_1}, \cdots, x^t_{r_{2k+1}}\}$ in X^t and operating

$$u_i \leftarrow x^t_{r_1} + \sum_{l=1}^{k} F_l(x^t_{r_{2l}} - x^t_{r_{2l+1}}) \tag{1.1}$$

where $F_1, \cdots, F_k$ are algorithm's parameters defined during the initialization. Usually $F_1 = \cdots = F_k = F$ and F is known as the _scaling factor_;

2. DE/best/k/Z, let $x^t_{best} \in X^t$ be the individual of X^t with the minimum value by f. Here u_i is generated by uniformly choosing $2k$ distinct individuals $\{x^t_{r_1}, \cdots, x_{r^t_{2k}}\}$

in $X^t - \{x_{best}^t\}$ and operating

$$u_i \leftarrow x_{best}^t + \sum_{l=1}^{k} F_l(x_{r_{2l-1}}^t - x_{r_{2l}}^t) \tag{1.2}$$

where $F_1, \cdots, F_k$ are previously defined parameters of the algorithm;

3. DE/current/k/Z, where u_i is generated by uniformly choosing $2k$ distinct individuals $\{x_{r_1}^t, \cdots, x_{r_{2k}^t}\} \subset X^t - \{x_i^t\}$ and operating

$$u_i \leftarrow x_i^t + \sum_{l=1}^{k} F_l(x_{r_{2l-1}}^t - x_{r_{2l}}^t) \tag{1.3}$$

where $F_1, \cdots, F_k$ are previously defined parameters of the algorithm.

One may also choose an intermediary configuration, an example is the DE/current-to-best/k/Z, where $2k$ distinct individuals $\{x_{r_1}^t, \cdots, x_{r_{2k}^t}\} \subset X^t - \{x_i^t, x_{best}^t\} \subset X^t$ are uniformly chosen and we operate

$$u_i \leftarrow \lambda x_{best}^t + (1 - \lambda)x_i^t + \sum_{l=1}^{k} F_l(x_{r_{2l-1}}^t - x_{r_{2l}}^t) \tag{1.4}$$

where $\lambda \in (0, 1)$, as well as $F_1, \cdots, F_k$, are previously defined parameters of the Differential Evolution.

1.1.2 Crossover

There are two main crossover variants, the binomial and the exponential, denoted by DE/X/Y/bin and DE/X/Y/exp, respectively. Let $x_i^t = \begin{bmatrix} x_1^i & \cdots & x_n^i \end{bmatrix}^T$, $u_i = \begin{bmatrix} u_1^i & \cdots & u_n^i \end{bmatrix}^T$ and $o_i = \begin{bmatrix} o_1^i & \cdots & o_n^i \end{bmatrix}^T$. In the binomial variant we uniformly choose n values $v_1, \cdots, v_n$ on the interval $(0, 1)$ and set

$$o_s^i = \begin{cases} u_s^i, & \text{if } v_s \leq C_r \\ x_s^i, & \text{if } v_s > C_r \end{cases} \qquad \text{for all } s = 1, \cdots, n \tag{1.5}$$

where C_r is a parameter of the algorithm defined during the initialization, known as *crossover parameter*.

In the exponential variant we choose two numbers, k and l, the first is uniformly chosen in the set $\{1, \cdots, n\}$ and the second is chosen by a geometric distribution with success rate C_r. Then we set:

$$o_k = \begin{cases} u_k, & \text{if } k - 1 = k + m \pmod{n} \text{ for some } m \in \{1, \cdots, l\} \\ x_k, & \text{otherwise} \end{cases} \tag{1.6}$$

1.1.3 Complexity

If the stop criterion is given by a fixed number T_f of iterations one may calculate the time complexity of the algorithm in terms of the time complexity of the objective function f.

For each individual x_i^t a DE/X/k/Z mutation needs to choose, without replacement, $O(k)$ individuals from a list with $O(N)$ individuals, the time complexity of choosing k elements out of N is $O\left(k\left(1 + \log \frac{N}{k}\right)\right)$ [9], after that the algorithm computes u_i in $O(nk)$. Therefore, the mutation is $O\left(k\left(n + \log \frac{N}{k}\right)\right)$. The crossover operator has complexity $O(n)$ because it iterates over n dimensions with $O(1)$ operations. All the operations are executed for each one of the N individuals every iteration, then, excluding the selection (that relies on the evaluation of f), the final time complexity is $O\left(T_f \times N \times k \times \left(n + \log \frac{N}{k}\right)\right)$, and the amount of evaluations of f is $T_f \times N$. Assuming that the size of the population is, at most, exponential on n, then the algorithm is $O\left(T_f \times N \times k \times n\right)$.

The space complexity is simply $O(nN)$ since one needs to store only the $i - th$ population and the u_i or o_i individuals for every $j = 1, \cdots, n$ depending on which stage of the algorithm the individual is. It is also useful to store the image of each individual by f to prevent unnecessary computations of f.

1.1.4 Other variants

Two other small variants are present in the literature. In the mutation operator it is common that when computing u_i the uniformly chosen individuals must also be distinct from x_i. And in the crossover step it is common to see variants where at least one of the entries must suffer crossover. Those variants are often ignored when analysing the algorithm since for high values of N and n their impacts are negligible [10].

1.1.5 Important definitions

We now state the algorithm in a more rigorous way, naming some important sets and events. The Differential Evolution algorithm variant to be studied here is the $DE/rand/1/bin$ variant defined by the pseudo-code presented in Algorithm 2.

Algorithm 2: Differential Evolution algorithm, $DE/rand/1/bin$ variant.

Input: N, n, F, C_r, $f(x)$
```
/* N:  number of individuals in population            */
/* n:  decision variable dimension                     */
/* F:  scaling factor                                  */
/* C_r:  crossover rate                                */
/* f(x):  objective function                           */
```
Output: x^*
```
/* Generate initial population (N individuals in ℝⁿ):  */
```
1   $X^0 \leftarrow$ **initial_population**(N, n)
```
/* Let X⁰ = vec(x₁,··· ,x_N) where xᵢ ∈ ℝⁿ for all i = 1,··· ,n  */
```
2   $t \leftarrow 0$
3   **while** ***not*** *stopping condition* **do**
4   **for** $i \leftarrow 1 : N$ **do**
5       $i_1 \leftarrow i$
```
      /* Select 3 elements, without replacement, from the set
         {1,2,...,N} − {i₁}, with uniform probability:      */
```
6       $\{i_2, i_3, i_4\} \leftarrow$ **rand_select**$(3, \{1, 2, \ldots, N\} - \{i\})$
```
      /* Retrieve the individuals to be submitted to crossover and
         mutation, formally we have 𝓑ᵢ = (i₁,i₂,i₃,i₄):     */
```
7       $\bar{x}_1 \leftarrow X^t[i_1]$
8       $\bar{x}_2 \leftarrow X^t[i_2]$
9       $\bar{x}_3 \leftarrow X^t[i_3]$
10      $\bar{x}_4 \leftarrow X^t[i_4]$
```
      /* Build the set 𝓐ᵢ ⊂ {1,2,...,n} by a random choice such that
         each element is put in 𝓐ᵢ with probability C_r:   */
```
11      $\mathcal{A}_i \leftarrow$ **rand_subset**(n, C_r)
```
      /* Perform mutation and crossover:                    */
```
12      $o_i \leftarrow \sum_{j \notin \mathcal{A}_i} \langle \bar{x}_1, e_j \rangle e_j + \sum_{j \in \mathcal{A}_i} \langle \bar{x}_2 + F.(\bar{x}_3 - \bar{x}_4), e_j \rangle e_j$
```
      /* Perform elitist selection:                         */
```
13   **if** $f(o_i) < f(\bar{x}_i)$ **then**
14        $\bar{x}_i \leftarrow o_i$`/* If oᵢ is better then x̄ᵢ updates the value of x̄ᵢ  */`

15     $t \leftarrow t + 1$
```
    /* Store the population for the next iteration:          */
```
16     $X^t \leftarrow \text{vec}(\bar{x}_1, \ldots, \bar{x}_N)$
```
  /* Get the best point that was visited:                   */
```
17   $x^* \leftarrow \arg\min_{i \in \{1, \ldots, N\}} f(x_i)$

We can represent a population ρ of $N \geq 4$ individuals $x_1, \cdots, x_N \in \mathbb{R}^n$ as a set of column vectors $\rho = (x_1, \cdots, x_N) \in \mathbb{R}^{n \times N}$[1]. We also define $B_{n \times N}$ as the unit ball centered on the origin.

Let $k \leq n$ be a positive integer, $\mathcal{A}$ be a subset of $\{1, 2, \cdots, n\}$ such that $|\mathcal{A}| = k$ and $\mathcal{B} = (i_1, i_2, i_3, i_4)$ a sequence of four distinct integers in $\{1, 2, \cdots, N\}$, we say that $(\mathcal{A}, \mathcal{B})$

---

[1]When we see a population as a point in $\mathbb{R}^{n \times N}$ we use the greek letter ρ. When we see it as a state during the execution of the algorithm we use X^t.

is a k-configuration. We also say that $\lambda = \{(\mathcal{A}_i, \mathcal{B}_i)\}_{i=1}^N$ is an iteration rule if $(\mathcal{A}_i, \mathcal{B}_i)$ is a k_i-configuration for every $i = 1, \cdots, N$ such that $\mathcal{B}_i = (i, \cdot, \cdot, \cdot)$. By following Algorithm (2) we can see the role played by an iteration rule $\{(\mathcal{A}_i, \mathcal{B}_i)\}_{i=1}^N$ at each iteration: tuples $\mathcal{B}_i$ store which individuals were chosen for the mutation step and sets $\mathcal{A}_i$ store which coordinates from the original individual i were used by the crossover. An iteration rule fully characterizes an iteration.

A population $\rho = (x_1, \cdots, x_N) \in \mathbb{R}^{n \times N}$ is $(\mathcal{A}, \mathcal{B})$-contracting if

$$f(o) < f(x_{i_1}), \text{ where } o = \sum_{j \notin \mathcal{A}} \langle x_{i_1}, e_j \rangle e_j + \sum_{j \in \mathcal{A}} \langle x_{i_2} + F(x_{i_3} - x_{i_4}), e_j \rangle e_j \qquad (1.7)$$

and e_j is the j-th canonical vector. We also define $\Delta_{\mathcal{A}, \mathcal{B}}$ to be the set of all $(\mathcal{A}, \mathcal{B})$-contracting populations $\rho \in B_{N \times n}$ and Λ_k^n to be the set of all populations $\rho \in B_{N \times n}$ such that $\rho \notin \Delta_{\mathcal{A}, \mathcal{B}}$ for all $(\mathcal{A}, \mathcal{B})$ k-configuration. A population $\rho \in \Lambda_k^n$ is called a k-fixed population.

Finally, we define a separable function $f : \mathbb{R}^n \to \mathbb{R}$ as a function such that for every $i = 1, \cdots, n$ and every choice of $(x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n), (x'_1, \cdots, x'_{i-1}, x'_{i+1}, \cdots, x'_n) \in \mathbb{R}^{n-1}$ we have

$$S_i = \operatorname{argmin}_{x_i \in \mathbb{R}} f(x_1, \cdots, x_n) = \operatorname{argmin}_{x'_i \in \mathbb{R}} f(x'_1, \cdots, x'_n). \qquad (1.8)$$

Notice that if f is separable, then

$$\operatorname{argmin}_{x \in \mathbb{R}^n} f(x) = S_1 \times S_2 \times \cdots \times S_n. \qquad (1.9)$$

Examples of separable functions are additive functions of the form

$$f(x_1, \cdots, x_n) = f_1(x_1) + \cdots + f_n(x_n)$$

or multiplicative functions of the form

$$f(x_1, \cdots, x_n) = \prod_{i=1}^n f_i(x_i),$$

where f_i is a real function for every $i = 1, \cdots, n$.

1.2 Some properties and remarks

In the next chapters some analytical results about the algorithm will be presented, but before that we introduce the plots that will be used on the text and also show an
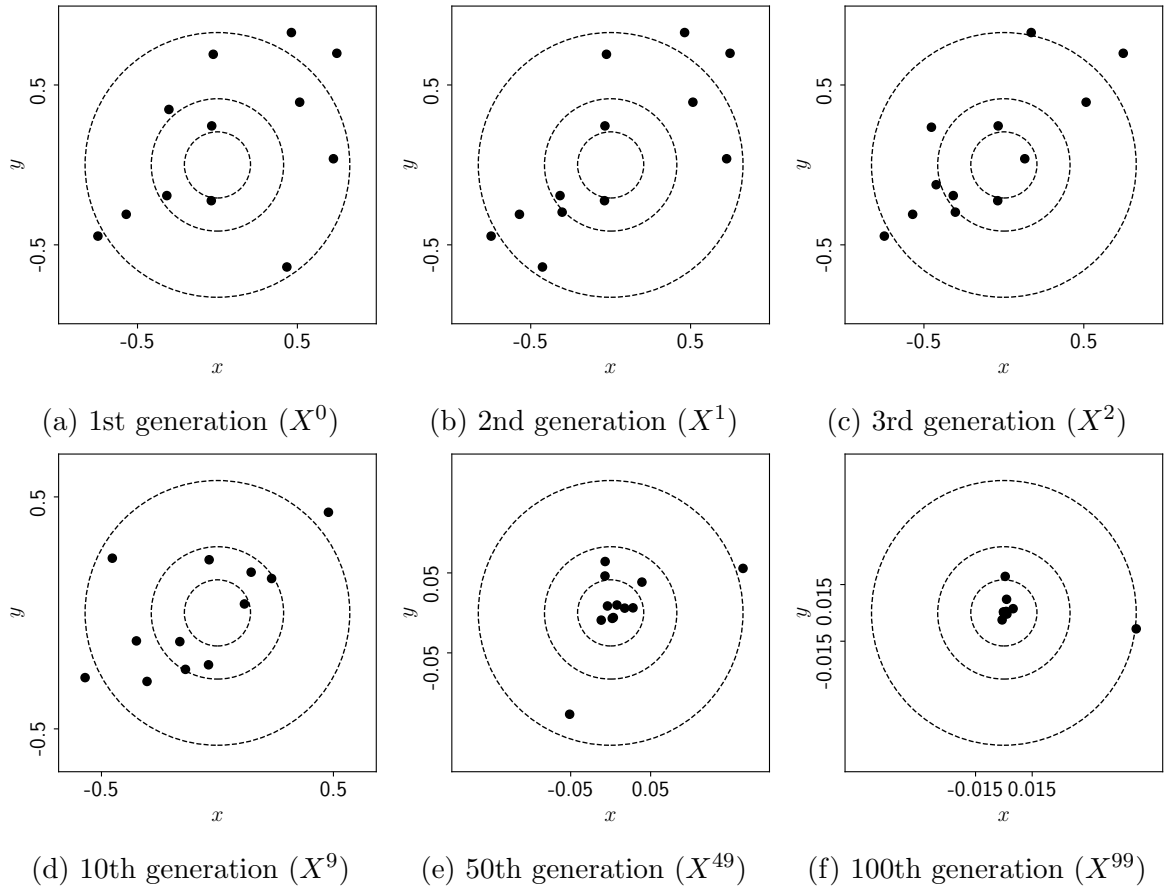
(a) 1st generation (X^0)   (b) 2nd generation (X^1)   (c) 3rd generation (X^2)

(d) 10th generation (X^9)   (e) 50th generation (X^{49})   (f) 100th generation (X^{99})

Figure 1.1: Typical execution of the DE algorithm for $f(x) = ||x||^2$ with $n = 2$, $N = 12$, $F = 0.75$ and $C_r = 0.5$.

execution of the algorithm. Figure 1.1 shows an execution of the algorithm for the function $f(x) = ||x||_2^2$.

To visualize the evolution of the algorithm in higher dimensions we plot two curves: the best function value reached by the algorithm (see Figures 1.2(a) and 1.2(c)) and the distance of the best individual to the optimum (see Figures 1.2(b) and 1.2(d)). To construct, for example, Figures 1.2(a) and 1.2(b) 50 executions of the algorithm were run. For each generation t we compute the 10% (lower dotted), 25% (lower dashed), 50% (solid), 75% (upper dashed) and 90% (upper dotted) quantiles of the best f value over all executions, producing five curves that relates each generation with its best f value (Figure 1.2(a)), the process to plot the distance of the best individual to the optimum is analogous (Figure 1.2(b)). To compare executions with different parameters or different algorithms we set the x-axis to be the number of function evaluations. Figures 1.2(c) and 1.2(d) were constructed using the mean of 50 runs, each one with a budget of 10000 evaluations.

A first behavior that one may observe is that a population can get stuck around a local optimum depending on the parameter F and on the function f (see subsection 1.3.2). That behavior can be justified by the elitist selection.
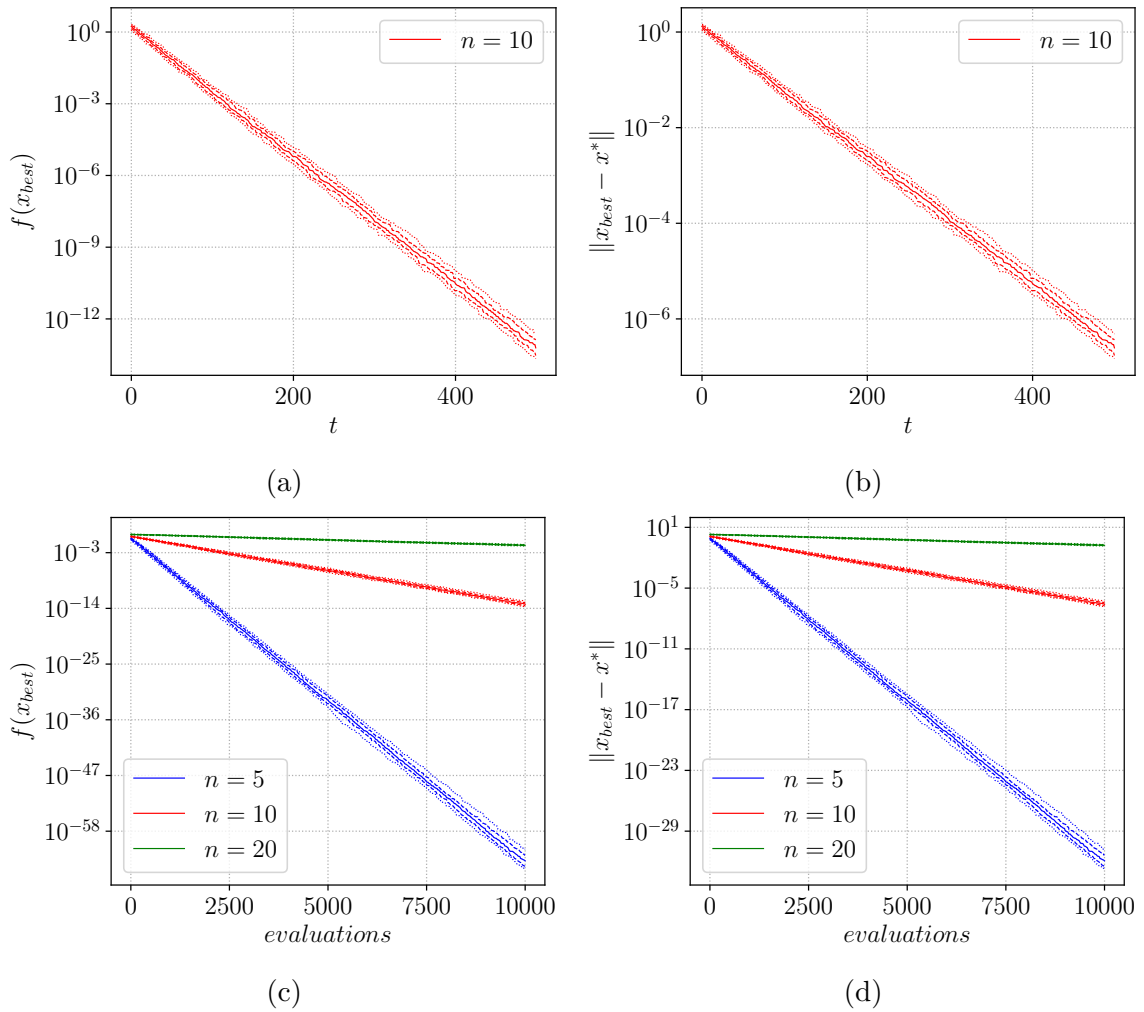
Figure 1.2: (a) Value of the function on the best individual (x_{best}^t) at each generation t. (b) Distance of the best individual x_{best}^t to the optimum x^* at each generation. (c) Value of the function for the best individual for each amount of function evaluations. (d) Distance of the best individual to the optimum x^* for each amount of function evaluations. All with $N = 2n$, $C_r = 0.5$ and $F = 0.75$. Each curve was produced with 50 executions, representing the 10% (lower dotted), 25% (lower dashed), 50% (solid), 75% (upper dashed) and 90% (upper dotted) quantiles.

One may also notice that the algorithm can't distinguish functions that have the same relative positions, i.e., if $f(x) < f(y)$ if and only if $g(x) < g(y)$ then the algorithm can't distinguish f and g.

Finally, Figures 1.2(c) and 1.2(d) shows the impact of the number of dimensions of the domain in the algorithm's performance. That kind of behavior was reported and justified previously for other optimization algorithms (see [2] for an example) and is known as the *curse of dimensionality*.

The *curse of dimensionality* was a term coined by Richard Bellman, the father of Dynamic Programming, to describe the exponential trade-off between the number of dimensions of the search space and the amount of function evaluations needed to find the

optimum. In this work we aim to study that behavior for the Differential Evolution.

1.3   Convergence

An evolutionary algorithm that aims to minimize a function $f : \mathcal{D} \to \mathbb{R}$ by iterating each generation using only the previous one can be seen as a homogeneous Markov chain[2] $(X^t : \Omega \to E)$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with state space given by $E = \mathcal{D}^N$ (where N is the number of individuals) and with a kernel K that stands for the transition rules (mutation, crossover and selection).

To define the kernel we need a measurable space $(E, \mathcal{F})$, that can be done in several ways depending on the domain $\mathcal{D}$. For example, if $\mathcal{D} = \mathbb{R}^n$ one can set $\mathcal{F}$ as the borelians, if $|\mathcal{D}| < \infty$ we can choose the power set.

In this section we will see a simple criterion developed by Rudolph [13] that allows us to ensure the convergence of an evolutionary algorithm to the optimum. We will also see an objective function, designed by Hu et. al [7], for which the DE may not converge to the optimum depending on the initial population.

1.3.1   Convergence definitions

Let f^* be the minimum value of f, define $b : E \to \mathbb{R}$ as $b((x_1, \cdots, x_N)) = \min_{1 \leq i \leq N} f(x_i)$ and $d : E \to \mathbb{R}$ as $d(X) = b(X) - f^*$. Now we can define what it means for an algorithm to converge:

Definition 1.1 (Convergence of an evolutionary algorithm, from [13]). *Let $(X^t)_{t \geq 0}$ be the Markov chain defined by an evolutionary algorithm. We say that the algorithm converges in probability to the optimum if*

$$\lim_{t \to \infty} \mathbb{P}(d(X^t) > \epsilon) = 0 \tag{1.10}$$

for any $\epsilon > 0$ and for all initial distribution of X^0. We say that an algorithm converges completely to the optimum if

$$\sum_{t=1}^{\infty} \mathbb{P}(d(X^t) > \epsilon) < \infty \tag{1.11}$$

---

[2]See Appendix A

for any $\epsilon > 0$ and for all initial distribution of X^0.

Note that complete convergence is just almost everywhere convergence and implies convergence in probability. Before the main result we need a lemma:

Lemma 1.1 ([13]). *Let $\epsilon > 0$ and $A_\epsilon = \{\rho \in E : d(\rho) < \epsilon\}$. If there exists $\delta = \delta(\epsilon) \in (0, 1]$ such that $K(\rho, A_\epsilon) \geq \delta$ for all $\rho \in A_\epsilon^c$ (i.e., $\rho \notin A_\epsilon$) and $K(\rho, A_\epsilon) = 1$ for all $\rho \in A_\epsilon$, then*

$$K^{(t)}(\rho, A_\epsilon) \geq 1 - (1 - \delta)^t \tag{1.12}$$

for all $t \geq 1$.

Proof. We start by using induction to show that $K^{(t)}(\rho, A_\epsilon) = 1$ for all $t \geq 1$ and all $\rho \in A_\epsilon$. The base case is covered by the hypothesis, so let's assume it holds for t to show it for $t + 1$. Using that $K(\rho, A_\epsilon^c) = 0$ we get:

$$K^{(t+1)}(\rho, A_\epsilon) = \int_E K^{(t)}(y, A_\epsilon) K(\rho, dy) \tag{1.13}$$

$$= \int_{A_\epsilon} K^{(t)}(y, A_\epsilon) K(\rho, dy) + \int_{A_\epsilon^c} K^{(t)}(y, A_\epsilon) K(\rho, dy) \tag{1.14}$$

$$= \int_{A_\epsilon} K(\rho, dy) + 0 = K(\rho, A_\epsilon) = 1 \tag{1.15}$$

Now we show, also by induction, that $K^{(t)}(\rho, A_\epsilon) \geq 1 - (1 - \delta)^t$ for all $t \geq 1$ and all $\rho \in E$. If $t = 1$ then $K^{(1)}(\rho, A_\epsilon) = K(\rho, A_\epsilon) \geq \delta = 1 - (1 - \delta)^1$. Assuming it for t, then:

$$K^{(t+1)}(\rho, A_\epsilon) = \int_E K^{(t)}(y, A_\epsilon) K(\rho, dy) \tag{1.16}$$

$$= \int_{A_\epsilon} K^{(t)}(y, A_\epsilon) K(\rho, dy) + \int_{A_\epsilon^c} K^{(t)}(y, A_\epsilon) K(\rho, dy) \tag{1.17}$$

$$= K(\rho, A_\epsilon) + \int_{A_\epsilon^c} K^{(t)}(y, A_\epsilon) K(\rho, dy) \tag{1.18}$$

$$\geq K(\rho, A_\epsilon) + (1 - (1 - \delta)^t) \int_{A_\epsilon^c} K(\rho, dy) \tag{1.19}$$

$$= K(\rho, A_\epsilon) + K(\rho, A_\epsilon^c) - (1 - \delta)^t K(\rho, A_\epsilon^c) \tag{1.20}$$

$$= 1 - (1 - \delta)^t (1 - K(\rho, A_\epsilon)) \tag{1.21}$$

$$\geq 1 - (1 - \delta)^{t+1} \tag{1.22}$$

$\square$

Theorem 1.1 ([13]). *An evolutionary algorithm whose kernel fulfills the hypothesis of Lemma 1.1 converges completely to the minimum regardless of the initial distribution $p : \mathcal{F} \to [0, 1]$.*

Proof. Let $(X^t)_{t\geq 0}$ be a sequence of populations generated by the algorithm, we want to show that $(d(X^t))_{t\geq 0}$ converges completely to zero. Let $\epsilon > 0$ and $\delta = \delta(\epsilon)$, then

$$\mathbb{P}(X^t \in A_\epsilon) = \int_E K^{(t)}(\rho, A_\epsilon)p(d\rho) \geq (1 - (1 - \delta)^t)\int_E p(d\rho) = 1 - (1 - \delta)^t \qquad (1.23)$$

and finally

$$\sum_{t=1}^{\infty}\mathbb{P}(d(X^t) > \epsilon) \leq \sum_{t=1}^{\infty}(1 - \delta)^t = \frac{1 - \delta}{\delta} < \infty. \qquad (1.24)$$

$\square$

Example 1.1. *Let $f : [0,1]^m \to \mathbb{R}$ be a continuous function to be optimized by the following algorithm: let $X^t = \{x_1^t, \cdots, x_N^t\}$ be the population at time t, at each iteration uniformly and independently pick N points $o_1^t, \cdots, o_N^t$ and generate X^{t+1} by setting x_i^{t+1} as x_i^t if $f(x_i^t) \leq f(o_i^t)$ or as o_i^t otherwise. Then, for every $\epsilon > 0$,*

$$K(\rho, A_\epsilon) = \begin{cases} 1, \textit{ if } \rho \in A_\epsilon \\ 1 - (1 - m(A_\epsilon))^N, \textit{ if } \rho \notin A_\epsilon \end{cases}$$

where m is the Lebesgue measure. Choosing $\delta = \delta(\epsilon) = 1 - (1 - m(A_\epsilon))^N \in (0, 1]$ and using Theorem 1.1 we can see that this algorithm converges.

Unfortunately, we can't use Theorem 1.1 to study the convergence of the DE algorithm. Although the condition $K(\rho, A_\epsilon) = 1$ hols for every $\rho \in A_\epsilon$, it is possible that $K(\rho, A_\epsilon) = 0$ for some $\rho \notin A_\epsilon$, we just need ρ to be far enough from A_ϵ.

One may think about K by looking at the mutation (K_m), crossover (K_c) and selection (K_s) kernels separately. Operating first with the mutation, then with the crossover and using an elitist selection we compute:

$$K(\rho, A_\epsilon) = \begin{cases} 1, \text{ if } \rho \in A_\epsilon \\ \int_E K_c(y, A_\epsilon)K_m(\rho, dy), \text{ if } \rho \notin A_\epsilon \end{cases} \qquad (1.25)$$

If the crossover has a probability $p < 1$ and $\forall \epsilon > 0 \; \exists \delta = \delta(\epsilon)$ such that $K_m(\rho, A_\epsilon) \geq \delta \; \forall \rho \in E$, then

$$\int_E K_c(y, A_\epsilon)K_m(\rho, dy) \geq \int_{A_\epsilon} K_c(y, A)K_m(\rho, dy) \qquad (1.26)$$

$$\geq (1 - p)\int_{A_\epsilon} K_m(\rho, dy) \geq (1 - p)\delta \qquad (1.27)$$

and so K is on the hypothesis of Theorem 1.1.

Theorem 1.2. *If an evolutionary algorithm has elitist selection, a crossover probability of $p \leq 1$ and the mutation kernel K_m fulfills the hypothesis of Lemma 1.1 then the algorithm converges completely.*

Proof. Follows from the observations above. □

That way, by looking at each operation at a time we can obtain information about the dynamics of the algorithm. This idea will be explored later to study the algorithm on high dimensions.

1.3.2 The non-convergence of the algorithm

In [7] the authors designed a family of functions for which the convergence to the optimum isn't assured regardless of the initial distribution. Let k be a positive integer and define:

$$f_k(x) = \begin{cases} -kx - 1, \text{ if } -\frac{2}{k} < x < -\frac{1}{k} \\ kx + 1, \text{ if } -\frac{1}{k} \leq x < 0 \\ -\frac{x}{k} + 1, \text{ if } 0 \leq x < k - 1 \\ \frac{1}{k}, \text{ if } k - 1 \leq x \leq k \end{cases} \tag{1.28}$$
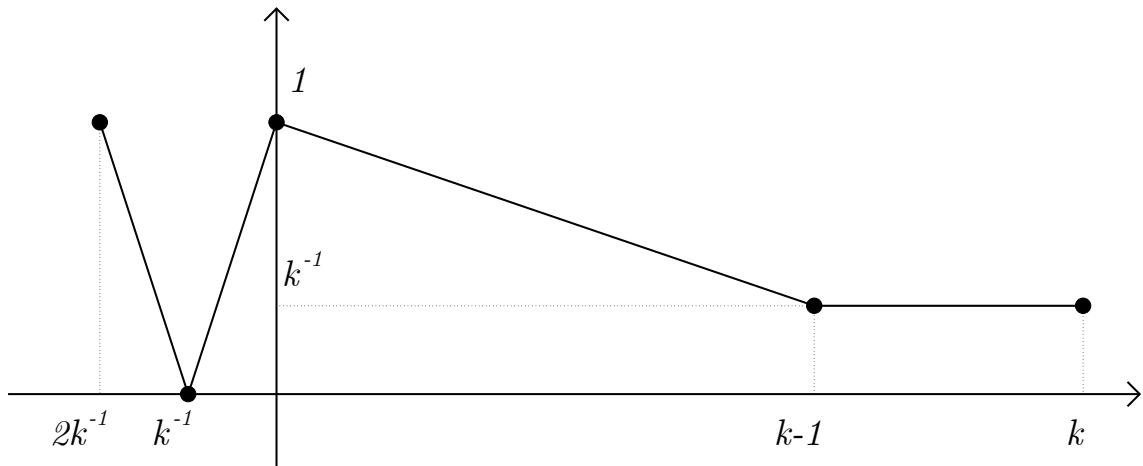
Figure 1.3: The f_k function.

Take any version of the DE that uses l difference vectors during the mutation (i.e. $DE//l$) and set $F \leq \frac{1}{l}$. Note that if $X^t \in \left[\frac{k}{2}, k\right]^N$ then the operations of mutation and crossover reach only points on $[0, k]$. Since every point in $\left[0, \frac{k}{2}\right]$ is worse than any point in $\left[\frac{k}{2}, k\right]$ we have that $X^{t+1} \in \left[\frac{k}{2}, k\right]$. Taking $\epsilon = \frac{1}{k}$ and

$$S = \left\{ (x_1, \cdots, x_N) \in E : \frac{k}{2} \leq x_i \leq 0 \, \forall i = 1, \cdots, N \right\}$$

we have that:

$$\mathbb{P}(X^t \in A_\epsilon) \leq 1 - P\left(X^0 \in S\right) \tag{1.29}$$

Then, if $\mathbb{P}(X^0 \in S) > 0$ the DE algorithm won't converge even in probability to the optimum. For example, let X^0 be uniformly distributed on $[0,k]^N$, then $\mathbb{P}(X^0 \in S) = \prod_{i=1}^N \mathbb{P}\left(\frac{k}{2} \leq x_i \leq k\right) = 2^{-N} > 0$.

1.3.3 Population diversity analysis

Since the selection step is strongly dependent on the objective function f, it is difficult to produce general results about the algorithm. On the other hand, the first two operators (mutation and crossover) don't need f. In [18] Zaharie studied these first operations and obtained results on critical values for parameters.

On evolutionary algorithms the concept of population diversity plays an important role. During the execution of an evolutionary algorithm we want to have populations that encode as many information as possible. For instance, a population $\rho = (x_1, \cdots, x_N) \in \mathbb{R}^{n \times N}$ where $x_i = (x_i^1, \cdots, x_i^n)$ for all $i = 1, \cdots, n$ and $x_i^1 = x_j^1$ for all $1 \leq i, j \leq N$ can't explore the first coordinate of $\mathbb{R}^n$ since all of its first coordinates are the same. The following definition encapsulates the idea of population diversity:

Definition 1.2 (Population variance, [18])**.** *Let* $\rho = (x_1, \cdots, x_N) \in \mathbb{R}^{1 \times N}$ *be a population of scalars, its variance is defined by*

$$Var(\rho) = \frac{1}{N} \sum_{i=1}^N (x_i - \mathbb{E}\rho)^2 \tag{1.30}$$

where $\mathbb{E}\rho = \frac{1}{N} \sum_{i=1}^N x_i$. *If* $\rho = (x_1, \cdots, x_N) \in \mathbb{R}^{n \times N}$ *is a population of vectors* $x_i = (x_i^1, \cdots, x_i^n)$ *we define its variance as*

$$Var(\rho) = \frac{1}{n} \sum_{j=1}^n Var(\rho_j) \tag{1.31}$$

where $\rho_j = (x_1^j, x_2^j, \cdots, x_N^j)$.

Notice that if $\rho = (x_1, \cdots, x_N)$ is a scalar population then the populations $U = (u_1, \cdots, u_N)$ generated after the mutation step and $O = (o_1, \cdots, o_N)$ generated by U after crossover are random vectors. We also define $\mathbb{E}(\rho) = \frac{1}{N} \sum_{i=1}^N x_i$ and $\mathbb{E}(\rho^2) = \frac{1}{N} \sum_{i=1}^N x_i^2$.

Remark 1.1. *In the following theorem, Zaharie studied the population variance of the Differential Evolution with a minor modification: during the mutation step for* x_i*, the algorithm now selects three distinct individuals* $x_{i_1}, x_{i_2}, x_{i_3}$ *to construct* $u_i = x_{i_1} + F(x_{i_2}, x_{i_3})$

without requiring that $i \neq i_j$ for all $j = 1, 2, 3$. Higher values of N produce lower probabilities that $i = i_j$ for some $j \in \{1, 2, 3\}$, then for high values of N that variant behaves closely to the algorithm defined in 2.

Theorem 1.3 ([18]). *Let $\rho = (x_1, \cdots, x_N)$ be a scalar population, $U = (u_1, \cdots, u_N)$ the result of the mutation step and $O = (o_1, \cdots, o_N)$ the result of the crossover step between ρ and U, then*

$$\mathbb{E}(\mathit{Var}(U)) = \left(2F^2 + \frac{N-1}{N} \right) \mathit{Var}(\rho) \tag{1.32}$$

and

$$\mathbb{E}(\mathit{Var}(O)) = \left(1 + 2C_r F^2 - \frac{2C_r}{N} + \frac{C_r^2}{N} \right) \mathit{Var}(\rho). \tag{1.33}$$

Proof. See Appendix B. □

1.3.4 Critical region for parameters

Theorem B provides a critical region for the parameters F, N and C_r. We must set

$$1 + 2C_r F^2 - \frac{2C_r}{N} + \frac{C_r^2}{N} \geq 1 \tag{1.34}$$

to avoid loosing population diversity. Since the analysis was performed ignoring the selection and reproduction step it is interesting to experimentally verify if the impact of this step is relevant.

Usually, the inequality 1.34 is used to obtain a critical value of F, given by $F_{crit} = \sqrt{\frac{1}{N} - \frac{C_r}{2N}}$. In [18] the author also performs experiments to verify if premature convergence occurs when $F < F_{crit}$, founding premature convergence in most of the executions meeting $F < F_{crit}$. Se say that premature convergence occurs during an experiment when the improvements became meaningless. Since it is a behavior experimentally observed, premature convergence occurs both when the algorithm is really converging to an undesirable population (that doesn't have an optimum individual) or when the algorithm is converging to a desirable population, but very slowly.

Chapter 2

Using the Sphere Function

In this chapter we set $f : \mathbb{R}^n \to \mathbb{R}$ as the sphere function $f(x) = \langle x, x \rangle = ||x||_2^2$. Using this f we try to understand how the algorithm behaves for high values of n, looking for probabilities of improvement and fixed points.

2.1   Motivation

Figure 2.1 exemplifies how different values of k changes the way a k-configuration $(\mathcal{A}, \mathcal{B})$ acts on a population ρ.
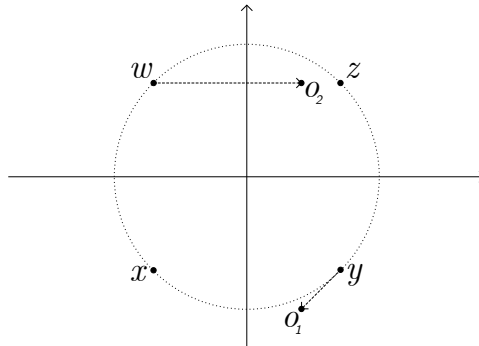
Figure 2.1: The individual w can't be enhanced by $o_1 = u = y + F(x - z)$ using all two dimensions (case $(\mathcal{A}, \mathcal{B}) = (\{1, 2\}, (w, y, x, z))$), regardless of the F parameter (the point o_1 is always outside the level set). But using only one dimension and a suitable value of F we can update w to o_2 (case $(\mathcal{A}, \mathcal{B}) = (\{1\}, (w, y, x, z))$). Note that $\mathcal{A} = \{1\}$ means we are performing the crossover only on the first coordinate and $\mathcal{B} = (w, y, x, z)$ that the mutation will produce an individual $u = y + F(x - z)$ to crossover with w.

To analyse the effect of a k-mutation we modified the DE to do only k-mutations and looked at the following values:

- m_{ic}, the mean over all executions of the number of times that an individual was replaced by its offspring;

- m_{ir}, the mean value of the ratio between the objective function on best individual of the first iteration and on the best of the last iteration;

- m_s, the number of executions that reach a population $\rho \in \Lambda_k^n$.

Tables 2.1, 2.2 and 2.3 were constructed using 100 executions of the algorithm per k, n combination, a stop criterion of 100 generations and $F = 0.75$. Note that the improvement rate m_{ir} gets worse as k increases, nevertheless, for the tested values of N lower than $2n$ the mean number of improvements m_{ic} decreases and then starts to grow again. The results also point to the idea that k-fixed populations are common for values of k close to n.

k	m_{ic}	m_{ir}	m_s	k	m_{ic}	m_{ir}	m_s	k	m_{ic}	m_{ir}	m_s
1	146.11	0.049	-	1	293.19	0.011	-	1	576.57	0.007	-
2	144.85	0.039	-	2	268.4	0.009	-	2	518.8	0.009	-
3	133.9	0.041	-	3	237.18	0.011	-	3	454.31	0.011	-
4	130.27	0.056	-	4	212.52	0.013	-	4	398.46	0.016	-
5	132.13	0.108	-	5	197.84	0.014	-	5	351.73	0.021	-
6	139.77	0.183	-	6	190.82	0.020	-	6	318.06	0.027	-
7	153.23	0.312	-	7	208.09	0.045	-	7	291.51	0.034	-
8	140.61	0.634	-	8	-	-	9	8	-	-	60

Table 2.1: $n = 8$

k	m_{ic}	m_{ir}	m_s	k	m_{ic}	m_{ir}	m_s	k	m_{ic}	m_{ir}	m_s
2	266.02	0.098	-	2	517.65	0.098	-	2	1026.71	0.096	-
4	210.89	0.107	-	4	399.4	0.118	-	4	779.1	0.123	-
6	172.31	0.131	-	6	315.33	0.154	-	6	609.27	0.164	-
8	152.23	0.150	-	8	253.63	0.208	-	8	478.26	0.234	-
10	138.99	0.162	-	10	210.6	0.253	-	10	378.45	0.305	-
12	145.19	0.166	-	12	187.02	0.283	-	12	305.37	0.390	-
14	184.48	0.219	-	14	196.78	0.233	-	14	271.14	0.418	-
16	-	-	31	16	-	-	75	16	-	-	94

Table 2.2: $n = 16$

k	m_{ic}	m_{ir}	m_s	k	m_{ic}	m_{ir}	m_s	k	m_{ic}	m_{ir}	m_s
4	397.5	0.342	-	4	782.49	0.353	-	4	1551.5	0.346	-
8	253.28	0.440	-	8	483.55	0.454	-	8	954.1	0.472	-
12	171.56	0.554	-	12	318.16	0.603	-	12	609.14	0.613	-
16	121.29	0.662	-	16	210.85	0.720	-	16	389.63	0.756	-
20	97.17	0.715	-	20	147.71	0.816	-	20	255.05	0.862	-
24	98.75	0.659	-	24	117.19	0.833	-	24	177.67	0.932	-
28	126.28	0.514	-	28	123.91	0.753	-	28	144.92	0.913	-
32	-	-	74	32	-	-	98	32	-	-	97

Table 2.3: $n = 32$

In the DE algorithm the choice of k on each crossover operation depends on the parameter C_r. Therefore, by analysing k-configurations we might expect glimpses on how to choose the parameter C_r. Some experimental research was done with this purpose. In [12], the authors recommended $C_r < 0.2$ for separable functions and $C_r > 0.9$ for non-separable functions, but they used a fixed scaling factor $F = 0.9$ and didn't study the influence of n on the choice of these parameters. In [15] the author gives similar advice to the user, but adds that the value of F must be on the interval $[0.5, 1]$ and that N should be equal to $10n$. A more recent article [2] discusses the well known *curse of dimensionality* with experimental evidence, saying that since the time complexity is super-linear on n the use of population sizes N lower then n is needed. The article also points out that it seems to be possible to compensate lower values of N with smarter values of F. Despite of the abundance of empirical evidence, few articles give rigorous justifications of their results.

2.2 Probabilities of improvement

2.2.1 Improvements from k-mutations

Figure 2.1 shows that it is possible to create configurations from which the DE algorithm can't provide any improvement using 2-configurations, but it doesn't answer how frequent these configurations are, the next theorem starts to answer that.

Theorem 2.1 (Resende and Takahashi). *For a fixed scaling factor F, let $p_k(F)$ be the*

probability that a randomly uniformly chosen population $\rho \in B_{n \times N}$ belongs to $\Delta_{\mathcal{A},\mathcal{B}}$ for a fixed k-configuration $(\mathcal{A}, \mathcal{B})$, then $p_k(F)$ is independent of both n and N and

$$p_k(F) = \frac{\gamma^{-\frac{k}{2}}\alpha^k}{k\beta\left(\frac{k}{2}, \frac{k}{2}+1\right)} + 1 - I_\alpha\left(\frac{k}{2}, \frac{k}{2}+1\right) \tag{2.1}$$

where

$$\beta(a,b) = \int_0^1 r^{a-1}(1-r)^{b-1}dr \tag{2.2}$$

is the beta function,

$$I_x(a,b) = \frac{\int_0^x r^{a-1}(1-r)^{b-1}dr}{\beta(a,b)} \tag{2.3}$$

is the normalized incomplete beta function,

$$\gamma = 2F^2 + 1 \geq 1 \tag{2.4}$$

and $\alpha = \frac{\gamma}{1+\gamma}$.

Moreover, $p_k(F) \leq Dr^k$ for some $D \in \mathbb{R}$ and $0 < r < 1$. Therefore, $p_k(F) \to 0$ as k goes to ∞.

Before proving the theorem we need the following lemma:

Lemma 2.1. *Let $V_n = vol(B_n)$ be the volume of a n-dimensional ball of unitary radius, then:*

$$\int_0^\pi \cdots \int_0^\pi sen^{n-2}\theta_{n-2} \cdots sen\theta_1 d\theta_{n-2} \cdots d\theta_1 = \frac{n}{2\pi}V_n$$

Proof. It follows from the definition of V_n as an integral in spherical coordinates. $\square$

Proof of Theorem 2.1, first part. Using $\rho = (x_1, \cdots, x_N)$ as a column vector and assuming w.l.g. that $\mathcal{A} = \{1, 2, \cdots, k\}$ and $\mathcal{B} = (1, 2, 3, 4)$, the inequality (1.7) can be seem as $\rho^T M \rho < 0$, where

$$M_{nN \times nN} = \left[\begin{array}{c:c:c:c:c} -\mathbb{I}_{n,k} & 0 & 0 & 0 & 0 \\ \hdashline 0 & \mathbb{I}_{n,k} & F\mathbb{I}_{n,k} & -F\mathbb{I}_{n,k} & 0 \\ \hdashline 0 & F\mathbb{I}_{n,k} & F^2\mathbb{I}_{n,k} & -F^2\mathbb{I}_{n,k} & 0 \\ \hdashline 0 & -F\mathbb{I}_{n,k} & -F^2\mathbb{I}_{n,k} & F^2\mathbb{I}_{n,k} & 0 \\ \hdashline 0 & 0 & 0 & 0 & 0 \end{array}\right] , \mathbb{I}_{n,k} = \left[\begin{array}{c:c} \mathbb{I}_k & 0 \\ \hdashline 0 & 0 \end{array}\right]_{n \times n} \tag{2.5}$$

and $\mathbb{I}_k$ as the $k \times k$ identity matrix.

The matrix M is symmetric and, therefore, diagonalizable in the form $S^T D S$, with S orthogonal. The eigenvalues of M are 0 with multiplicity $nN - 2k$, $2F^2 + 1$ with multiplicity k and -1 with multiplicity k.

Let $C = \Delta_{\mathcal{A},\mathcal{B}} \cap B_{n \times N}$. Since S is orthogonal, $vol(C) = vol(S(C))$. The probability that $p \in B_{n \times N}$ belongs to C is given by $\frac{vol(C)}{V_{nN}} = \frac{vol(S(C))}{V_{nN}}$. Using that $S(B_{n \times N}) = B_{n \times N}$ we get

$$S(C) = \left\{(\xi, \eta, \omega) \in B_{n \times N} : ||\xi||^2 < \gamma^{-1}||\eta||^2\right\} \tag{2.6}$$

where $\omega \in \mathbb{R}^{nN-2k}$, $\xi, \eta \in \mathbb{R}^k$ and $\gamma = 2F^2 + 1$.

Using χ as the indicator function we have, by Fubini (see, for example, [1]):

$$vol(S(\Delta_{\mathcal{A},\mathcal{B}})) = \int_{||\omega||^2 \leq 1} \int_{||(\xi,\eta)||^2 \leq 1 - ||\omega||^2} \chi_{S(C)} d\xi d\eta d\omega \tag{2.7}$$

$$= vol(A) \int_{||\omega||^2 \leq 1} \left(1 - ||\omega||^2\right)^k d\omega \tag{2.8}$$

where $A = \{(\xi, \eta) \in B_{2k} : ||\xi||_2^2 < \gamma^{-1}||\eta||_2^2\}$.

We compute the last integral using spherical coordinates and the lemma:

$$\int_0^1 \int_0^{2\pi} \int_0^\pi \cdots \int_0^\pi (1 - r^2)^k r^{nN-2k-1} \prod_{i=1}^{nN-2k-2} \sin^i \theta_i d\theta_1 \cdots d\theta_{nN-2k-1} dr \tag{2.9}$$

$$= (2n - k)V_{nN-2k}\beta\left(2n - k, k + 1\right) \tag{2.10}$$

The volume of A is calculated in two parts, fixing η we look at which intervals the inequality $||\xi||_2^2 < 1 - ||\eta||_2^2$ dominates $||\xi||_2^2 < \gamma^{-1}||\eta||_2^2$ and vice-versa. The first inequality dominates when $||\eta||_2^2 > \alpha$ (where $\alpha = \gamma(1 + \gamma)^{-1}$) and the last otherwise.

$$vol(A) = \int_{||\eta||_2 \leq \alpha} V_k(\gamma^{-\frac{1}{2}}||\eta||)^k d\eta + \int_{\alpha < ||\eta||^2 \leq 1} V_k(1 - ||\eta||^2)^{\frac{k}{2}} d\eta \tag{2.11}$$

$$= kV_k^2\left(\gamma^{-\frac{k}{2}} \int_0^{\sqrt{\alpha}} r^{2k-1} dr + \int_{\sqrt{\alpha}}^1 (1 - r^2)^{\frac{k}{2}} r^{k-1} dr\right) \tag{2.12}$$

Then:

$$p_k = \frac{vol(S(\Delta_{\mathcal{A},\mathcal{B}}))}{V_{nN}} = \frac{\gamma^{-\frac{k}{2}}\alpha^k}{k\beta\left(\frac{k}{2}, \frac{k}{2} + 1\right)} + 1 - I_\alpha\left(\frac{k}{2}, \frac{k}{2} + 1\right) \tag{2.13}$$

$\square$

Proof of Theorem 2.1, part 2. Using that $k\beta\left(\frac{k}{2}, \frac{k}{2} + 1\right) \geq 2^{-k}$[1]:

$$\frac{\gamma^{-\frac{k}{2}}\alpha^k}{k\beta\left(\frac{k}{2}, \frac{k}{2} + 1\right)} \leq r_1^k \tag{2.14}$$

where $r_1 = 2\frac{\gamma^{\frac{1}{2}}}{1+\gamma}$. Since $F > 0$ it is true that $0 < \frac{2\gamma^{\frac{1}{2}}}{1+\gamma} < 1$, then the first term of p_k goes to zero.

Now, we need to show that $1 - I_\alpha\left(\frac{k}{2}, \frac{k}{2} + 1\right) \to 0$ as $k \to \infty$. Let $s \in \mathbb{R}$, the inequalities $(r(1 - r))^s \leq r^{s-1}(1 - r)^s \leq (r(1 - r))^{s-1}$ holds for $0 \leq r \leq 1$.

Let $g : [0, 1] \to \mathbb{R}$ be the map $r \mapsto r(1-r)$, then $g \in L^1([0, 1])$ and so $||g||_p \to ||g||_\infty$ when $p \to \infty$[2]. Since $||g||_\infty = \frac{1}{4}$ we have that $\forall \epsilon > 0 \,\exists k_0(\epsilon) \in \mathbb{R}$ such that for all $k \geq k_0$ we get:

$$\left(\frac{1}{4} - \epsilon\right)^k \leq \int_0^1 (r(1 - r))^k dr \leq \left(\frac{1}{4} + \epsilon\right)^k \tag{2.15}$$

---

[1]Easily follows from the sharper inequality $\beta(x, y) \geq \frac{1}{xy}$ [6].
[2]Note that this holds because the interval $[0, 1]$ has finite measure.

Since $\frac{1}{2} < \alpha < 1$ and g is a decreasing function in $\left[\frac{1}{2}, 1\right]$ we take ϵ such that $\alpha(1-\alpha) < \frac{1}{4} - \epsilon$ to get that:

$$1 - I_\alpha \left(\frac{k}{2}, \frac{k}{2} + 1 \right) = \frac{\int_\alpha^1 r^{\frac{k}{2}-1}(1-r)^{\frac{k}{2}} dr}{\int_0^1 r^{\frac{k}{2}-1}(1-r)^{\frac{k}{2}} dr} \tag{2.16}$$

$$\leq \frac{(1-\alpha)\,\alpha^{\frac{k}{2}-1}(1-\alpha)^{\frac{k}{2}}}{\int_0^1 (r(1-r))^{\frac{k}{2}} dr} \tag{2.17}$$

$$\leq \frac{(1-\alpha)\,\alpha^{\frac{k}{2}-1}(1-\alpha)^{\frac{k}{2}}}{\left(\frac{1}{4} - \epsilon\right)^{\frac{k}{2}}} \tag{2.18}$$

$$\leq \left(\frac{1}{\alpha} - 1 \right) r_2^k \tag{2.19}$$

Where $r_2 = \left(\frac{\alpha(1-\alpha)}{\frac{1}{4}-\epsilon} \right)^{\frac{1}{2}} = \left(\frac{1}{\frac{1}{4}-\epsilon} \right)^{\frac{1}{2}} \frac{\gamma^{\frac{1}{2}}}{1+\gamma} < 1$, showing that the last term of p_k also goes to zero. Setting $r = \max(r_1, r_2) < 1$ and $D = \frac{1}{\alpha}$ we have that $p_k \leq Dr^k$. $\qquad\square$

Remark 2.1. *Theorem 2.1 implies that p_k exponentially approaches zero as k grows. The decreasing of p_k with k is significant even for low values of k, as it can be seen in Figure 2.2. The same figure shows that there is a strong dependence between the behavior of p_k and F.*
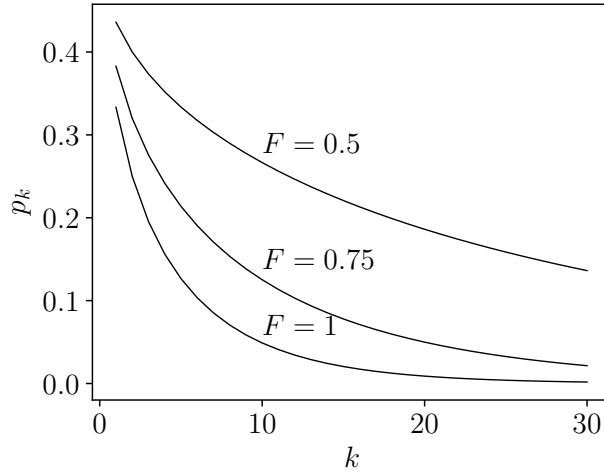
Figure 2.2: Variation of $p_k(F)$ as a function of k, for different values of F.

Using Theorem 2.1, it is possible to estimate the probability $\tilde{p}_k(F)$ that at least one individual in a randomly uniformly chosen population with N individuals will be improved in one iteration if using only k-crossovers. Noticing that given two k-configurations $(\mathcal{A}, \mathcal{B})$ and $(\mathcal{A}', \mathcal{B}')$, the events $\rho \in \Delta_{\mathcal{A},\mathcal{B}}$ and $\rho \in \Delta_{\mathcal{A}',\mathcal{B}'}$ are not necessarily independent, then:

$$p_k(F) \leq \tilde{p}_k(F) \leq N p_k(F) \tag{2.20}$$

Since $p_k(F) \to 0$ as $k \to \infty$ and assuming that the population size N is polynomial on n, it is true that $\tilde{p}_k(F) \to 0$ when $k \to \infty$, showing that the number of evasive configurations grows with the number of dimensions used on the mutation.

2.2.2   Introducing the crossover probability

Now, allowing the crossover operator to use or not any dimension, the Theorem 2.2 presents the expression of the probability $p_{C_r,n}(F)$ that in a randomly uniformly chosen population $\rho \in B_{n \times N}$ a given individual will be improved after one iteration.

Theorem 2.2 (Resende and Takahashi). *Consider the objective function $f(x) = ||x||_2^2$ and a population $\rho = \{x_1, \cdots, x_N\}$ such that it is randomly uniformly distributed inside the set $B_{n \times N}$. Let the individuals labeled as x_{i_1}, x_{i_2}, x_{i_3}, x_{i_4} be chosen randomly, with uniform probability and without replacement, from the population. Let also the offspring o be defined by equation (1.7), with the elements of set $\mathcal{A}$ chosen randomly from the set $\{1, \ldots, n\}$ by a Bernoulli trial with probability of success given by C_r. For a fixed scaling factor F, let $p_{C_r,n}(F)$ denote the probability that $f(o) < f(x_1)$. Then:*

$$p_{C_r,n}(F) = \sum_{k=1}^{n} \binom{n}{k} C_r^k (1 - C_r)^{n-k} p_k(F) \tag{2.21}$$

Proof. The number of elements in the set $\mathcal{A}$, given by k, becomes an outcome of a random variable sampled from a binomial distribution with success probability C_r. The expression (2.21) comes from the application of (2.1), summing over all possible values of k. $\square$

Theorem 2.3 shows an upper bound for $p_{C_r,n}(F)$ as a function of the number of problem dimensions n.

Theorem 2.3. *Let $C_r \in [0,1]$ be a real number and $p_{C_r,n}(F)$ be defined as in equation (2.21), then $p_{C_r,n}(F) \le Cs^n$, where $C \in \mathbb{R}$ comes from Theorem 2.1 and $s \in (0,1)$.*

Proof of Theorem 2.3. We know that $p_k \le Cr^k$ for some $C \in \mathbb{R}$ and $0 < r < 1$, then:

$$p_{C_r,n} \le C \sum_{k=0}^{n} \binom{n}{k} (rC_r)^k (1 - C_r)^{n-k} \tag{2.22}$$

Since $(rC_r)^k (1 - C_r)^{n-k} \le s^n \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}$[3] with $s = 1 - C_r + rC_r$:

$$p_{C_r,n} \le Cs^n \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \le Cs^n \tag{2.23}$$

---

[3]To see that take the map $h : x \mapsto x^k(s-x)^{n-k}$ with $x \in [0,s]$, then we have $h'(x) = 0 \Rightarrow x = s\frac{k}{n}$.

Where the last inequality comes from $\left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \leq \frac{1}{2^n}$ and $\sum_{k=0}^{n} \binom{n}{k} = 2^n$. $\square$

Define $\tilde{p}_{C_r,n}(F)$ as the probability that a randomly uniformly chosen population with N individuals can be improved on the first iteration of the DE. We estimate:

$$p_{C_r,n}(F) \leq \tilde{p}_{C_r,n}(F) \leq N p_{C_r,n}(F) \tag{2.24}$$

And, analogously to $\tilde{p}_k(F)$, it is true that $\tilde{p}_{C_r,n}(F) \to 0$ as $n \to \infty$. In the next subsections we analyse some implications of these results. Note that Theorem 2.2 rigorously states the curse of dimensionality to the Differential Evolution algorithm since for any fixed parameter configuration the algorithm fails to provide enhancements as n grows.

2.3 Fixed points and improvement quality

Proposition 2.1. *Let* $\mathbb{P}(A) = \frac{vol(A)}{vol(B_{n \times N})} = vol(A)$ *for every borelian A, where vol is the Lebesgue measure. For a fixed $N \geq 4$ and $n \geq 1$ it holds that*

$$\lim_{F \to \infty} \mathbb{P}\left(\Lambda_1^n\right) = 1 \tag{2.25}$$

and for a fixed $F > 0$ and $N \geq 4$ it holds that

$$\lim_{n \to \infty} \mathbb{P}\left(\Lambda_n^n\right) = 1 \tag{2.26}$$

Proof of Proposition 2.1. First notice that $\Lambda_1^n = B_{n \times N} - \cup_{(\mathcal{A},\mathcal{B})}\Delta_{\mathcal{A},\mathcal{B}}$, where the union is over all 1-configurations. Considering that the number of 1-configurations is $4!\binom{N}{4}n$, then:

$$\mathbb{P}\left(\Lambda_1^n\right) \geq 1 - \sum_{(\mathcal{A},\mathcal{B})} \mathbb{P}(\Delta_{\mathcal{A},\mathcal{B}}) = 1 - 4!n\binom{N}{4}p_1(F) \tag{2.27}$$

By definition:

$$p_1(F) = \frac{\gamma^{-\frac{1}{2}}\alpha}{\beta\left(\frac{1}{2}, \frac{1}{2} + 1\right)} + 1 - I_\alpha\left(\frac{1}{2}, \frac{1}{2} + 1\right) \tag{2.28}$$

where $\gamma = 2F^2 + 1$ and $\alpha = \frac{\gamma}{1+\gamma}$. Since $\lim_{F \to \infty} \alpha = 1$, then, by Lebesgue dominated convergence theorem, $\lim_{F \to \infty} I_\alpha\left(\frac{1}{2}, \frac{1}{2} + 1\right) = 1$ and $\lim_{F \to \infty} \gamma^{-\frac{1}{2}} = 0$, thereby $\lim_{F \to \infty} p_1(F) = 0$. Using equation (2.27) we conclude that $\lim_{F \to \infty} \mathbb{P}\left(\Lambda_1^n\right) = 1$.

We prove the second limit analogously. Since $\Lambda_n^n = B_{n \times N} - \cup_{(\mathcal{A},\mathcal{B})} \Delta_{\mathcal{A},\mathcal{B}}$, where the union is over all n-configurations and considering that the number of n-configurations is $4! \binom{N}{4}$, then

$$\mathbb{P}(\Lambda_n^n) \geq 1 - \sum_{(\mathcal{A},\mathcal{B})} \mathbb{P}(\Delta_{\mathcal{A},\mathcal{B}}) = 1 - 4! \binom{N}{4} p_n(F) \to 1 \tag{2.29}$$

since $p_n(F) \to 0$ when $n \to \infty$. $\qquad \square$

Remark 2.2. *Note that if (1.7) holds then the norm of at least one coordinate of o decreases in relation to x_{i_1}, so $\Lambda_k^n \subset \Lambda_1^n$ for all $k = 1, \cdots, n$. Then the set of fixed points of the DE with crossover rate $C_r \in (0,1)$ is just Λ_1^n.*

Theorem 2.1 shows that the relative size of the set of fixed points goes to one as n increases (justifying the experimental results from Tables 2.1, 2.2 and 2.3) and that the value of F is critical to control the amount of fixed points on the algorithm.

The probabilities $p_k(F)$ and $p_{C_r,n}(F)$ give us information about the existence or not of an improvement, but don't tell us about how good or bad this improvement can be. Let $\delta \in (0,1)$, an offspring o of w is δ-better then w if $f(o) \leq \delta f(w)$, a population $\rho = (x_1, \cdots, x_N)$ is δ-improvable if there is an individual x_i capable of produce an offspring o_i that is δ-better then x_i. The next theorem shows that the improvements can be as bad as we want them to be.

Proposition 2.2. *For every $\delta \in (0,1)$ there is a set $B_\delta \in B_{n \times N}$ of positive probability such that every $\rho \in B_\delta$ is not δ-improvable.*

Proof of Proposition 2.2. If $\rho = (x_1, \cdots, x_n)$ is δ-improvable and $x_i \neq 0$ for every $i = 1, \cdots, n$, then there exists some x_i and a possible offspring o_i such that $\delta \geq \frac{f(o_i)}{f(x_i)}$. Therefore the minimum possible δ is given by $\min \frac{f(o_i)}{f(x_i)}$, where the minimum is taken over all individuals x_i and over of all their possible offspring o_i. Let $x^* \neq 0$ be an individual such that $\rho^* = (x^*, \cdots, x^*) \in B_{n \times N}$ and $\rho = (x_1, \cdots, x_n)$ be a population such that $||\rho - \rho^*||_2 < \epsilon$ for some $\epsilon > 0$. By the definition of the offspring and the triangular inequality

$$||o_i - x^*||_2 < (1 + 2F)\epsilon \tag{2.30}$$

and since $||\rho - \rho^*||_2 < \epsilon$:

$$||x_i - x^*||_2 < \epsilon \tag{2.31}$$

Then

$$\lim_{\epsilon \to 0} \frac{f(o_i)}{f(x_i)} = 1 \tag{2.32}$$

for every x_i $i = 1, \cdots, n$ and every possible offspring. Since the number of individuals and possible offspring is finite we define $B_\delta \in B_{n \times N}$ as the ball centered on ρ^* with radius $\epsilon_\delta = \min \epsilon_{x_i,o_i}$ where ϵ_{x_i,o_i} is such that $\frac{f(o_i)}{f(x_i)} > \delta$ for every $\epsilon < \epsilon_{x_i,o_i}$. $\qquad \square$

2.4   The dependence between improvements

Let $(\mathcal{A}_1, \mathcal{B}_1), \cdots, (\mathcal{A}_s, \mathcal{B}_s)$ be $k_1, k_2, \cdots, k_s$ configurations, respectively. We want to study the probability that a randomly uniformly chosen $\rho \in B_{n \times N}$ belongs to $\cap_{i=1}^{s} \Delta_{\mathcal{A}_i, \mathcal{B}_i}$. Given a k-contraction $(\mathcal{A}, \mathcal{B})$ and a population $\rho = (x_1, \cdots, x_N)$ we say that $x_{i,d}$ is an active entry for $(\mathcal{A}, \mathcal{B})$ if $i \in \mathcal{A}$ and $d \in \mathcal{B}$. We say that the configurations

$$(\mathcal{A}_1, \mathcal{B}_1), \cdots, (\mathcal{A}_s, \mathcal{B}_s)$$

act disjointly on a population ρ if they have no active entry in common. Since configurations may share active entries one may expect some dependence.

Let X be the random variable that counts the number of common active entries between all the s configurations above. We may write $X = \sum X_{x_{i,d}}$ where $X_{x_{i,d}}$ is the indicator variable of the event $x_{i,d}$ is a common active entry for all the s configurations. Setting $\mathbb{E}$ to be the expectation over all choices of $k_1, \cdots, k_s$ configurations, we have

$$\mathbb{E}X = \sum \mathbb{E}X_{x_{i,d}} = nN\mathbb{E}X_{x_{1,1}} = nN \prod_{i=1}^{s} \frac{4k_i}{nN} \tag{2.33}$$

since $\mathbb{P}(X_{x_{i,d}} = 1) = \prod_{i=1}^{s} \frac{4k_i}{nN}$ because we first verify if they share the same i and then the same d. Using the Markov Inequality:

$$\mathbb{P}(X = 0) \geq 1 - nN \prod_{i=1}^{s} \frac{4k_i}{nN} \tag{2.34}$$

On Tables 2.1, 2.2 and 2.3 one may observe that, for $N \leq n$ the quantity m_{ic} decreases until certain k and then starts to increase. A possible explanation for this behavior is that when k starts to increase the improvements become more dependent and the occurrence of one improvement leads, with high probability, to the occurrence of another one.

If we don't restrict ourselves only to k-configurations, allowing the crossover operator to select different values of k, a similar analysis leads to $\mathbb{E}X = nN \left(\frac{4C_r}{N}\right)^s$, showing that lower values of C_r are also effective promoting configurations that act disjointly.

2.5   Choosing parameters

For a fixed n, the value of $p_{C_r,n}(F)$ is a function of C_r and F, Figure 2.3 shows how those variables are related for different values of n. The existence of a simple and

explicit expression for $p_{C_r,n}$ gives a strategy to choose C_r: given the values of F and n one may choose $C_r = C_r^*$, where C_r^* is the one that maximizes $p_{C_r,n}(F)$.
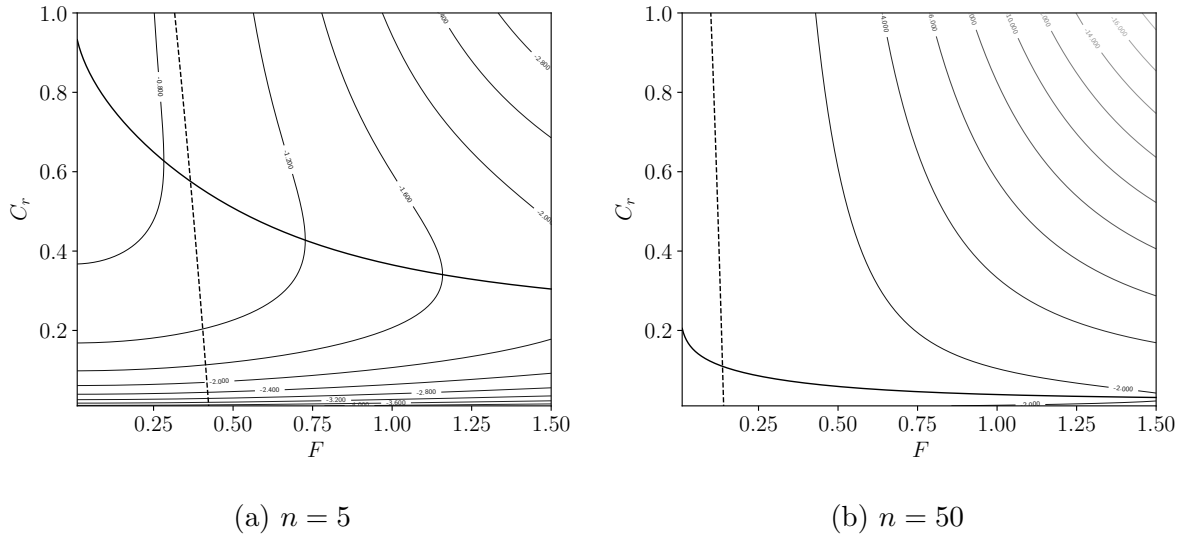

(a) $n = 5$ (b) $n = 50$

Figure 2.3: Contour plots of $\ln\left(p_{C_r,n}\right)$. The dashed line is the critical F for each C_r given by equation 1.34 and the thicker line is the C_r value that maximizes $p_{C_r,n}$ for each F.

One must keep in mind that C_r^* only maximizes the probability that a single choice of four individuals in ρ will produce an improvement after one iteration assuming that ρ is uniformly distributed on $B_{n \times N}$. It does not take into account the dependencies between individuals when looping through several choices and it also does not take into account that after various iterations we may have a population distribution not even close to an uniform distribution. And, of course, the value of $p_{C_r,n}$ was calculated for the function $f(x) = ||x||^2$ and might not work as well for other functions. Therefore, the choice of $C_r = C_r^*$ is a heuristics.

Until now we have evaluated some expressions that led us to a heuristics. In the next chapter some experiments will help us clarify how the algorithm runs on certain scenarios and how helpful our heuristics is.

Chapter 3

Experimental results

The analytical results from the last chapter concerns only about the first iteration of the algorithm, as we will see in the next chapter the analytical understanding of the interactions between generations seems to require new developments on dynamical systems. Therefore, to support the value of our analytical results to the full run we performed some experiments.

3.1 Empirical enhancements for the sphere function

A numerical experiment is performed in order to measure the empirical proportion of enhancements attained with the operation (1.7) when the individuals x_i are distributed uniformly[1] on the unit ball B_n. The function f_1 is defined as:

$$f_1(x) = x^T x \tag{3.1}$$

Function f_1 has a unique point of minimum, which corresponds to the origin of the space. The minus gradient of this function induces a vector field which has a zero-dimensional attractor (the point of minimum) and which points toward a line passing through this point, from any point of the space.

All theoretical predictions should hold for this function. Different combinations of n and C_r are adopted, and all simulations employ the scaling factor $F = 0.75$. For each combination of parameters, the sequence of steps described in the following *protocol for 1-step experiments* is executed 10000 times:

---

[1]Notice that this is different from the condition assumed in Theorem 2.2, which requires $\rho \in B_{n \times N}$. The experiment is performed with the more usual condition $\rho \subset B_n$, and the results obtained are essentially the same that would be obtained with the exact condition.

Protocol for 1-step experiments

(i) a population ρ with N individuals is generated randomly choosing N individuals independently and uniformly in B_n;

(ii) each individual in ρ is chosen once as x_1, and (x_2, x_3, x_4) are chosen randomly from the remainder individuals;

(iii) for each x_1, a set $\mathcal{A}$ is generated by n Bernoulli trials performed sequentially on the elements of the set $\{1, \ldots, n\}$ with probability of each such element being included in $\mathcal{A}$ given by C_r;

(iv) the operation defined by (1.7) is executed with the configuration $(\mathcal{A}, \mathcal{B})$ where $\mathcal{B} = (1, 2, 3, 4)$, generating an offspring o;

(v) if $f(o) < f(x_1)$, then the counter of the number of enhancements achieved with that combination of n and C_r is increased by one.

Table 3.1 presents the results of this experiment, comparing the value of $p_{C_r,n}$ obtained by evaluating expression (2.21) with the empirical value $\hat{p}_{C_r,n}$ obtained by dividing the count of the number of enhancements by the number of trials. This experiment shows that the expression (2.21) provides an accurate prediction of the probability of enhancement of DE individuals for uniformly distributed populations in the minimization of function f_1. We also point that $p_{C_r,n}$ belongs to the 95% confidence interval[2] for all experiments.

Another set of experiments is performed in order to examine what happens in the case of different objective functions. The particular case of functions which are not separable is focused. The following unit vectors are defined for being used in the construction of functions with suitable features:

$$
v_{ns} = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad v_s = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.2}
$$

The vector v_{ns} will be used to define an eigenvector of non-separable functions, and the vector v_s will define an eigenvector of separable functions.

---

[2] The $(1 - \alpha)$-confidence interval was estimated by the classical normal approximation as $\hat{p}_{C_r,n} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_{C_r,n}(1-\hat{p}_{C_r,n})}{n}}$, where $z_{1-\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ percentile of the standard normal distribution. For $\alpha = 5\%$ we have $z \approx 1.96$.

n	C_r	$p_{C_r,n}$ eq. (2.21)	$\hat{p}_{C_r,n}$	95% confidence interval
8	0.125	0.2313	0.2311	(0.2228, 0.2394)
16	0.125	0.2790	0.2798	(0.2710, 0.2886)
32	0.125	0.2473	0.2481	(0.2396, 0.2566)
8	0.500	0.2469	0.2470	(0.2385, 0.2555)
16	0.500	0.1581	0.1593	(0.1521, 0.1665)
32	0.500	0.0738	0.0741	(0.0690, 0.0792)
8	0.875	0.1719	0.1719	(0.1645, 0.1793)
16	0.875	0.0862	0.0863	(0.0808, 0.0918)
32	0.875	0.0257	0.0259	(0.0228, 0.0290)

Table 3.1: Results from a numerical experiment for testing the expression (2.21) for the probability of enhancement $p_{C_r,n}$ in the optimization of function f_1. Each line shows: the problem dimension n; the crossover rate C_r; the analytical value of $p_{C_r,n}$ provided by eq. (2.21); the empirical estimation of $p_{C_r,n}$, indicated by $\hat{p}_{C_r,n}$. All runs were performed with the scaling factor $F = 0.75$.

In order to establish two limiting cases, the functions f_2 and f_3 are defined. Function f_2 corresponds to the squared distance from x to the line which crosses the origin, in the direction of vector v:

$$f_2(x) = w^T w \tag{3.3}$$

in which w is defined as:

$$w = x - v\, x^T v$$

The minus gradient of f_2 induces a vector field pointing towards the line $x = \kappa\, v$. Therefore, this vector field has a 1-dimensional attractor, given by this line. The function f_2 becomes f_2^s for $v = v_s$, and becomes f_2^{ns} for $v = v_{ns}$.

Function f_3 corresponds to the squared distance from x to the $(n-1)$-dimensional hyperplane that crosses the origin and is perpendicular to vector v:

$$f_3(x) = (x^T v)^2 \tag{3.4}$$

The minus gradient of f_3 induces a vector field that points towards the $(n-1)$-dimensional hyperplane defined by $x^T v = 0$. Therefore, this vector field has an $(n-1)$-dimensional attractor, represented by this hyperplane. Function f_3 becomes f_3^s for $v = v_s$, and becomes f_3^{ns} for $v = v_{ns}$.

Finally, the quadratic function f_4 is defined:

$$f_4(x) = x^T Q\, x \tag{3.5}$$

(a) f_4^{ns} with $\phi = 10^{-1}$.   (b) f_4^{ns} with $\phi = 10^{-4}$.   (c) f_2.

(d) f_4^{ns} with $\phi = 10$.   (e) f_4^{ns} with $\phi = 10^4$.   (f) f_3.
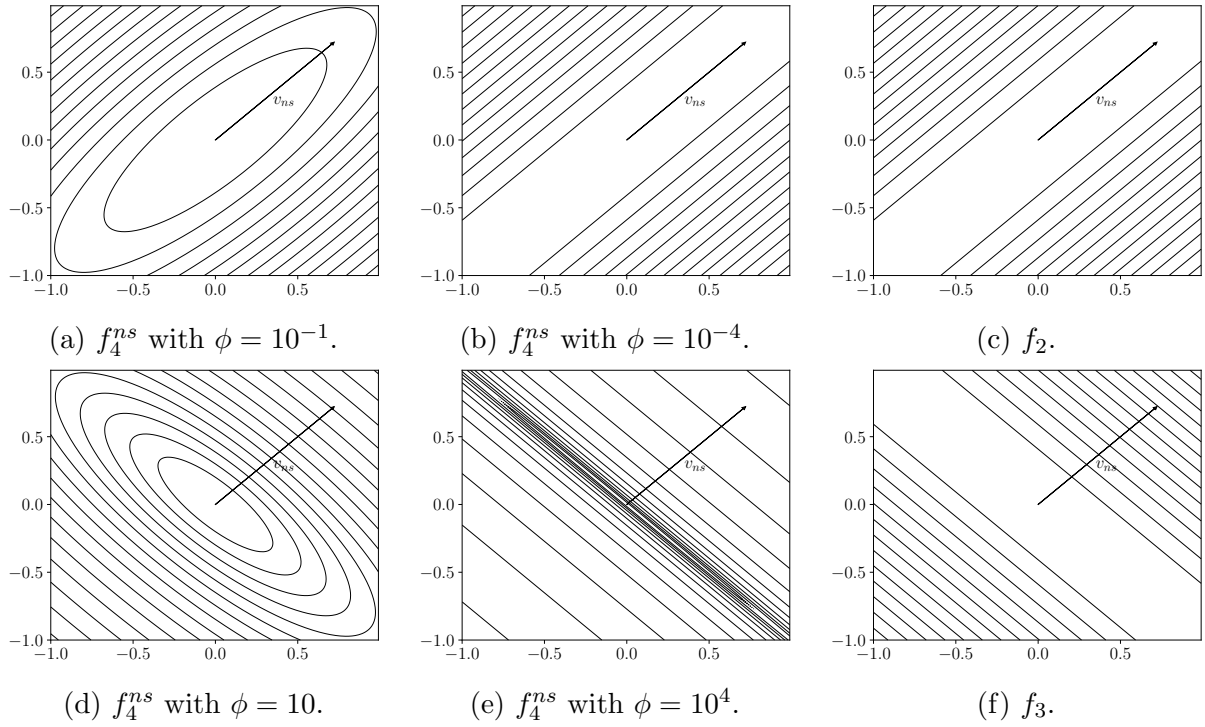
Figure 3.1: Contours of functions f_2, f_3 and f_4^{ns}. Notice that the levels curves of function f_4^{ns} approximates the level curves of function f_2 as $\phi \to 0^+$ or the level curves of function f_3 as $\phi \to \infty$.

with Q given by:

$$Q = V^T D V$$

$$V = \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix}$$

$$v_i^T v_j = 0 \;\forall j \neq i; \qquad v_i^T v_i = 1$$

$$D = \text{diag}(\phi, 1, 1, \dots, 1)$$

Function f_4 becomes f_4^s for $v_1 = v_s$, and becomes f_4^{ns} for $v_1 = v_{ns}$. It should be noticed that:

$$\phi = 1 \Rightarrow f_4(x) = f_1(x) \qquad \lim_{\phi \to 0^+} f_4(x) = f_2(x) \qquad \lim_{\phi \to \infty} f_4(x) = f_3(x) \qquad (3.6)$$

When $\phi \approx 0$, the structure of functions generated by f_4 may be understood as being composed of two scales: in the first scale, the line $x = \kappa v$ works as an attractor. In the second scale, the point $x = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}^T$ becomes an atractor, inside the first atractor. When $\phi \gg 0$, that structure also behaves as being composed of two scales: in the first scale, the hyperplane $x^T v = 0$ works as an attractor. In the second scale, the point $x = \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix}^T$ becomes an atractor inside the first atractor.

The first experiment with non-separable functions is performed considering the functions f_2^{ns} and f_4^{ns}, the last one with $\phi = 10^{-2}$ and $\phi = 10^{-4}$. Those functions represent the situation in which the vector field of function enhancement is 1-dimensional, in an

n	C_r	$p_{C_r,n-1}$ f_1	$\hat{p}_{C_r,n}$ f_2^{ns}	$\hat{p}_{C_r,n}$ f_4^{ns} $\phi = 10^{-4}$	$\hat{p}_{C_r,n}$ f_4^{ns} $\phi = 10^{-2}$
9	0.125	0.2313	0.2506	0.2513	0.2498
17	0.125	0.2790	0.2859	0.2853	0.2853
33	0.125	0.2473	0.2477	0.2474	0.2481
9	0.500	0.2469	0.2483	0.2491	0.2482
17	0.500	0.1581	0.1579	0.1587	0.1593
33	0.500	0.0738	0.0736	0.0742	0.0742
9	0.875	0.1719	0.1715	0.1723	0.1721
17	0.875	0.0862	0.0870	0.0870	0.0860
33	0.875	0.0257	0.0258	0.0260	0.0256

Table 3.2: Empirical probability of enhancement $\hat{p}_{C_r,n}$, for functions with 1-dimensional attractor. The columns represent: the problem dimension n; the value of C_r; the value of $p_{C_r,n-1}$ (calculated analytically), for function f_1; the empirical values of $p_{C_r,n}$, indicated by $\hat{p}_{C_r,n}$ for functions f_2^{ns} and f_4^{ns}. In the case of f_4^{ns}, the values $\phi = 10^{-4}$ and $\phi = 10^{-2}$ are considered. All runs were performed with the scaling factor $F = 0.75$.

exact sense in the case of function f_2^{ns} and in an approximate sense in the case of the two instances of function f_4^{ns}. Table 3.2 presents in each row: the problem dimension n; the crossover rate C_r; the analytical value of $p_{C_r,n-1}$ (which represents the probability of enhancement of a function f_1 in a problem with dimension $n-1$); the empirical evaluation of the probability of enhancement, indicated by $\hat{p}_{C_r,n}$, for functions f_2^{ns} and f_4^{ns} in n dimensions. In the case of f_4^{ns}, the results are presented for $\phi = 10^{-4}$ and $\phi = 10^{-2}$. In all cases, the estimate $\hat{p}_{C_r,n}$ is calculated using 10000 runs of the *protocol for 1-step experiments*, with the scaling factor $F = 0.75$.

From Table 3.2, it is interesting to notice that:

$$p_{C_r,n-1}[f_1] \approx \hat{p}_{C_r,n}[f_2^{ns}] \approx \hat{p}_{C_r,n}[f_4^{ns}(\phi = 10^{-4})] \approx \hat{p}_{C_r,n}[f_4^{ns}(\phi = 10^{-2})] \qquad (3.7)$$

for all pairs (n, C_r), with an error order greater than 10^{-3} only in the cases $(9, 0.125)$ and $(17, 0.125)$. This suggests that when the geometric pattern of the possible enhancements is organized as a vector field pointing towards a line in dimension n, the probability of enhancement behaves in the same way as in the case when the geometrical pattern of enhancements is organized as a vector field pointing towards one point in dimension $n-1$. It is worthy to mention that another set of experiments was conducted with functions f_2^s and f_4^s. In that experiment, the relations (3.7) held in all cases, including the two ones in which the former experiment presented some discrepancy.

The second experiment with non-separable functions is conducted considering the functions f_3^s, f_3^{ns}, f_4^s and f_4^{ns}. Both f_4^s and f_4^{ns} are considered with $\phi = 10^2$ and $\phi = 10^4$. Those functions represent the situation in which the vector field of function enhancements is $(n-1)$-dimensional, in an exact sense in the case of functions f_3^{ns} and f_3^s, and in an approximate sense in the case of all instances of functions f_4^{ns} and f_4^s. Table 3.3 presents

n	C_r	$p_{C_r,1}$ f_1	$\hat{p}_{C_r,n}$ f_3^{ns}	$\hat{p}_{C_r,n}$ f_4^{ns} $\phi = 10^2$	$\hat{p}_{C_r,n}$ f_4^{ns} $\phi = 10^4$	$\hat{p}_{C_r,n}$ f_3^s	$\hat{p}_{C_r,n}$ f_4^s $\phi = 10^2$	$\hat{p}_{C_r,n}$ f_4^s $\phi = 10^4$
9	0.125	0.0478	0.3204	0.3146	0.3198	0.0481	0.2490	0.2498
17	0.125	0.0478	0.4146	0.4708	0.4142	0.0477	0.2905	0.2922
33	0.125	0.0478	0.4789	0.4239	0.4781	0.0478	0.2618	0.2648
9	0.500	0.1914	0.4226	0.4701	0.4237	0.1920	0.3043	0.3143
17	0.500	0.1914	0.4245	0.3790	0.4240	0.1910	0.2520	0.2701
33	0.500	0.1914	0.4233	0.3434	0.4214	0.1913	0.1952	0.2281
9	0.875	0.3349	0.3926	0.3574	0.3910	0.3346	0.3270	0.3572
17	0.875	0.3349	0.3931	0.3574	0.3908	0.3352	0.2919	0.3445
33	0.875	0.3349	0.3935	0.2853	0.3897	0.3349	0.2471	0.3362
9	1.000	0.3828	0.3827	0.3437	0.3816	0.3826	0.3439	0.3826
17	1.000	0.3828	0.3829	0.3141	0.3802	0.3834	0.3148	0.3820
33	1.000	0.3828	0.3817	0.2682	0.3800	0.3821	0.2671	0.3811

Table 3.3: Empirical probability of enhancement $\hat{p}_{C_r,n}$, for non-separable functions with $(n-1)$-dimensional attractor. The columns represent: the problem dimension n; the value of C_r; the analytical value of $p_{C_r,1}$ (the probability of enhancement of function f_1 in the case of a 1-dimensional problem); the empirical values of $p_{C_r,n}$, indicated by $\hat{p}_{C_r,n}$ for functions f_3^{ns}, f_3^s, f_4^{ns} and f_4^s. In the case of f_4^{ns} and f_4^s, the values $\phi = 10^2$, $\phi = 10^4$ are considered. All runs were performed with the scaling factor $F = 0.75$ and initial population distributed uniformly in B_n.

in each row: the problem dimension n; the crossover rate C_r; the analytical value of $p_{C_r,1}$ (the probability of enhancement of function f_1 in the case of a 1-dimensional problem); the empirical evaluation of the probability of enhancement, indicated by $\hat{p}_{C_r,n}$, for functions f_3^{ns}, f_3^s, f_4^{ns} and f_4^s. In all cases, the estimate $\hat{p}_{C_r,n}$ is based on 10000 runs of the *protocol for 1-step experiments*, using the scaling factor $F = 0.75$.

Function f_3 represents an extremal situation in which the possible enhancements are organized as a vector field pointing towards a hyperplane of dimension $n-1$. The enhancements occur, in this case, in a single dimension, along the normal direction to the hyperplane. An interesting information presented in Table 3.3 is that $\hat{p}_{C_r,n}$ for function f_3^s, in n dimensions, corresponds to the value of $p_{C_r,1}$ – the probability of enhancement in a 1-dimensional f_1 problem. It should be noticed that the minimization of function $f_3^s(x)$ means the same as the minimization of the absolute value of the first component of x, no matter what happens with the other $n-1$ components – which explains why the approximate equality $\hat{p}_{C_r,n}[f_3^s] \approx p_{C_r,1}[f_1]$ holds for all values of n and C_r. An important fact about DE algorithm is revealed when the same experiment is performed with function f_3^{ns}. This function corresponds to a simple rotation of f_3^s, now putting the hyperplane's normal vector in the direction $v_{ns} = [\ 1\ \cdots\ 1\ ]^T$. This means that, in the case of f_3^{ns}, all components of x matter for enhancing the solution. By virtue of this, the probabilities of enhancement $p_{C_r,n}[f_3^{ns}]$ become significantly greater than in the case of function f_3^s.

n	C_r	$\hat{p}_{C_r,n}$ f_1	$\hat{p}_{C_r,n}$ f_2^{ns}	$\hat{p}_{C_r,n}$ f_3^{ns}	$\hat{p}_{C_r,n}$ f_4^{ns} $\phi = 10^{-4}$	$\hat{p}_{C_r,n}$ f_4^{ns} $\phi = 10^{-2}$	$\hat{p}_{C_r,n}$ f_4^{ns} $\phi = 10^{2}$	$\hat{p}_{C_r,n}$ f_4^{ns} $\phi = 10^{4}$
9	0.125	0.3508	0.3504	0.3487	0.3502	0.3492	0.3493	0.3496
17	0.125	0.4470	0.4474	0.4480	0.4467	0.4471	0.4492	0.4482
33	0.125	0.4910	0.4879	0.4941	0.4905	0.4907	0.4921	0.4946
9	0.500	0.4986	0.4985	0.4978	0.4966	0.4980	0.5000	0.4980
17	0.500	0.4983	0.4965	0.5000	0.4979	0.4981	0.4996	0.5000
33	0.500	0.4969	0.4912	0.4998	0.4928	0.4931	0.5014	0.4996
9	0.875	0.4986	0.4997	0.4995	0.4988	0.4986	0.4995	0.4994
17	0.875	0.4990	0.4975	0.4993	0.4983	0.4984	0.4997	0.5009
33	0.875	0.4958	0.4956	0.5006	0.4961	0.4946	0.5000	0.4999

Table 3.4: Estimated probability of enhancement $p_{C_r,n}$, when the initial population is far away the function enhancement attractor. The columns represent: the problem dimension n; the value of C_r; the empirical values of $p_{C_r,n}$, indicated by $\hat{p}_{C_r,n}$ for functions f_1, f_2^{ns}, f_3^{ns} and f_4^{ns}, this last one with $\phi = 10^2$ and $\phi = 10^{-2}$. All runs were performed with the scaling factor $F = 0.75$. The initial population, in all cases, is distributed uniformly in a unit ball centered in a point located at a distance $d = 1000$ from the function attractor.

This feature will have an important role in the global behavior of DE algorithm. The last observation on Table 3.3 refers to the comparison between $\hat{p}_{C_r,n}[f_3^s]$ and the two instances of $\hat{p}_{C_r,n}[f_4^s]$. It can be seen that $\hat{p}_{C_r,n}[f_3^s] < \hat{p}_{C_r,n}[f_4^s]$ for all combinations of (C_r, n), which may be explained by noticing that, in the case of functions f_4^s, there is the possibility that an enhancement occur due to some components of vector x other than the first one.

A third experiment is conducted considering the functions f_1, f_2^{ns}, f_3^{ns} and f_4^{ns}, this last one with $\phi = 10^2$ and $\phi = 10^{-2}$. Now, the element to be investigated is the effect of the initial population being distant from the function attractor. In this way, the population is generated with uniform distribution inside a unit ball whose center is located at a distance of 1000 in relation to each attractor. In the cases of functions f_1, f_3^{ns} and f_4^{ns} with $\phi = 10^2$, the center is located on the direction of vector $v_{ns} = [\begin{array}{ccc} 1 & \cdots & 1 \end{array}]^T$, and in the cases of functions f_2^{ns} and f_4^{ns} with $\phi = 10^{-2}$ it is located on a direction orthogonal to v_{ns}. Table 3.4 presents in each row: the problem dimension n; the crossover rate C_r; the empirical evaluation of the probability of enhancement, indicated by $\hat{p}_{C_r,n}$, for functions f_1, f_2^{ns}, f_3^{ns} and f_4^{ns}, this last one considering the cases $\phi = 10^2$ and $\phi = 10^{-2}$. In all cases, the estimate $\hat{p}_{C_r,n}$ is calculated by 10000 executions of the *protocol for 1-step experiments*, using the scaling factor $F = 0.75$.

The main conclusion that can be extracted from Table 3.4 is that, for all considered functions, the probability of enhancement at a large distance from the attractor is approximately equal to 0.5. This effect can be explained by the following reasoning: (i) at a large distance, the function contour curves in the neighborhood of the DE population becomes similar to hyperplanes; (ii) every time an offspring o is compared to the respective individual x_{i_1}, the offspring may be on either sides of the contour curve that

passes on x_{i_1}, with the same probability. Therefore, the probability of enhancement is 0.5. It should be noticed that even the cases of (n, C_r) equal to $(9, 0.125)$ and $(17, 0.125)$ don't contradict this reasoning: the smaller probability of enhancement, in that cases, comes from the cases in which the offspring remains equal to x_{i_1}, which occur due to the non-negligible probability of no component of the vector resulting from mutation being selected for crossover, with that combinations of parameters. Notice that, discounting those situations of no mutation, the analytical values of the probability of enhancement for the optimization of f_1 in the case of initial population far from the origin become $p_{C_r,n}^\infty = \frac{1}{2} - \frac{1}{2}(1 - C_r)^n$ and then $p_{0.125,9}^\infty = 0.3497$ and $p_{0.125,17}^\infty = 0.4483$.

3.2   The dynamics of a full run

The analysis of the probability of enhancement of the individuals suggests that the dynamics of DE population may be described as a multi-step process. For instance, in the case of the simple function f_1, considering an initial population which is launched far from the point of minimum, the steps are:

1. The population approaches the attractor (the point of minimum). In this process, it becomes more spread.

2. After having included the attractor within its convex hull, the population starts to contract around the attractor, until the convergence.

In this subsection, some experiments considering $n = 3$ are performed in order to allow the visualization of the results. This multi-phase behavior of DE algorithm are examined in those experiments.

Figure 3.2 shows a sequence of populations, that start in a unit ball whose center is situated at a distance 20 from the point of minimum. The population becomes more spread until it reaches that point, and then starts a contraction towards it.

Figure 3.3 shows the evolution of the distance of the population to the point of minimum of function f_1 through the iterations, in a logarithm scale. Five curves are shown, representing the last individual among: the 10% best ones, the 25% best ones, the 50% best ones, the 75% best ones and the 90% best ones. The DE algorithm is run with a population of 50 individuals, $C_r = 0.5$ and $F = 0.75$.

It can be seen that, in the first 20 iterations, the distance to the point of minimum decreased very slowly – which corresponds to the phase of population expansion. After
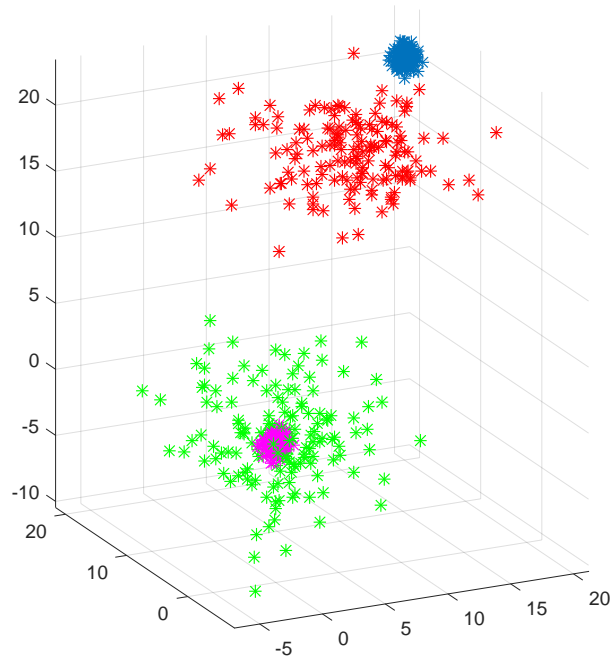
Figure 3.2: Sequence of populations. The initial population is shown in blue, and the next ones following the order: red, green, magenta. The point of minimum, which is the origin, is located within the cloud of the last population.
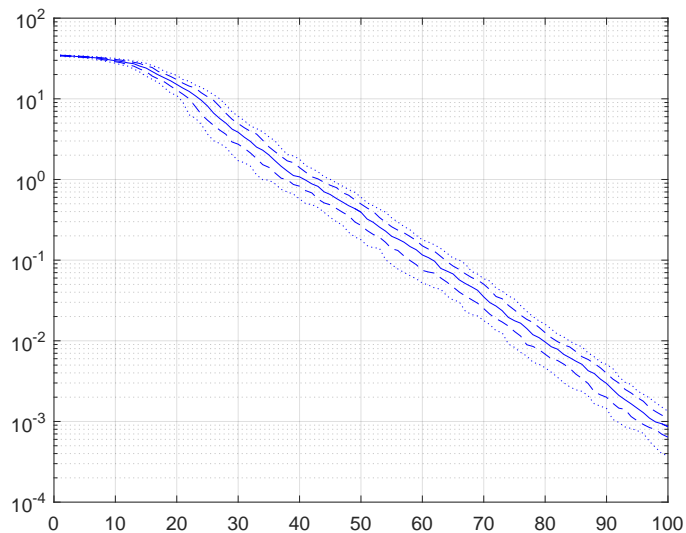

Figure 3.3: Distance from the individuals to the point of minimim of function f_1. Five curves are shown, representing the last individual among: the 10% best ones, the 25% best ones, the 50% best ones, the 75% best ones and the 90% best ones, in each iteration of the optimization process.

the 20th iteration, that distance started a process of exponential decrease, at a rate of

more than 10^3 for each 60 iterations.

One should notice that the description of this two-phase behavior is consistent with the information provided in Tables 3.1 and 3.4. A high probability of enhancement when the population is far from the attractor means that a large proportion of the new individuals are accepted, leading to an increase on the spread of population. A smaller probability of enhancement when the population cloud includes the attractor means that only the new individuals that are created towards the point of minimum are accepted.

Table 3.5 shows the empirical estimation of the probability of enhancement, measured on the DE population in some moments of the optimization process. This measurement is performed in the following way:

- The DE algorithm is run until it reaches each value of iteration indicated in the first column of Table 3.5. Once reaching that number of iterations, the algorithm is interrupted, and the estimation of the probability of enhancement on that iteration is performed as explained next. After each interruption, the algorithm proceeds from the point it was interrupted.

- When the algorithm is interrupted, the current population of DE on that iteration is used as the initial population in an experiment in which only one iteration of DE is performed, and the number of enhancements is registered. This experiment is repeated 2000 times, with the same initial population. The average number of enhancements per individual per repetition is taken as the empirical probability of enhancement of that population on that specific iteration number.

This table shows that before starting the optimization process, the probability of enhancement of the initial population is 0.4379, which is approximately equal to the analytical value of the probability of enhancement calculated for function f_1 with initial population far from the origin, $p_{0.5,3}(\infty) = 0.4375$. This probability decreases until nearly the 45th iteration, when it reaches a value near to 0.30, which is kept up to the end of the run. The analytical value of $p_{C_r,n}$ for function f_1 and initial population with uniform distribution in a unit ball around the origin is $p_{0.5,3} = 0.2980$.

The results presented in Table 3.5 suggest that, in the case of function f_1, each one of the two phases of the process of convergence towards the problem solution is characterized by a particular value of the probability of enhancement: the first phase, when far from the attractor, by $\frac{1}{2}$ and the second phase, after including the attractor within its convex hull, by $p_{C_r,n}$.

More complex multi-phase dynamics are observed in the case of objective functions with attractors which have multiple scales with higher dimensions in the initial scales. An experiment is performed for the optimization of function f_4^{ns}, $\phi = 10^{-4}$, with the initial population within a unit ball around the origin and parameters $N = 150$, $C_r = 0.5$ and $F = 0.75$. Figure 3.4 shows the initial population and the 20th iteration population. It can

iteration	$\hat{p}_{C_r,n}$
0	0.4379
15	0.4334
30	0.3676
45	0.3018
60	0.3011
75	0.2966
90	0.2994
105	0.2976
120	0.2951
135	0.3018
150	0.2933

Table 3.5: Empirical estimates of the probability of enhancement measured on the DE population on different moments along the process of optimization of function f_1 with initial population within a unit ball whose center is located at a distance 20 in relation to the origin. The DE parameters are $C_r = 0.5$, $F = 0.75$, and the problem dimension is $n = 3$. The analytical value of $p_{C_r,n}$ for function f_1 and initial population with uniform distribution in a unit ball around the origin, $p_{0.5,3} = 0.2980$, is approximated by the empirical estimates $\hat{p}$ after iteration 45.

be seen that the initial population is located within a unit sphere, while the 20th iteration population becomes distributed along a line (the function first-scale attractor). Figure 3.5 shows the evolution of the distances from the individuals to the point of minimum and from the individuals to the attractor line. The range of distances to the the point of minimum initially grows approximately on the first 25 iterations, and then starts to decrease exponentially. The distances to the attractor line initially decrease fast on the first 25 iterations, and then their rate of decrease reduces, but after the 25th iteration the decrease is still exponential. In all steps after the 25th iteration, all individuals keep moving on the attractor line, in a process that becomes approximately one-dimensional.

Once again, it should be noticed that the behavior presented in figures 3.4 and 3.5 is consistent with the information presented in Table 3.2. That table essentially says that, on the initial stages of the evolution of population in this problem, the probability of an enhancement which moves the individual towards the attractor line is greater than the probability of an enhancement which moves it towards the point of minimum. Therefore, there is an average movement towards the attractor line before the final convergence towards the minimum.

A very important feature of the function f_4^{ns} is that, if the population becomes distributed exactly on the first-scale attractor (a line), the improvements on the individuals of such a population will require that all dimensions are changed simultaneously. Any movement which involves less dimensions will move the point away from the attractor, being non-enhancing. In the usual case when the population becomes distributed approximately on the attractor, there will be movements involving less than n dimensions which
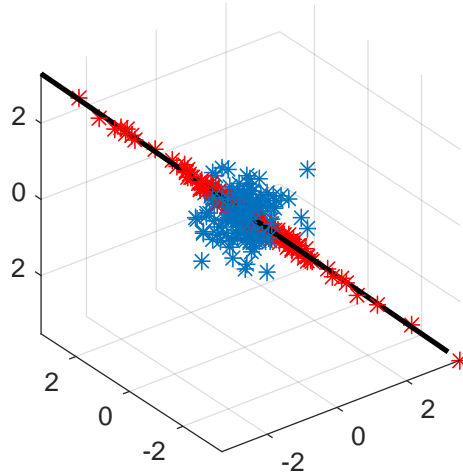
Figure 3.4: The initial population (blue) and the 20th iteration population (red) of a DE in the minimization of f_4^{ns} with $\phi = 10^{-4}$. A black line represents the first-scale attractor of the objective function.
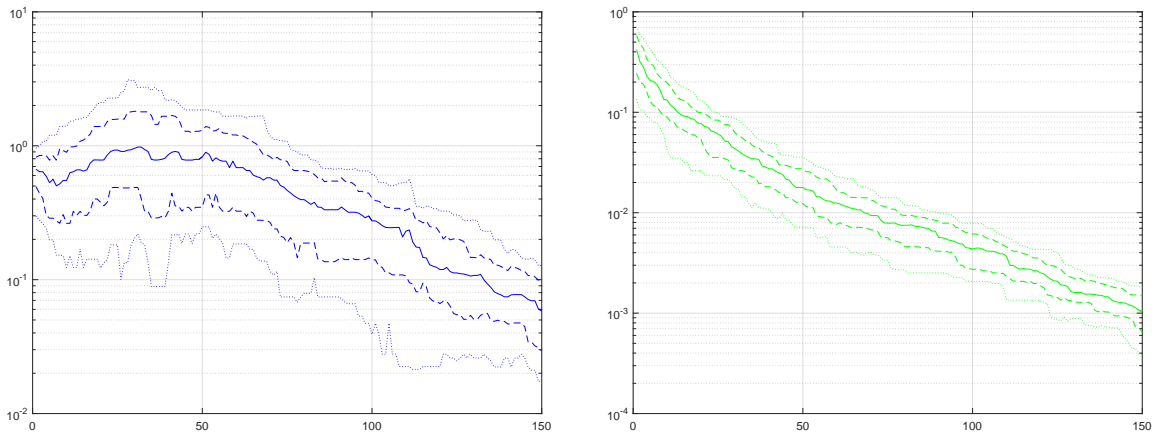

Figure 3.5: Distance, in logarithm scale, from the individuals: (left) to the point of minimim of function f_4^{ns}, $\phi = 10^{-4}$; and (right) to the line that represents the first attractor of the function. Five curves are shown, representing the last individual among: the 10% nearest ones, the 25% nearest ones, the 50% nearest ones, the 75% nearest ones and the 90% nearest ones, in each iteration of the optimization process.

may represent an enhancement, but those enhancements will be necessarily small. Table 3.6 presents the empirical probability of enhancement for this experiment and also for similar experiments considering different values of C_r. Those empirical probabilities are measured in the same way as in the experiments reported in Table 3.5. The table also presents the average number of dimensions that are involved in the enhancements, $\bar{k}$, as a proportion of the number of problem dimensions, n.

It can be seen that, for $C_r \in \{1/2, 2/3, 1\}$, except in the iterations 0 and 15, more than 90% of the dimensions were involved in the enhancements. As $n = 3$, this means that the large majority of enhancements involved all the three dimensions. Even

iteration	$C_r = 1/3$		$C_r = 1/2$		$C_r = 2/3$		$C_r = 1$	
	$\bar{k}$	$\hat{p}$	$\bar{k}$	$\hat{p}$	$\bar{k}$	$\hat{p}$	$\bar{k}$	$\hat{p}$
0	0.4797	0.2741	0.5535	0.3240	0.6703	0.3491	1	0.3189
15	0.5328	0.0682	0.7812	0.0681	0.9512	0.1006	1	0.2759
30	0.6546	0.0304	0.9056	0.0424	0.9779	0.0868	1	0.2760
45	0.7263	0.0212	0.9190	0.0419	0.9737	0.0856	1	0.2754
60	0.7821	0.0168	0.9252	0.0402	0.9721	0.0866	1	0.2754
75	0.8102	0.0151	0.9314	0.0406	0.9759	0.0862	1	0.2719
90	0.8319	0.0141	0.9336	0.0398	0.9752	0.0862	1	0.2790
105	0.8367	0.0141	0.9339	0.0405	0.9772	0.0869	1	0.2726
120	0.8292	0.0148	0.9328	0.0398	0.9765	0.0861	1	0.2740
135	0.8042	0.0157	0.9258	0.0399	0.9754	0.0854	1	0.2760
150	0.8020	0.0156	0.9277	0.0400	0.9729	0.0869	1	0.2734

Table 3.6: The average number of dimensions involved in each enhancement, $\bar{k}$, presented as a fraction of n, and the empirical probability of enhancement, $\hat{p}$, considering the DE populations occurring on different moments along the process of optimization of function f_4^{ns}, $\phi = 10^{-4}$ with initial population within a unit ball around the origin, for different values of C_r. The scaling factor is set as $F = 0.75$, and the problem dimension is $n = 3$.

in the case of $C_r = 1/3$, in which most of the mutations involved only one dimension, the enhancements after iteration 75 involved 80% of the dimensions, on average. It should be noticed that, in all cases, the population was committed to approach the function attractor on the first iterations, which allowed enhancements involving 1 or 2 dimensions. After the population becoming disposed along the attractor, most of the enhancements should involve all 3 dimensions. This means that the average number of dimensions involved in enhancements grows when the population reaches the stage in which it is disposed along the attractor and starts to contract along it.

It is interesting to notice that the formula for $p_k(F)$, presented in equation (2.1), seems to provide some accurate information regarding this experiment. The idea is to produce an estimate of the steady-state value of $\hat{p}$ in the cases presented in Table 3.6 by the formula:

$$p(C_r) = p_k(F)|_{k=3, F=0.75} \cdot (C_r)^n|_{n=3} \qquad (3.8)$$

This formula calculates the probability of enhancement as the product of the probability of 3 dimensions being selected for crossover, given by $(C_r)^n$, by the probability of enhancement if 3 dimensions are involved in crossover, given by (2.1). The application of this formula leads to: $p(1) = 0.2759$, $p(2/3) = 0.0819$, $p(1/2) = 0.0345$, $p(1/3) = 0.0102$. It can be seen that $p(1)$ is quite accurate, and $p(2/3)$ is still reasonable accurate. The values of $p(1/2)$ and $p(1/3)$, on the other hand, underestimate the respective empirical values because, in those cases, there is a significant number of crossover trials with less than 3 dimensions involved, which leads to some enhancements provided by those trials that become relevant, although small. The difference between the behavior of the DE algorithm in the case $C_r = 1/3$ and the case $C_r = 1$ can be observed in Figure 3.6. The

much faster convergence of the case $C_r = 1$ is explained both by the higher probability of enhancement and by the higher size of enhancement, compared with the enhancements that occur by crossover with less than 3 dimensions.
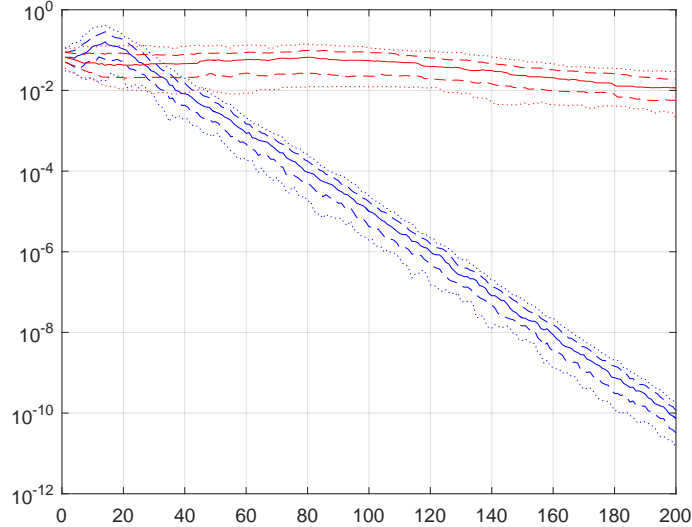

Figure 3.6: Distance, in logarithm scale, from the individuals to the point of minimim of function f_4^{ns}, $\phi = 10^{-4}$. Blue: DE algorithm with $C_r = 1$. Red: DE algorithm with $C_r = 1/3$. Five curves are shown in each case, representing the last individual among: the 10% nearest ones, the 25% nearest ones, the 50% nearest ones, the 75% nearest ones and the 90% nearest ones, in each iteration of the optimization process.
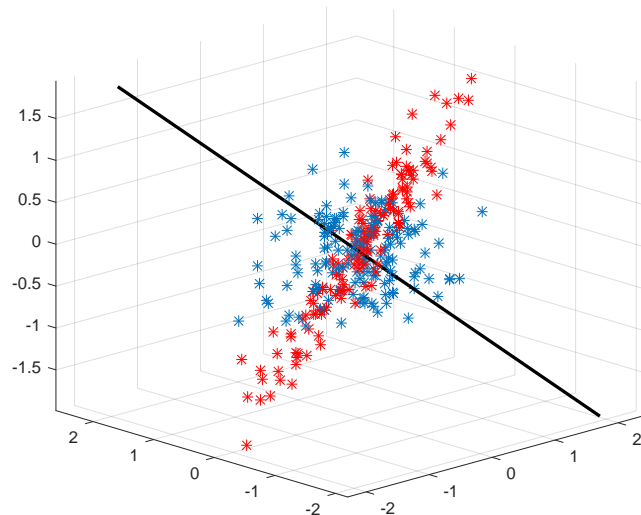

Figure 3.7: The initial population (blue) and the 20th iteration population (red) of a DE in the minimization of f_4^{ns} with $\phi = 10^4$. A black line represents the normal direction to the plane which is the first-scale attractor of the objective function.

Another experiment is performed for the optimization of function f_4^{ns}, $\phi = 10^4$, with the initial population within a unit ball around the origin and parameters $N = 150$, $C_r = 0.5$ and $F = 0.75$. Figure 3.7 shows the initial population and the 20th iteration

population. It can be seen that the initial population is located within a unit sphere, while the 20th iteration population becomes distributed along a plane (the function first-scale attractor) whose normal direction is also presented as a black line. Figure 3.8 shows the evolution of the distances from the individuals to the point of minimum and from the individuals to the attractor plane. The range of distances to the the point of minimum initially grows approximately on the first 25 iterations, and then starts to decrease exponentially. The distances to the attractor plane initially decrease fast on the first 25 iterations, and then reduces the rate of decrease; after the 25th iteration the decrease is still exponential. In all steps after the 25th iteration, all individuals keep moving on the attractor plane, in a $(n-1)$-dimensional process.
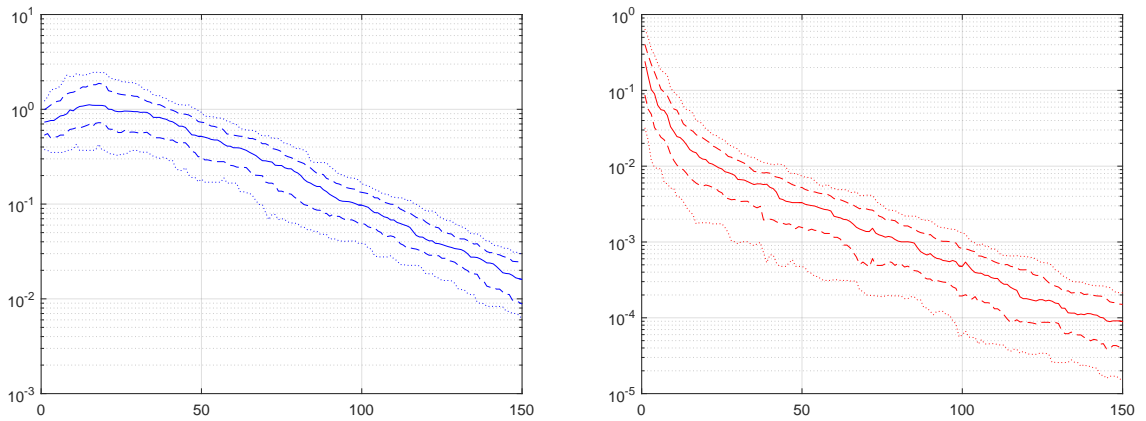

Figure 3.8: Distance, in logarithm scale, from the individuals to: (left) the point of minimim of function f_4^{ns}, $\phi = 10^4$; and (right) the plane that represents the first attractor of the function. Five curves are shown, representing the last individual among: the 10% nearest ones, the 25% nearest ones, the 50% nearest ones, the 75% nearest ones and the 90% nearest ones, in each iteration of the optimization process.

As in the case of function f_4^{ns} with $\phi = 10^{-4}$, also for f_4^{ns} with $\phi = 10^4$ it occurs that if the population becomes distributed exactly on the first-scale attractor (now a plane), the improvements on the individuals of such a population will require that all dimensions are changed simultaneously. Any movement which involves less dimensions will move the point away from the attractor, being non-enhancing. Again, in the usual case when the population becomes distributed approximately on the attractor, there will be movements involving less than n dimensions which may represent an enhancement, but those enhancements will be necessarily small. Table 3.7 presents the empirical probability of enhancement for this function, considering different values of C_r. Those empirical probabilities are measured in the same way as in the experiments reported in Table 3.5. The table also presents the average number of dimensions that are involved in the enhancements, $\bar{k}$, as a proportion of the number of problem dimensions, n.

iteration	$C_r = 1/3$		$C_r = 1/2$		$C_r = 2/3$		$C_r = 1$	
	$\bar{k}$	$\hat{p}$	$\bar{k}$	$\hat{p}$	$\bar{k}$	$\hat{p}$	$\bar{k}$	$\hat{p}$
0	0.4643	0.3115	0.5563	0.3855	0.6787	0.3974	1.0000	0.3885
15	0.5997	0.0433	0.8715	0.0515	0.9575	0.0956	1.0000	0.2700
30	0.7187	0.0223	0.9158	0.0417	0.9739	0.0873	1.0000	0.2757
45	0.7686	0.0181	0.9295	0.0397	0.9724	0.0878	1.0000	0.2723
60	0.8046	0.0160	0.9320	0.0405	0.9710	0.0869	1.0000	0.2739
75	0.8274	0.0152	0.9361	0.0398	0.9745	0.0885	1.0000	0.2725
90	0.8184	0.0151	0.9408	0.0392	0.9748	0.0882	1.0000	0.2727
105	0.8228	0.0147	0.9333	0.0398	0.9748	0.0857	1.0000	0.2753
120	0.8250	0.0146	0.9292	0.0407	0.9764	0.0869	1.0000	0.2761
135	0.8250	0.0148	0.9188	0.0401	0.9724	0.0876	1.0000	0.2795
150	0.8354	0.0144	0.9445	0.0399	0.9771	0.0887	1.0000	0.2726

Table 3.7: The average number of dimensions involved in each enhancement, $\bar{k}$, presented as a fraction of n, and the empirical probability of enhancement, $\hat{p}$, considering the DE populations occurring on different moments along the process of optimization of function f_4^{ns}, $\phi = 10^4$ with initial population within a unit ball around the origin, for different values of C_r. The scaling factor is set as $F = 0.75$, and the problem dimension is $n = 3$.

The most noticeable feature of Table 3.7 is that it does not differ significantly from Table 3.6 in any entry. This suggests that the probability of enhancement, in the case of non-separable functions with two scales does not depend on the dimension of the attractor of the first scale. Again, a close approximation of those probabilities can be performed by using equation (2.1).

One last experiment is performed in order to examine the role of the non separateness of functions in the former results. The function f_4^s is employed, with $\phi = 10^{-4}$ and $\phi = 10^4$, initial population within a unit ball around the origin and parameters $N = 150$ and $F = 0.75$. Now, only $C_r = 1/3$ and $C_r = 1$ are considered. The results are presented in Table 3.8. It can be seen that the cases with $C_r = 1$ both lead to the same results of the non-separable case. On the other hand, now the cases with $C_r = 1/3$ have a greater probability of enhancement, which approaches $\hat{p} = 0.247$. This empirical probability is near to the value $p_{Cr,n}|_{C_r=1/3,n=3} = 0.2515$.

Figure 3.9 shows that the DE algorithm performs similarly with $C_r = 1/3$ and with $C_r = 1$, in this experiment, both for the case $\phi = 10^4$ and $\phi = 10^{-4}$. The version with $C_r = 1$, in both cases, presents a delay due to the need for the population reaching the function attractor before starting the contraction phase. In the case of the version with $C_r = 1/3$, the contraction starts from the beginning.

Remark 3.1. *It should be noticed that figures 3.3, 3.6, 3.8 and 3.9 show an interesting pattern: the algorithm, in all those cases, seems to reach a kind of steady-state behavior after a transient initial phase. In this steady-state, the relative distance between quantiles becomes unchanged iteration after iteration, in the logarithm scale. As in all those figures*

iteration	$\phi = 10^{-4}$ $C_r = 1/3$		$\phi = 10^{-4}$ $C_r = 1$		$\phi = 10^4$ $C_r = 1/3$		$\phi = 10^4$ $C_r = 1$	
	$\bar{k}$	$\hat{p}$	$\bar{k}$	$\hat{p}$	$\bar{k}$	$\hat{p}$	$\bar{k}$	$\hat{p}$
0	0.4653	0.2604	1.0000	0.3202	0.4723	0.2640	1.0000	0.3874
15	0.4640	0.2421	1.0000	0.2769	0.4670	0.2465	1.0000	0.2759
30	0.4753	0.2493	1.0000	0.2748	0.4776	0.2447	1.0000	0.2702
45	0.4782	0.2472	1.0000	0.2790	0.4770	0.2469	1.0000	0.2765
60	0.4745	0.2480	1.0000	0.2756	0.4765	0.2438	1.0000	0.2735
75	0.4762	0.2469	1.0000	0.2771	0.4777	0.2433	1.0000	0.2712
90	0.4773	0.2478	1.0000	0.2778	0.4779	0.2427	1.0000	0.2740
105	0.4781	0.2438	1.0000	0.2761	0.4768	0.2429	1.0000	0.2722
120	0.4773	0.2469	1.0000	0.2731	0.4780	0.2428	1.0000	0.2686
135	0.4759	0.2475	1.0000	0.2768	0.4767	0.2461	1.0000	0.2708
150	0.4757	0.2419	1.0000	0.2748	0.4780	0.2464	1.0000	0.2754

Table 3.8: The average number of dimensions involved in each enhancement, $\bar{k}$, presented as a fraction of n, and the empirical probability of enhancement, $\hat{p}$, considering the DE populations occurring on different moments along the process of optimization of function f_4^s, with $\phi = 10^4$ and $\phi = 10^{-4}$ with initial population within a unit ball around the origin, for different values of C_r. The scaling factor is set as $F = 0.75$, and the problem dimension is $n = 3$.
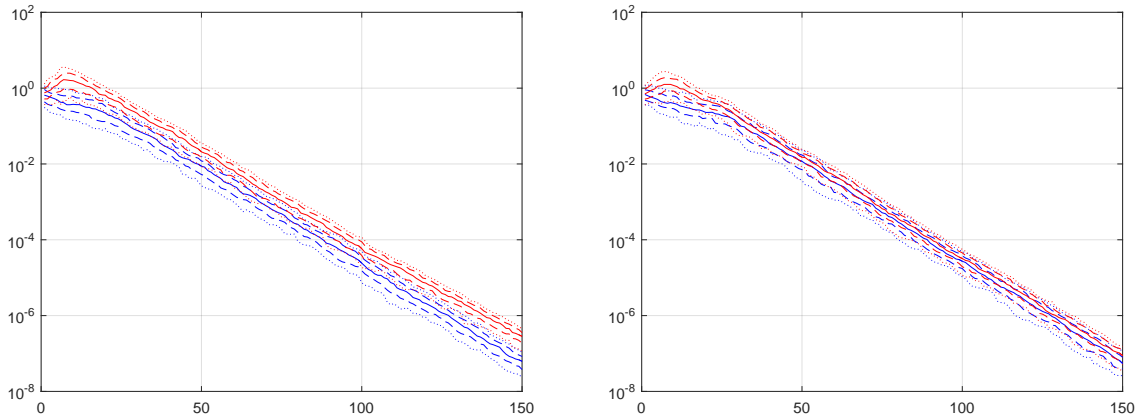

Figure 3.9: Distance, in logarithm scale, from the individuals to the point of minimum of function f_4^s, with $\phi = 10^4$ (left) and $\phi = 10^{-4}$ (right). Blue: DE algorithm with $C_r = 1$. Red: DE algorithm with $C_r = 1/3$. Five curves are shown in each case, representing the last individual among: the 10% nearest ones, the 25% nearest ones, the 50% nearest ones, the 75% nearest ones and the 90% nearest ones, in each iteration of the optimization process.

there is a constant negative slope in the distances which are represented, those data suggest that in the steady-state the population becomes distributed according to a geometric pattern which is modified by a contraction which preserves the geometry and produces the convergence of population towards the point of optimal. This conjecture is also consistent

with the existence of a similar steady-state behavior of the probability of convergence, as indicated in tables 3.5, 3.6, 3.7 and 3.8.

3.3 Behavior in Large Dimensions

Define $\tilde{p}_{pop}(C_r, F, n, N)$ as the probability that at least one individual from a random population of N individuals uniformly distributed within $B_{n \times N}$ is improved on the first iteration of the DE with parameters F and C_r, in the case of objective function f_1 with decision variable space of dimension n. Assuming independence between the probability of enhancement of different individuals, the expression of $\tilde{p}_{pop}(C_r, F, n, N)$ is given by:

$$\tilde{p}_{pop}(C_r, F, n, N) = 1 - (1 - p_{C_r,n}(F))^N \tag{3.9}$$

The Theorem 2.3 shows that the probability of an individual being improved after one iteration approaches zero, for large values of n. And, analogously to what happens with $\tilde{p}_k(F)$ when $k \to \infty$, it is true that $\tilde{p}_{C_r,F,n,N} \to 0$ as $n \to \infty$. Expression (3.9) reveals what is the effect of growing the population size N in compensating the effect of growth of n. The data corresponding to the evaluation of this expression is presented in Table 3.9. It becomes clear that, when $n > 256$, for most populations generated randomly there will be no enhancement in the first iteration of DE, so the algorithm will not even start.

The evaluation of (3.9), as presented in Table 3.9 suggests that the DE algorithm will not be useful in arbitrarily large dimensions. This table also suggests that the DE may be useful up to a maximal dimension that is somewhere between $n = 100$ and $n = 200$. In this section, the behavior of the DE algorithm is examined as the problem dimension grows up to a value of n within that range. In order to establish a baseline for comparison, a basic Evolutionary Strategy (ES) algorithm is employed, as described next.

n	$p_{C_r,n}(F)$	$p_{pop}(C_r, F, n, N)$ $N = n$	$p_{pop}(C_r, F, n, N)$ $N = 3n$
2	0.2714	0.4691	0.8504
4	0.2998	0.7596	0.9861
8	0.2469	0.8966	0.9989
16	0.1581	0.9363	0.9997
32	0.0738	0.9141	0.9994
64	0.0193	0.7127	0.9763
128	0.0016	0.1897	0.4679
256	1.54e-05	0.0039	0.0117
512	1.80e-09	9.21e-07	2.7623e-06
1024	1.14e-16	1.14e-13	3.4106e-13

Table 3.9: Probability of enhancement of a population, for different problem dimensions. n is the problem dimension. $p_{C_r,n}(F)$ is the probability of enhancement of a given individual, from a population with uniform distribution, calculated by expression (2.21). p_{pop} is the probability of enhancement of at leas one individual in a population with $N = n$ individuals with uniform distribution. In all cases, it was considered $C_r = 0.5$ and $F = 0.75$.

Algorithm 3: Basic Evolution Strategy algorithm, $(\lambda + 1)$ variant.

---

1 Generate initial population of N_{ES} individuals within a sphere of radius $R = R_0$ around the initial point x_0, with uniform probability

2 $x_{op} \leftarrow$ the point in population with the best objective function value

3 **while** *not stopping condition* **do**

4   Generate population of N_{ES} individuals within a sphere of radius R around the point x_{op}, with uniform probability

5 $n_{enh} \leftarrow$ number of new individuals better than x_{op}

6   **if** $n_{enh} > 0$ **then**

7 $x_{op} \leftarrow$ the point in population with the best objective function value

8   **else**

9 $R \leftarrow R/1.2$

10   **if** $n_{enh} > N_{ES}/5$ **then**

11 $R \leftarrow 1.5R$

The Figure 3.10(a) presents the comparison of one run of DE algorithm with one run of the basic ES algorithm for the minimization of function f_1 considering a decision variable space of $n = 5$ dimensions, with both algorithms having a budget of 10000 function evaluations. Figure 3.10(b) presents a similar comparison, for function f_4^{ns} with $\phi = 10^{-4}$, also considering $n = 5$, and both algorithms having a budget of 100000 function evaluations. Both (a) and (b) are performed with ES population $N_{ES} = 5$, and DE population $N = 10$.
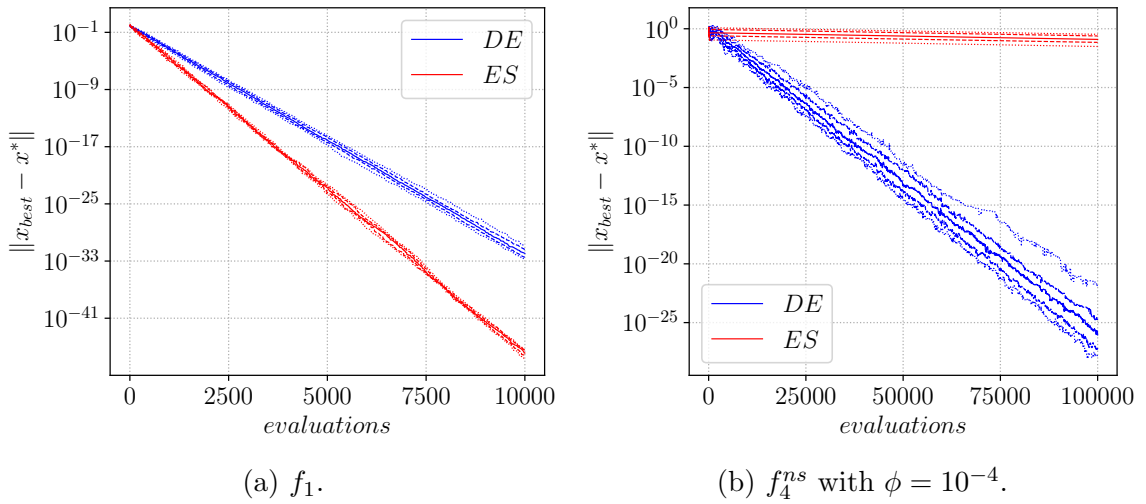
(a) f_1.

(b) f_4^{ns} with $\phi = 10^{-4}$.

Figure 3.10: (a) Distance, in logarithm scale, from the best solution to the point of minimum of function f_1; DE: blue, ES: red. (b) Distance, in logarithm scale, from the best solution to the point of minimum of function f_4^{ns}, with $\phi = 10^4$; DE: blue, ES: red. In all cases, $n = 5$. Each curve was produced with 50 executions, representing the 10% (lower dotted),25% (lower dashed),50% (solid), 75% (upper dashed) and 90% (upper dotted) quantiles. Both populations (for DE and ES) were initialized on $[-1, 1]^n$ and the ES with $R_0 = 1$.

The data presented in Figure 3.10 makes it clear that the DE algorithm is relevant mainly for problems in which the objective function has "elongated" contour curves. In fact, the distribution of the DE population according to the contour curves on each iteration, as studied in the last section, provides an implicit mechanism for compensating the curvature of function graphics. In this way, the DE is able to reach a much tighter approximation of the function optimal in functions like f_4^{ns}. In the case of functions with contour curves that are nearly spherical, like f_1, the basic ES outperforms the DE.

The issue to be studied in this section is the effect of the growth of the decision space dimension. In the case of function f_1, the ES outperforms the DE by a large margin in all experiments, for any value of n. In fact, when n reaches nearly 185, the DE starts to present a behavior in which the ability of enhancing the solution is lost: the initial solution is not improved at all, in most of runs. In those cases, the initial population of DE had become a *fixed point*. On the other hand, the experiments conducted here have found that the ES retains its capability of enhancing the objective function f_1 at least up to $n = 10000$.

Considering now the function f_4^{ns} with $\phi = 10^{-4}$, new experiments were performed for $n \in \{15, 25, 50, 100\}$. Again, the experiments are performed with ES population $N_{ES} = 5$, and DE population $N = 10$. Figure 3.11 shows the results of those experiments.

In all cases the DE outperforms the ES, reaching better approximations of the point of optimal. In all cases, the DE stopped the process of approaching the point of optimal when a fixed point was reached. In order to examine the issue of the fixed point, another
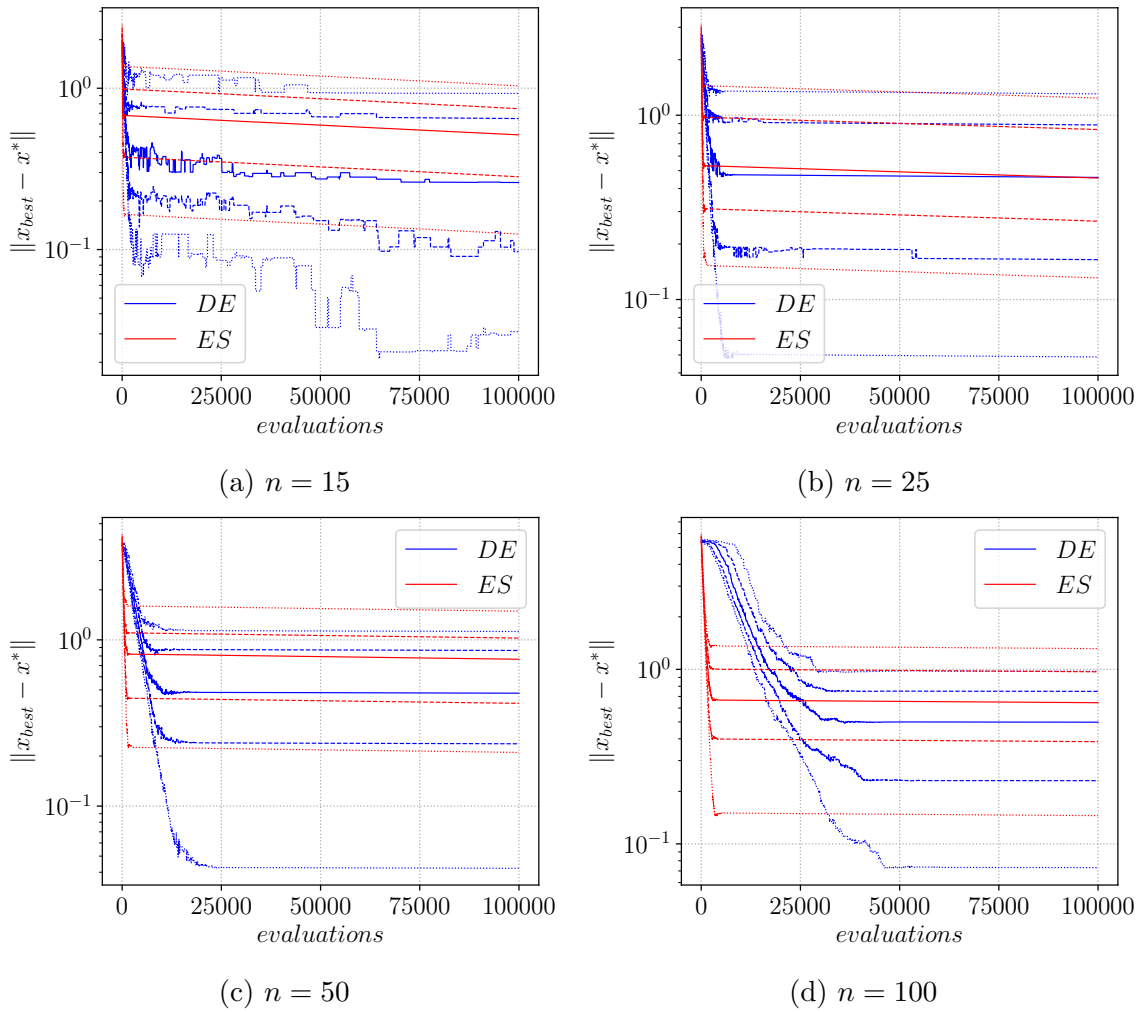
Figure 3.11: Distance, in logarithm scale, from the best solution to the point of minimum of function f_4^{ns} with $\phi = 10^{-4}$. In all cases: DE: blue, ES: red. Each curve was produced with 50 executions, representing the 10% (lower dotted),25% (lower dashed),50% (solid), 75% (upper dashed) and 90% (upper dotted) quantiles. Both populations (for DE and ES) were initialized on $[-1, 1]^n$ and the ES with $R_0 = 1$.

set of experiments were performed for $n = 15$, now varying the size of DE population within the set $N \in \{10, 20, 40, 80\}$. The results are presented in Figure 3.12.

Those experiments show that a small DE population, with $N = 10$, is susceptible to reach a fixed point earlier than larger populations. However, when the population becomes larger than $N = 40$, it seems that there is no further gain in increasing its size. In any case, the behavior of the DE population seems to present a decreasing convergence rate which approaches zero.
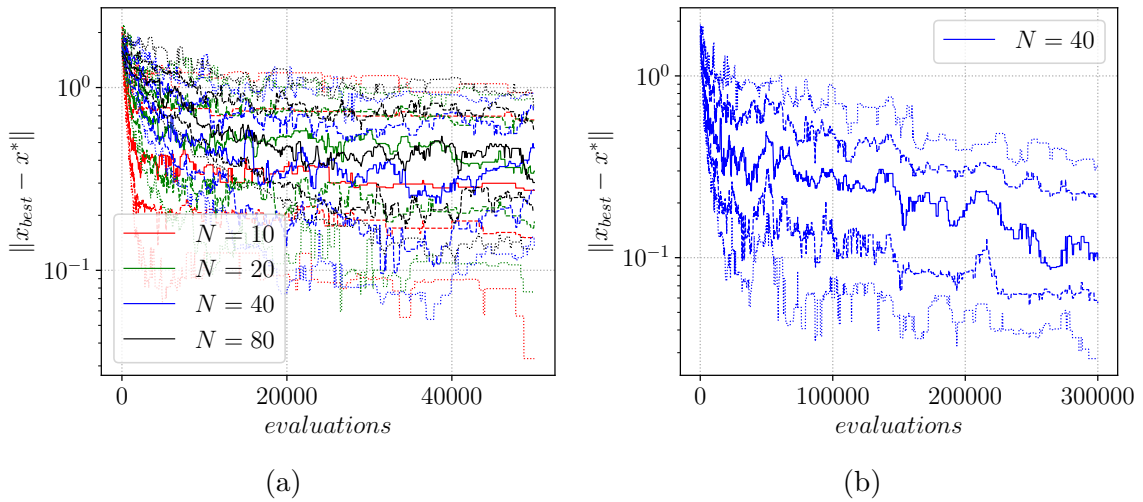
Figure 3.12: Distance, in logarithm scale, from the best solution of DE to the point of minimum of function f_4^{ns} with $\phi = 10^{-4}$ and $n = 15$, and a budget of 50000 function evaluations. (a) $N \in \{10, 20, 40, 80\}$. (b) The situation $N = 40$ is run again with a budget of 300000 function evaluations. Each curve was produced with 50 executions, representing the 10% (lower dotted),25% (lower dashed),50% (solid), 75% (upper dashed) and 90% (upper dotted) quantiles. All populations were initialized on $[-1, 1]^n$.

3.4 Coordinate transformation for function regularization

The former sections have provided evidence suggesting that:

- The DE algorithm performs better as the contour curves of the objective function become more spherical, around the point of optimal.

- After some iterations, the population of the DE becomes distributed according to the shape of the contour curves of the objective function, no matter what was its initial distribution.

This observation leads to the following idea for enhancing the performance of the DE algorithm:

1. After some iterations, compute a space coordinate transformation such that the distribution of the transformed population becomes as similar as possible to a sphere. This transformation involves the translation of the population by the vector $\mathbf{t}$, which is calculated as the mean of the current population. The translation matrix T is

defined as:

$$\mathbf{t} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{3.10}$$

$$T = \mathbf{t} \cdot \mathbf{1}_{1 \times N}$$

in which $\mathbf{1}_{1 \times N}$ denotes a row vector with N coordinates, with all coordinates filled with 1. A linear transformation matrix Q is calculated as:

$$Q = \left(\sqrt{(X - T) \cdot (X - T)^T} \right)^{-1} \tag{3.11}$$

in which $S = \sqrt{(X - T) \cdot (X - T)^T}$ is an $n \times n$ matrix such that $S \cdot S^T = (X - T) \cdot (X - T)^T$.

2. Perform the coordinate transformation, and continue the execution of DE algorithm on the transformed population. The transformed population $\tilde{X}$ is given by:

$$\tilde{X} = Q \cdot (X - T) \tag{3.12}$$

3. On each time a transformation is performed, compute also the total transformation whose inverse application leads back to the original coordinates. Start from the identity transformation:

$$\tilde{\mathbf{t}} = \mathbf{0}_{n \times 1} \qquad \tilde{Q} = \mathbb{I}_n \tag{3.13}$$

in which $\mathbf{0}_{n \times 1}$ denotes a zero column vector with n dimensions and $\mathbb{I}_n$ denotes the $n \times n$ identity matrix. On each time a transformation is performed, the total transformation is updated as:

$$\tilde{\mathbf{t}} = \tilde{\mathbf{t}} + Q^{-1} \cdot \mathbf{t}$$

$$\tilde{Q} = Q \cdot \tilde{Q} \tag{3.14}$$

4. Repeat periodically the coordinate transformation procedure.

5. On the end, transform the solution back to the original space coordinates. Assuming that the final solution, in the final coordinates, is stored in $\tilde{X}$, the final population is recovered as:

$$X = \tilde{Q}^{-1} \cdot \tilde{X} + \tilde{\mathbf{t}} \tag{3.15}$$

The following two algorithms clarify the idea, the coordinate transformation is shown in Algorithm 4 and Algorithm 5 incorporates the coordinate transformation in the DE algorithm.

A numerical experiment has been conducted in order to examine the effect of the coordinate change procedure. The DE algorithm was run on function f_4^{ns} with $\phi = 10^4$, on a decision space dimension $n = 50$. A population of $N = 100$ individuals was employed, and the parameters $C_r = 0.5$ and $F = 0.75$ were employed. In one run, the

---

Algorithm 4: Update coordinates.

Input: X, $\tilde{Q}$, $\tilde{\mathbf{t}}$, N
```
/* N:  number of individuals in population           */
/* X:  a population with N individuals in ℝⁿ          */
/* Q̃:  a n × n matrix                                 */
/* t̃:  a vector in ℝⁿ                                 */
```
Output: $X, \tilde{Q}, \tilde{\mathbf{t}}$
```
/* Get the mean of X:                                 */
```
1   $\mathbf{t} \leftarrow \frac{1}{N}\sum_{i=1}^{N} x_i$
```
/* Repeat t N times to construct T:                   */
```
2   $T \leftarrow \mathbf{t} \cdot \mathbf{1}_{1\times N}$
```
/* Compute an auxiliary variable S:                   */
```
3   $S \leftarrow \sqrt{(X-T)(X-T)^T}$
```
/* Compute Q:                                         */
```
4   $Q \leftarrow S^{-1}$
```
/* Update X, Q̃ and t̃:                                 */
```
5   $X \leftarrow Q(X-T)$
6   $\tilde{\mathbf{t}} \leftarrow \tilde{\mathbf{t}} + \tilde{Q}^{-1}\mathbf{t}$
7   $\tilde{Q} \leftarrow Q\tilde{Q}$

DE algorithm was as before, and in another run the coordinate transformation procedure was included, performing one coordinate transformation each 100 iterations. In order to establish a performance baseline, the DE algorithm was run also on function f_1, on the same dimension and using the same parameters. The results are presented in Figure 3.13.

The observation of Figure 3.13 suggests that the coordinate transformation procedure was able to avoid the effect of the population being attracted to a fixed point, which occurs in the case of the DE without that transformation. In fact, the rate of convergence that was obtained for function f_4^{ns} became almost the same that was achieved in the more favorable case of function f_1. This suggests that the contour surfaces of the function f_4^{ns} in the transformed coordinates should have become almost spherical, as intended.

3.5   Coordinate transformation with optimal C_r

In Section 2.5 we discussed how to use the analytical $p_{C_r,n}(F)$ value to optimize the choice of C_r for a given F or vice versa. Now we combine this method with the coordinate transformation from last section. Figures 3.14, 3.15, 3.16, 3.17 and 3.18 show how four versions of the algorithm behave for different test functions. The fist version is the original DE without modifications, the second is the DE with optimal C_r, the

---

Algorithm 5: Differential Evolution algorithm with coordinate transformation.

Input: N, n, F, C_r, $f(x)$
```
/* The parameters are the same as in Algorithm 2          */
```
Output: x^*
```
/* Generate initial population (N individuals in ℝⁿ):     */
```
1   $X^0 \leftarrow$ **initial_population**(N, n)
```
/* Let X⁰ = vec(x₁,···,x_N) where xᵢ ∈ ℝⁿ for all i = 1,···,n    */
```
2   $t \leftarrow 0$
```
/* Initializing coordinate transformation:                */
```
3   $\tilde{Q} \leftarrow \mathbb{I}_n$
4   $\tilde{\mathbf{t}} \leftarrow \mathbf{0}_{n \times 1}$
5   **while** ***not*** *stopping condition* **do**
6   **for** $i \leftarrow 1 : N$ **do**
7       $i_1 \leftarrow i$
```
      /* Select 3 elements, without replacement, from the set
         {1,2,...,N} − {i₁}, with uniform probability:      */
```
8       $\{i_2, i_3, i_4\} \leftarrow$ **rand_select**$(3, \{1, 2, \ldots, N\} - \{i\})$
```
      /* Retrieve the individuals to be submitted to crossover and
         mutation:                                          */
```
9       $\bar{x}_1 \leftarrow X^t[i_1]$
10      $\bar{x}_2 \leftarrow X^t[i_2]$
11      $\bar{x}_3 \leftarrow X^t[i_3]$
12      $\bar{x}_4 \leftarrow X^t[i_4]$
```
      /* Build the set 𝒜 ⊂ {1,2,...,n} by a random choice such that
         each element is put in 𝒜 with probability Cᵣ:      */
```
13      $\mathcal{A} \leftarrow$ **rand_subset**(n, C_r)
```
      /* Perform mutation and crossover:                    */
```
14      $o_i \leftarrow \sum_{j \notin \mathcal{A}} \langle \bar{x}_1, e_j \rangle e_j + \sum_{j \in \mathcal{A}} \langle \bar{x}_2 + F.(\bar{x}_3 - \bar{x}_4), e_j \rangle e_j$
```
      /* Perform elitist selection:                         */
```
15   **if** $f(\tilde{Q}^{-1} o_i + \tilde{\mathbf{t}}) < f(\tilde{Q}^{-1} \bar{x}_1 + \tilde{\mathbf{t}})$ **then**
16        $\bar{x}_i \leftarrow o_i$ /* If o_i is better then $\bar{x}_i$ updates the value of $\bar{x}_i$ */

17     $t \leftarrow t + 1$
```
    /* Store the population for the next iteration:         */
```
18     $X^t \leftarrow \text{vec}(\bar{x}_1, \ldots, \bar{x}_N)$
```
    /* Update coordinate system:                            */
```
19   **if** *its time to update coordinate system* **then**
20       $X^t, \tilde{Q}, \tilde{\mathbf{t}} \leftarrow$ **update_coordinates**$(X^t, \tilde{Q}, \tilde{\mathbf{t}}, N)$

```
  /* Get the best point that was visited in the last coordinate
     system:                                                */
```
21   $x^* \leftarrow \arg\min_{i \in \{1, \ldots, N\}} f(x_i)$
```
  /* Convert to the original coordinate system:            */
```
22   $x^* \leftarrow \tilde{Q}^{-1} \bar{x}^* + \tilde{\mathbf{t}}$

---

third is the DE with coordinate transformation and the last is the DE with coordinate transformation and optimal C_r.
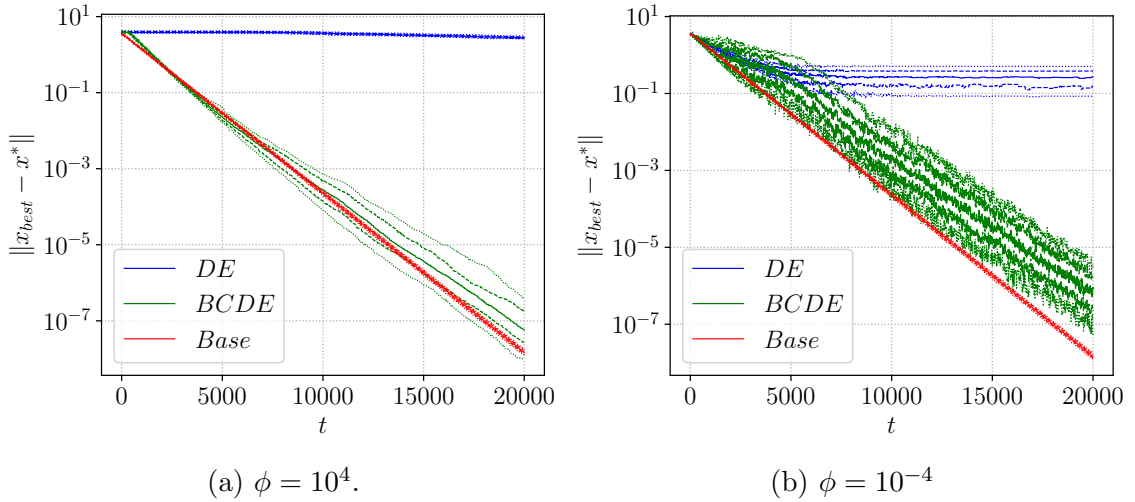
(a) $\phi = 10^4$.

(b) $\phi = 10^{-4}$

Figure 3.13: Distance, in logarithm scale, from the best solution of DE to the point of minimum of function f_4^{ns} with $n = 50$. Green: DE algorithm with coordinate transformation procedure. Blue: DE algorithm without coordinate transformation. Red: baseline data, from DE algorithm on function f_1. In all cases, the population is $N = 100$, and the parameters are defined as $C_r = 0.5$ and $F = 0.75$. Each curve was produced with 50 executions (each execution with a budget of 5×10^6 function evaluations), representing the 10% (lower dotted),25% (lower dashed),50% (solid), 75% (upper dashed) and 90% (upper dotted) quantiles. All populations were initialized on $[-1, 1]^n$.

The test bed is composed by:

1. Sphere function,

$$f_1^t(x) = ||x||_2^2 \tag{3.16}$$

   with minimum $f_1^t(x^*) = 0$ at $x^* = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}$. The population is initialized on $[-1, 1]^n$.

2. Rastrigin function,

$$f_2^t(x) = 10n + \sum_{i=1}^{n}[x_i^2 - 10cos(2\pi x_i)][-5.12, 5.12]^n \tag{3.17}$$

   with minimum $f_2^t(x^*) = 0$ at $x^* = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}$. The population is initialized on $[-5.12, 5.12]^n$.

3. Rosenbrock function,

$$f_3^t(x) = \sum_{i=1}^{n-1}[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2] \tag{3.18}$$

   with minimum $f_3^t(x^*) = 0$ at $x^* = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}$. The population is initialized on $[-5.12, 5.12]^n$.

4. Styblinski-Tang function,

$$f_4^t(x) = \frac{1}{2} \sum_{i=1}^{n} (x_i^4 - 16x_i^2 + 5x_i) - L, \qquad (3.19)$$

where

$$L = \min_{x \in \mathbb{R}^n} \frac{1}{2} \sum_{i=1}^{n} (x_i^4 - 16x_i^2 + 5x_i) = \min_{x \in \mathbb{R}} \frac{n}{2} (x^4 - 16x^2 + 5x) \approx -39.16599n, \quad (3.20)$$

with minimum $f_4^t(x^*) = 0$ at $x^* = \begin{bmatrix} c & c & \cdots & c \end{bmatrix}$, where $c \approx -2.90353$. The population is initialized on $[-5, 5]^n$.

5. Griewank function,

$$f_5^t(x) = 1 + \sum_{i=1}^{n} \frac{x_i^2}{4000} - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right) \qquad (3.21)$$

with minimum $f_5^t(x^*) = 0$ at $x^* = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}$. The population is initialized on $[-600, 600]^n$.

Each Figure shows at (a) a heat map with the two-dimensional version of the function and at (c) a cut of the two-dimensional version with x_2 constant containing the global optima. At (b) and (e) the best objective value, in logarithm scale, attained by each generation for $n = 5$ and $n = 50$, respectively. At (c) and (f) the distance, in logarithm scale, from the best solution of DE to the point of minimum of each function for $n = 5$ and $n = 50$, respectively.

All sub figures (b), (c), (e) and (f) were obtained with 50 runs, where each run used $N = 2n$, $F = 0.75$. The lines are the 10% (lower dotted), 25% (lower dashed), 50% (solid), 75% (upper dashed) and 90% (upper dotted) quantiles of the respective curves ($f(x_{best}^t)$ or $||x_{best}^t - x^*||$). The blue lines represent the DE without coordinate transformation and $C_r = 0.5$. The green lines represent the DE without coordinate transformation and $C_r = 0.5$. The red lines represent the DE without coordinate transformation and $C_r = C_r^*(F, n)$ (defined in section 2.5). The black lines represent the DE with coordinate transformation and $C_r = C_r^*(F, n)$.

In this work we archived two main results: the computation of C_r^* and the modified Differential Evolution algorithm with coordinate transformation. The value of C_r^* was calculated from the sphere function and takes profit of the separability of the sphere function. The modified DE introduces a coordinate transformation that tries to locally fit the level curves of any function onto something close to the level curves of the sphere function.

Even though it's expensive, the DE is a widely used algorithm, even to work with differentiable multimodal functions, because it tends to not get stuck in a local optimal
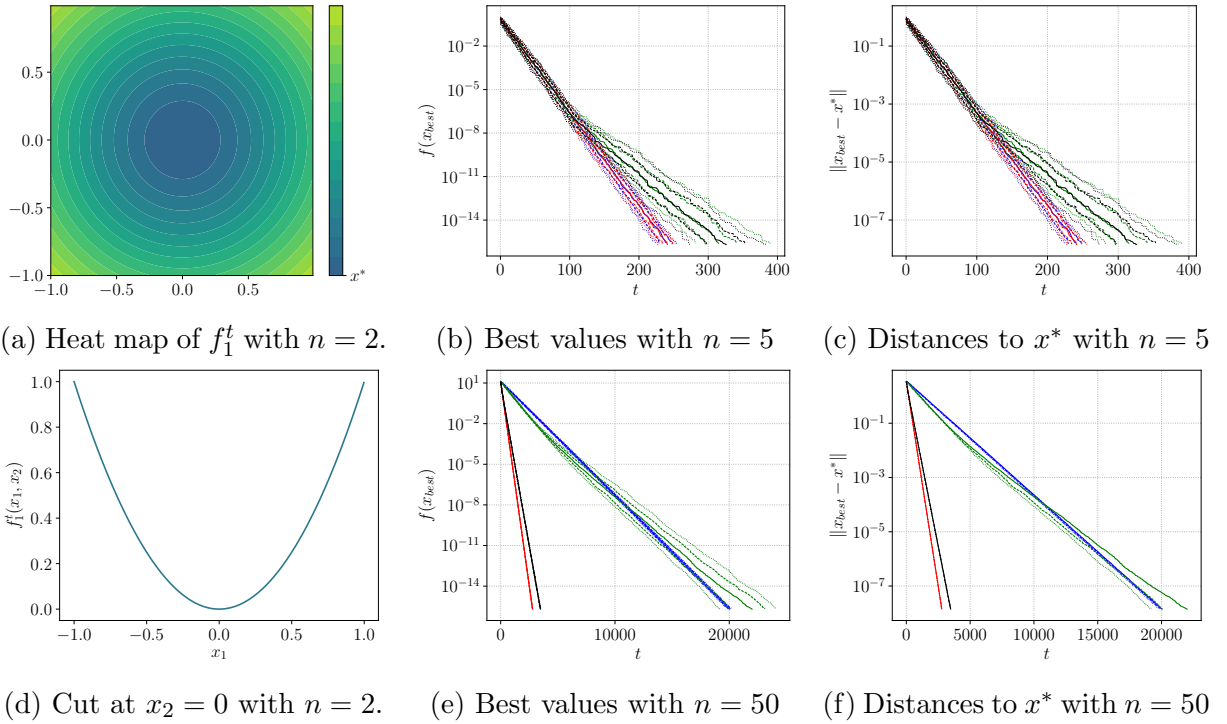
(a) Heat map of f_1^t with $n = 2$.　(b) Best values with $n = 5$　(c) Distances to x^* with $n = 5$

(d) Cut at $x_2 = 0$ with $n = 2$.　(e) Best values with $n = 50$　(f) Distances to x^* with $n = 50$

Figure 3.14: Plots for f_1^t: Sphere function. Each curve was produced from 50 trials, representing the 10% (lower dotted), 25% (lower dashed), 50% (solid), 75% (upper dashed) and 90% (upper dotted) quantiles. All populations were initialized on $[-1, 1]^n$ with $F = 0.75$. Blue: pure DE with $C_r = 0.5$; Green: DE with base transformation and $C_r = 0.5$; Red: pure DE with $C_r = C_r^*$; Black: DE with base transformation and $C_r = C_r^*$. The stopping criteria was given by $f(x_{best}) < \epsilon$, where $\epsilon \approx 2.2 \times 10^{-16}$ is the machine epsilon, or a maximum of 50000 iterations.

and look to the full picture. That is a feature that must be preserved by any modification or parameter choice.

Before analyzing the results we recall that for $F = 0.75$ we have $C_r^*(F, 5) \approx 0.421$ and $C_r^*(F, 50) \approx 0.046$. As said in Section 2.1, the experimental results in the literature point to values of $F \in [0.5, 1]$ and to values of $C_r < 0.2$ for separable functions and $C_r > 0.9$ for non-separable functions, all the results deal with low-dimensional search spaces, mainly $n < 10$. In [12] the author also studied elliptical non-separable functions, such as f_4^{ns}, saying that those functions offer the biggest challenge to the Differential Evolution.

When $n = 5$ the value of C_r^* is close to 0.5, justifying the behavior observed on all plots (b) and (c), where the solid and dashed lines were close to each other, even with a small difference in favor of the dashed lines. We can also notice that for $n = 5$ the coordinate transformation procedure wasn't helpful, it only outperforms the usual DE for the Rosenbrock function (f_3^t, Figures 3.16(b) and 3.16(c)) and only on the 10% quantile (e.g. the best executions with coordinate transformation are better then the best executions without). The geometry of the Rosenbrock function has elongated contour
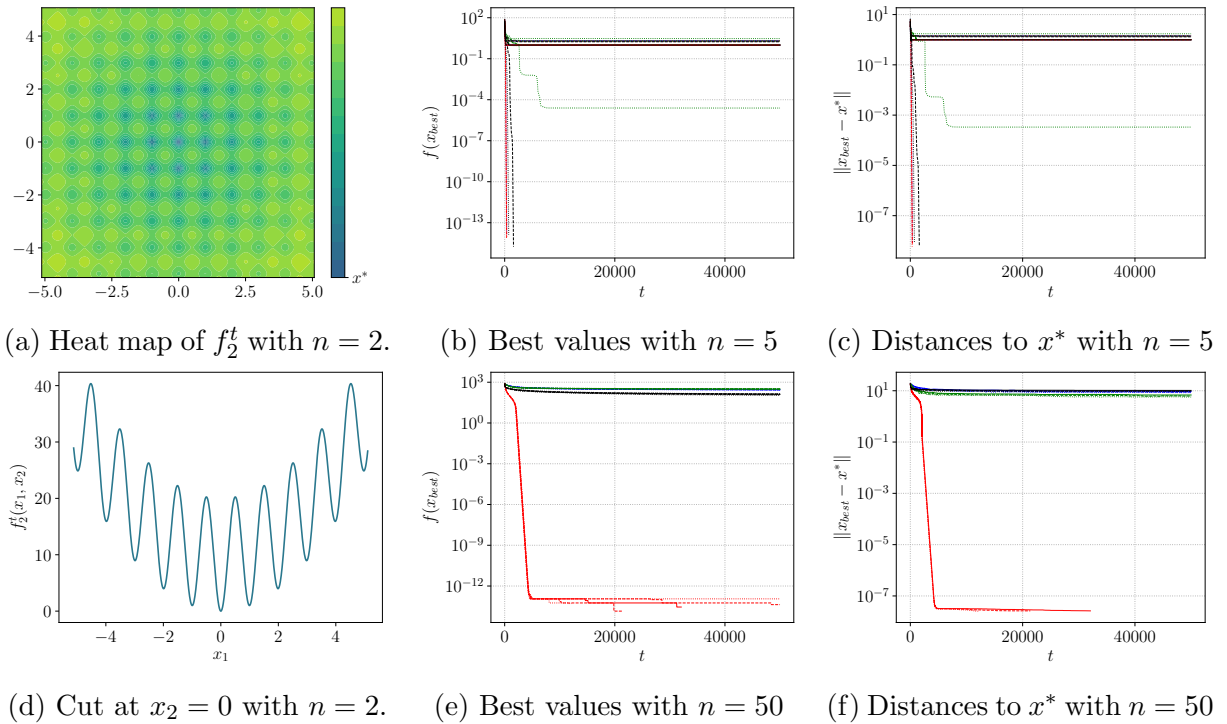
(a) Heat map of f_2^t with $n = 2$. (b) Best values with $n = 5$ (c) Distances to x^* with $n = 5$

(d) Cut at $x_2 = 0$ with $n = 2$. (e) Best values with $n = 50$ (f) Distances to x^* with $n = 50$

Figure 3.15: Plots for f_2^t: Rastrigin function. Each curve was produced from 50 trials, representing the 10% (lower dotted), 25% (lower dashed), 50% (solid), 75% (upper dashed) and 90% (upper dotted) quantiles. All populations were initialized on $[-5.12, 5.12]^n$ with $F = 0.75$. Blue: pure DE with $C_r = 0.5$; Green: DE with base transformation and $C_r = 0.5$; Red: pure DE with $C_r = C_r^*$; Black: DE with base transformation and $C_r = C_r^*$. The stopping criteria was given by $f(x_{best}) < \epsilon$, where $\epsilon \approx 2.2 \times 10^{-16}$ is the machine epsilon, or a maximum of 50000 iterations.

curves, justifying the improvement.

For the sphere function (f_1^t, Figures 3.14(b) and 3.14(c)) the transformation procedure doesn't bring any improvement, in fact, changing the basis of perfectly spherical level curves can just turn then onto a more stretched and rotated shape.

When $n = 50$ the value of C_r^* is far from 0.5. For the sphere function the DE with coordinate transformation is still worse then the pure DE for both $C_r = 0.5$ and $C_r = C_r^*$, but the choice of $C_r = C_r^*$ is better on both cases (f_1^t, Figures 3.14(e) and 3.14(f)).

The Rastrigin function is multimodal and locally separable, it has a global structure similar to the sphere function, but with waves coming down to the optimal (see Figures 3.15(a) and 3.15(d)). In that case the pure DE with optimal C_r outperforms the other three curves: the pure DE with $C_r = 0.5$ wasn't able to perform a good search as in the sphere function and the DE with coordinate transformation converges too fast to local optima. Functions f_4^t and f_5^t have a similar behavior.

The Rosenbrock function has an elongated and curved geometry, see Figure 3.16(a), that is well explored by the coordinate transformation procedure. The interesting point is that the coordinate transformation with $C_r = C_r^*$ loses to $C_r = 0.5$ since it converges
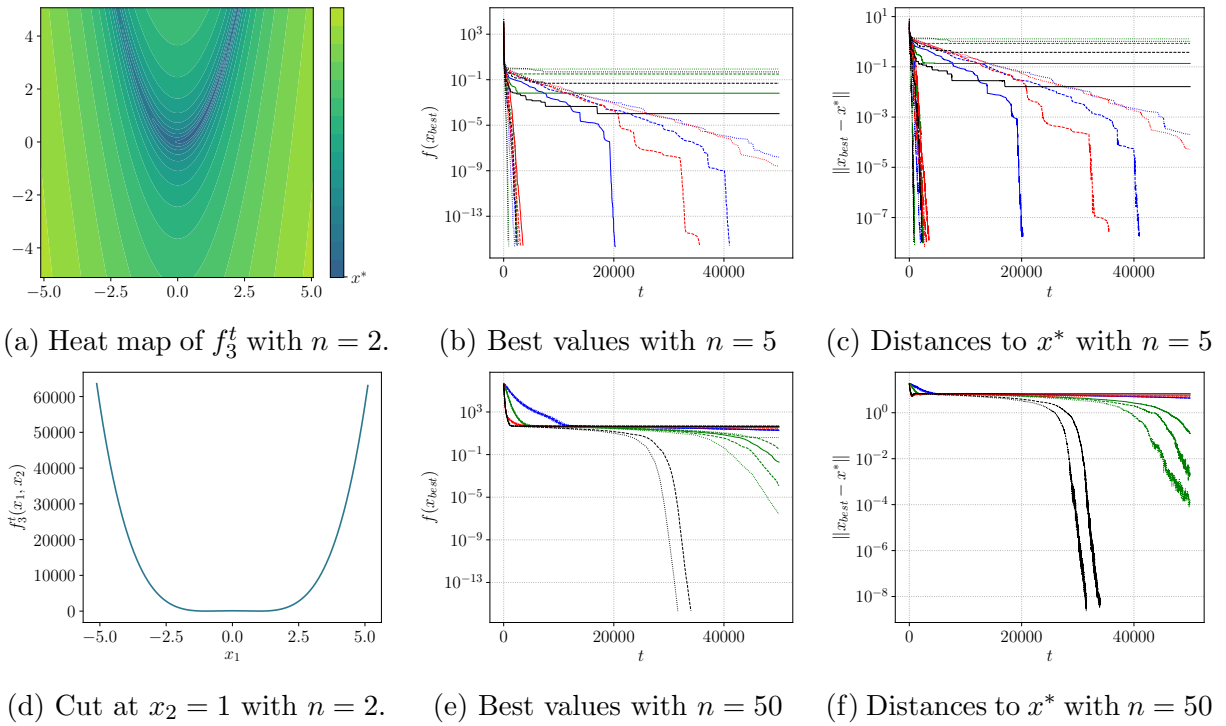
(a) Heat map of f_3^t with $n = 2$. (b) Best values with $n = 5$ (c) Distances to x^* with $n = 5$

(d) Cut at $x_2 = 1$ with $n = 2$. (e) Best values with $n = 50$ (f) Distances to x^* with $n = 50$

Figure 3.16: Plots for f_3^t: Rosenbrock function. Each curve was produced from 50 trials, representing the 10% (lower dotted), 25% (lower dashed), 50% (solid), 75% (upper dashed) and 90% (upper dotted) quantiles. All populations were initialized on $[-5.12, 5.12]^n$ with $F = 0.75$. Blue: pure DE with $C_r = 0.5$; Green: DE with base transformation and $C_r = 0.5$; Red: pure DE with $C_r = C_r^*$; Black: DE with base transformation and $C_r = C_r^*$. The stopping criteria was given by $f(x_{best}) < \epsilon$, where $\epsilon \approx 2.2 \times 10^{-16}$ is the machine epsilon, or a maximum of 50000 iterations.

too fast.

Usually, the Differential Evolution is used combined with other algorithms to perform a broad search over the search space that will then be refined by other techniques. In all experiments the pure DE with $C_r = C_r^*$ quickly approaches the global minimum. This suggests that one could use the pure DE with $C_r = C_r^*$ until it converges and then use the DE with coordinate transformation procedure to look for elongated shapes.
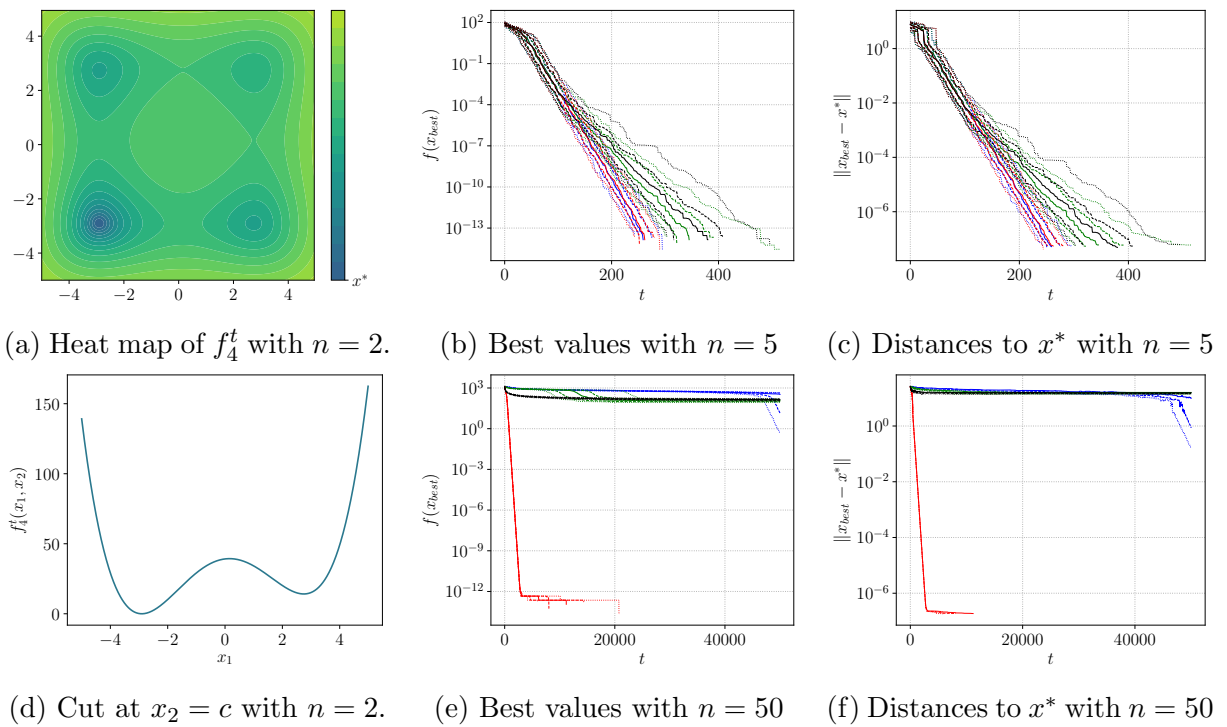
(a) Heat map of f_4^t with $n = 2$.

(b) Best values with $n = 5$

(c) Distances to x^* with $n = 5$

(d) Cut at $x_2 = c$ with $n = 2$.

(e) Best values with $n = 50$

(f) Distances to x^* with $n = 50$

Figure 3.17: Plots for f_4^t: Styblinski-Tang function. Each curve was produced from 50 trials, representing the 10% (lower dotted),25% (lower dashed),50% (solid), 75% (upper dashed) and 90% (upper dotted) quantiles. All populations were initialized on $[-5, 5]^n$ with $F = 0.75$. Blue: pure DE with $C_r = 0.5$; Green: DE with base transformation and $C_r = 0.5$; Red: pure DE with $C_r = C_r^*$; Black: DE with base transformation and $C_r = C_r^*$. The stopping criteria was given by $f(x_{best}) < \epsilon$, where $\epsilon \approx 2.2 \times 10^{-16}$ is the machine epsilon, or a maximum of 50000 iterations.
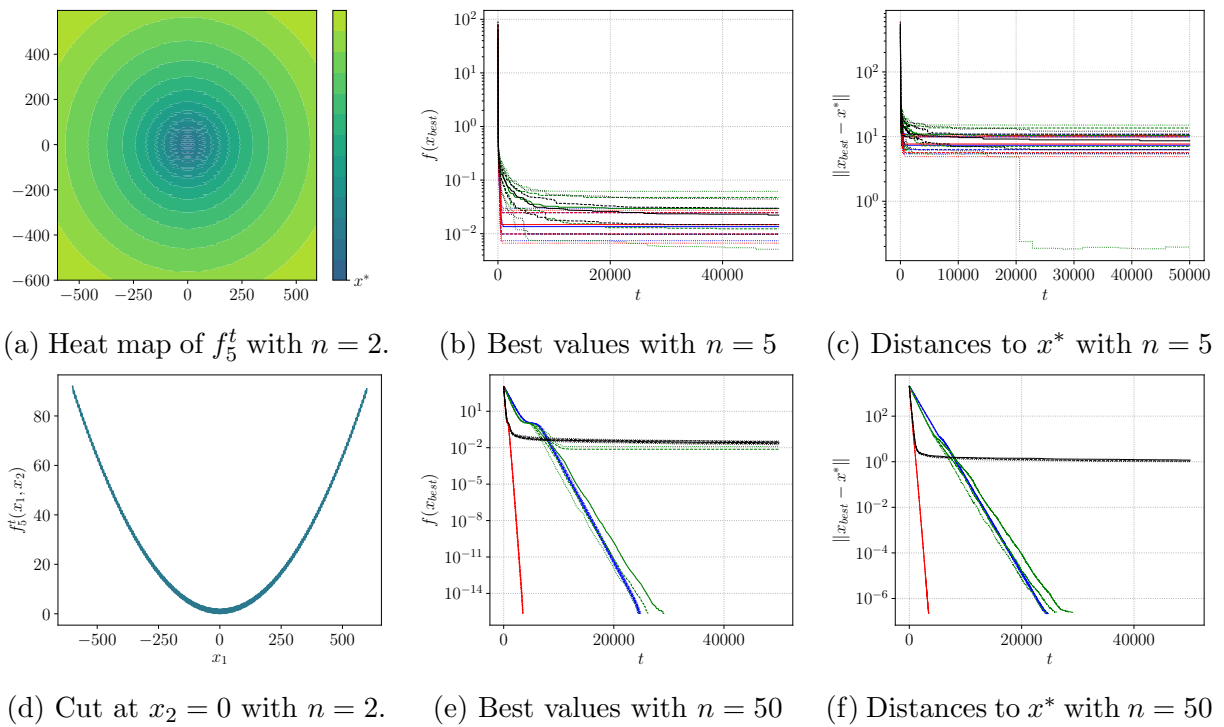
(a) Heat map of f_5^t with $n = 2$.   (b) Best values with $n = 5$ (c) Distances to x^* with $n = 5$

(d) Cut at $x_2 = 0$ with $n = 2$.   (e) Best values with $n = 50$ (f) Distances to x^* with $n = 50$

Figure 3.18: Plots for f_5^t: Griewank function. Each curve was produced from 50 trials, representing the 10% (lower dotted), 25% (lower dashed), 50% (solid), 75% (upper dashed) and 90% (upper dotted) quantiles. All populations were initialized on $[-600, 600]^n$ with $F = 0.75$. Blue: pure DE with $C_r = 0.5$; Green: DE with base transformation and $C_r = 0.5$; Red: pure DE with $C_r = C_r^*$; Black: DE with base transformation and $C_r = C_r^*$. The stopping criteria was given by $f(x_{best}) < \epsilon$, where $\epsilon \approx 2.2 \times 10^{-16}$ is the machine epsilon, or a maximum of 50000 iterations.

Chapter 4

The Differential Evolution as a Dynamical System

So far, all our results deal with the first iteration of the algorithm. It is natural to ask how the algorithm evolves over time, but this is not a simple problem. In this chapter we aim to pose this question in a more formal way, but not answering it. So, we introduce the notion of random dynamical systems and discuss how this tool can be used to model the Differential Evolution algorithm. We also discuss the current results and their limitations.

4.1 Definitions

Definition 4.1 (Metric dynamical system). *A metric dynamical system (MDS)* $\Theta = (\Omega, \mathcal{F}, \mathbb{P}, (\theta_t)_{t \in \mathbb{T}})$ *with time* $\mathbb{T}$ *(*$\mathbb{T} = \mathbb{Z}$ *or* $\mathbb{T} = \mathbb{R}$*) is a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *with a family of transformations* $\theta_t : \Omega \to \Omega$, $t \in \mathbb{T}$ *such that:*

1. *$\theta_0 = Id$ and $\theta_{t+s} = \theta_t \circ \theta_s$ for all $t, s \in \mathbb{T}$;*

2. *The map $(t, \omega) \mapsto \theta_t(\omega)$ is measurable;*

3. *The probability $\mathbb{P}$ is θ_t-invariant for all $t \in \mathbb{T}$, i.e. $\mathbb{P}(B) = \mathbb{P}(\theta_t B)$ for all $B \in \mathcal{F}$ and for all $t \in \mathbb{T}$.*

A well-known example of metric dynamical system is the left-shift operator. Let Σ be a finite set (called alphabet) with the power set σ-algebra and μ a probability measure. Let $\Omega = \Sigma^{\mathbb{Z}}$, $\mathcal{F}$ be given by the cylinder σ-algebra and $\mathbb{P}$ be the product measure, define $\theta : \Omega \to \Omega$ by $\theta((x_i)_{i \in \mathbb{Z}}) = (x_{i-1})_{i \in \mathbb{Z}}$. Using $\theta_t = \underbrace{\theta \circ \cdots \circ \theta}_{t \ times}$ and $\mathbb{T} = \mathbb{Z}$ it is easy to see that $\Theta = (\Omega, \mathcal{F}, \mathbb{P}, (\theta_t)_{t \in \mathbb{T}})$ is a metric dynamical system with time $\mathbb{Z}$.

Definition 4.2 (Random dynamical system)**.** *If $\Theta = (\Omega, \mathcal{F}, \mathbb{P}, (\theta_t)_{t \in \mathbb{T}})$ is a metric dynamical system we define a random dynamical system (RDS) on a measurable space $(X, \mathcal{B})$ over Θ as a map*

$$\varphi : \mathbb{T} \times \Omega \times X \to X \tag{4.1}$$

$$(t, \omega, x) \mapsto \varphi(t, \omega, x) \tag{4.2}$$

with the following properties:

1. *φ is measurable;*

2. *The maps $\varphi(t, \omega) : X \to X$ form a cocycle over $(\theta_t)_{t \in \mathbb{T}}$, i.e. $\varphi(0, \omega) = \mathrm{id}$ for all $\omega \in \Omega$ and $\varphi(t + s, \omega) = \varphi(t, \theta_s \omega) \circ \varphi(s, \omega)$ for all $s, t \in \mathbb{T}, \omega \in \Omega$.*

To see the Differential Evolution as a random dynamical system we first define a metric dynamical system. Let Σ be the set of all iteration rules. If

$$\lambda = ((\mathcal{A}_1, \mathcal{B}_1), \cdots, (\mathcal{A}_N, \mathcal{B}_N))$$

is an iteration rule of $k_1, \cdots, k_N$-configurations, respectively, we define a probability μ on $\mathcal{P}(\Sigma)$ by:

$$\mu(\lambda) = \prod_{i=1}^{N} \frac{C_r^{k_i}(1 - C_r)^{n-k_i}}{3!\binom{N-1}{3}} \tag{4.3}$$

Using $\Omega = \Sigma^{\mathbb{Z}}$, $\mathcal{F}$ as the cylinder σ-algebra and $\mathbb{P}$ as the measure induced by μ we construct, as in the last section, the left-shift metric dynamical system. Now we need to construct a random dynamical system. Let $X = \mathbb{R}^{n \times N}$ be the set of all populations and define

$$\mathcal{D} : \Sigma \times X \to X \tag{4.4}$$

$$(\lambda, \rho) \mapsto \mathcal{D}(\lambda, \rho) \tag{4.5}$$

to be the function that takes a rule λ and updates a population $\rho = (x_1, \cdots, x_N)$ by changing x_i to o_i when $f(o_i) < f(x_i)$. Now define $\varphi(1, \omega, \rho) = \mathcal{D}(\lambda_1, \rho)$ where $\omega = (\lambda_k)_{k \in \mathbb{N}}$ and extend φ by the recursion $\varphi(t, \omega, \rho) = \varphi(t - 1, \theta_1 \omega, \varphi(1, \omega, \rho))$, completing the definition of the Differential Evolution as a random dynamical system.

To study some properties of the functions $\mathcal{D}(\lambda, \cdot)$ we need first to write them carefully. For a fixed iteration rule $\lambda = ((\mathcal{A}_1, \mathcal{B}_1), \cdots, (\mathcal{A}_N, \mathcal{B}_N))$ we can define $\mathcal{O}(\lambda, \rho)$ as the function that associates each individual of a population with its offspring. Let $\mathcal{B}_i = (i_1, i_2, i_3, i_4)$, then

$$\mathcal{O} : \Sigma \times X \to X \tag{4.6}$$

$$(\lambda, \rho) \mapsto \rho' \tag{4.7}$$

where $\rho' = (o_1, \cdots, o_N)$ and

$$o_i = \sum_{j \notin \mathcal{A}_i} \langle x_{i_1}, e_j \rangle e_j + \sum_{j \in \mathcal{A}_i} \langle x_{i_2} + F(x_{i_3} - x_{i_4}), e_j \rangle e_j. \tag{4.8}$$

We now define $\mathcal{V} : X \to \mathbb{R}^N$ as $\mathcal{V}(\rho) = (f(x_1), \cdots, f(x_N))$ where $\rho = (x_1, \cdots, x_n)$ and finally $\mathcal{D}(\lambda, \rho) = (x'_1, \cdots, x'_N)$ where

$$x'_i = \begin{cases} o_i, \text{ if } \pi_i(\mathcal{V}(\mathcal{O}(\lambda, \rho)) - \mathcal{V}(\rho)) < 0 \\ x_i, \text{ otherwise} \end{cases} . \tag{4.9}$$

4.2 Properties of the algorithm

In the next two subsections we state some properties of the Differential Evolution that don't depend on the objective function f.

4.2.1 A linear structure

For a fixed iteration rule $\lambda = ((\mathcal{A}_1, \mathcal{B}_1), \cdots, (\mathcal{A}_N, \mathcal{B}_N))$ let $\Delta_i^1 = \Delta_{(\mathcal{A}_i, \mathcal{B}_i)}{}^1$ and $\Delta_i^0 = \Delta_{(\mathcal{A}_i, \mathcal{B}_i)}^c$ for all $i = 1, \cdots, N$. Define

$$\Delta^\xi = \bigcap_{i=1}^N \Delta_i^{\alpha_i} \tag{4.10}$$

for every tuple $\xi = (\alpha_1, \cdots, \alpha_N)$ of elements in $\{0, 1\}$. Notice that Δ^ξ is the set of populations $\rho = (x_1, \cdots, x_N)$ where the iteration rule λ can improve individuals $\{x_i : \alpha_i = 1\}$, but can't improve individuals $\{x_i : \alpha_i = 0\}$. Therefore, $\{\Delta^\xi : \xi\}$ is a finite partition of the search space.

Proposition 4.1. *Let Δ^ξ and $\mathcal{D}(\lambda, \cdot) : X \to X$ be defined as before. The restriction $\mathcal{D}(\lambda, \cdot)|_{\Delta^\xi}$ is linear[2].*

Proof of Proposition 4.1. Let $\rho = (x_1, \cdots, x_N) \in \Delta^\xi$, $\rho' = \mathcal{D}(\lambda, \rho) = (x'_1, \cdots, x'_N) \in X$ and $\xi = \{\alpha_1, \cdots, \alpha_N\}$. If $\alpha_i = 0$ then $x'_i = x_i$. If $\alpha_i = 1$ then

$$x'_i = \sum_{j \notin \mathcal{A}_i} \langle x_{i_1}, e_j \rangle e_j + \sum_{j \in \mathcal{A}_i} \langle x_{i_2} + F(x_{i_3} - x_{i_4}), e_j \rangle e_j \tag{4.11}$$

---

[1]Using your definition of $\mathcal{V}$ it is easy to see that $\Delta_{(\mathcal{A}_i, \mathcal{B}_i)} = \{\rho \in X : \pi_i(\mathcal{V}(\mathcal{D}(\rho)) - \mathcal{V}(\rho)) < 0\}$.

[2]The shape of the set $\Delta_{(\mathcal{A}_i, \mathcal{B}_i)}$ depends on the objective function f, but the linearity is a property of the algorithm for any objective function.

where $\mathcal{B}_i = (i_1, i_2, i_3, i_4)$, $i_1 = i$ and e_j is the j-th canonical vector. Thus, x_i' is a linear combination of $x_1, \cdots, x_N$ for every $i = 1, \cdots, N$. Then $\mathcal{D}(\lambda, \cdot)|_{\Delta^\xi}$ is linear. $\qquad\square$

4.2.2 Discontinuities

Since $\mathcal{D}(\lambda, \cdot)$ is piecewise linear then it is continuous in the interior of each set Δ^ξ. But it is not difficult to construct a configuration where discontinuity occurs. Take, for example, the configuration in Figure 4.1. Let $(\mathcal{A}, \mathcal{B})$ be a configuration with $\mathcal{A} = \{2\}$, $\mathcal{B} = (w, x, y, z)$ and $F = 1$, producing the offspring o of w in the figure. Also assume that the dotted circumference is the level curve for the value $f(o)$ and that the function is smaller inside the circle than outside.
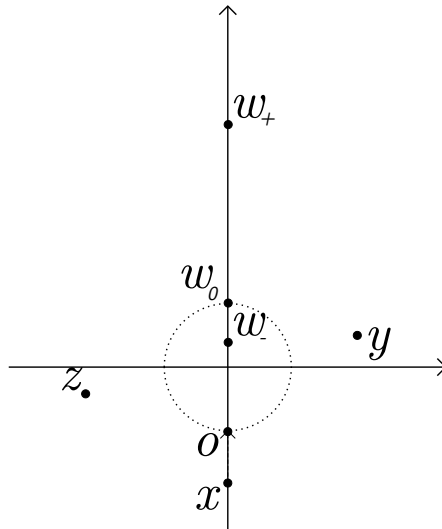

Figure 4.1: Example of discontinuity of $\mathcal{D}(\lambda, \cdot)$ at $\rho = (w_0, x, y, z)$. If w is inside the circle (lower level curves) then $\mathcal{D}(\lambda, \cdot)$ will keep w as it is. If w is outside the circle (higher level curves) then $\mathcal{D}(\lambda, \cdot)$ will exchange w for o. So, if w_+ is outside the circle on the y-axis and w_- is inside the circle, also on the y-axis, then $\lim_{w_+ \to w_0} \mathcal{D}(\lambda, \cdot) = (o, \cdot, \cdot, \cdot)$, but $\lim_{w_- \to w_0} \mathcal{D}(\lambda, \cdot) = (w_0, \cdot, \cdot, \cdot)$.

If w is inside the circle (e.g. $w = w_-$) then $\mathcal{D}(\lambda, \cdot)$ will keep w as it is. If w is outside the circle (e.g. $w = w_+$) then $\mathcal{D}(\lambda, \cdot)$ will exchange w for o. So, if λ is an iteration rule λ that has the configuration $(\mathcal{A}, \mathcal{B})$, then $\mathcal{D}(\lambda, \cdot)$ is discontinuous at $\rho = (w_0, x, y, z)$. In fact, we can even introduce a perturbation and find more discontinuous configurations, all that matters is that w transits from lower level curves to higher level curves without passing near o. The following proposition characterizes the set of discontinuities of $\mathcal{D}(\lambda, \cdot)$.

Proposition 4.2. *Let Δ^ξ and $\mathcal{D}(\lambda, \cdot) : X \to X$ be defined as before. The set $C^\xi = \{\rho \in \partial\Delta^\xi : \mathcal{D}(\lambda, \cdot) \text{ is continuous at } \rho\}$ has zero measure.*

Proof of Proposition 4.2. Let $\rho = (x_1, \cdots, x_N) \in C^\xi$. Since $\rho \in \partial\Delta^\xi$ then there exists a configuration $(\mathcal{A}_i, \mathcal{B}_i)$ such that $\rho \in \Delta_{(\mathcal{A}_i, \mathcal{B}_i)}$ and a sequence of populations $\rho_n \notin \Delta_{(\mathcal{A}_i, \mathcal{B}_i)}$, $n \geq 0$, such that $\rho_n \to \rho$. Note that it is also possible that $\rho \notin \Delta_{(\mathcal{A}_i, \mathcal{B}_i)}$ and $\rho_n \in \Delta_{(\mathcal{A}_i, \mathcal{B}_i)}$, but this case is analogous.

By continuity we have that $\mathcal{D}(\lambda, \rho_n) \to \mathcal{D}(\lambda, \rho)$. Since $\rho \in \Delta_{(\mathcal{A}_i, \mathcal{B}_i)}$ we know that $x_i' = o_i$ and since $\rho_n = (x_1{}^n, \cdots, x_N{}^n) \notin \Delta_{(\mathcal{A}_i, \mathcal{B}_i)}$ we also know that $x_i{}^{n'} = x_i{}^n$. Then $x_i{}^n - o_i \to 0$ when $\rho^n \to \rho$ and so $x_i = o_i$. Then $C^\xi \subset \cup_{i=1}^N \{\rho \in X : x_i = o_i\}$.

Since $\{\rho \in X : x_i = o_i\}$ is a proper subspace of X when $\mathcal{A}_i \neq \emptyset$, then C^ξ has zero measure. The case $\mathcal{A}_i = \emptyset$ has $\Delta_{(\mathcal{A}_i, \mathcal{B}_i)} = \emptyset$ and then $\rho \notin C^\xi$. $\square$

Then, in terms of measure, the set of continuous points of $\mathcal{D}(\lambda, \cdot)$ is basically $\cup_\xi (\Delta^\xi)^\circ$.

4.3 A brief literature review and expectations

Unfortunately, as pointed out by [10], there is a lack of well-developed theoretical foundations when it comes to evolutionary algorithms analysis. On the other hand, a lot of work has been done in Random Dynamical Systems. We now want to discuss what kind of results exist and what are the main barriers to apply them to the Differential Evolution algorithm.

It is easy to see that if f has a global minimum then $\mathcal{V}(\varphi(t, \omega, \rho))$ always converges (not necessarily to the minimum) since $\pi_i(\mathcal{V}(\varphi(t + 1, \omega, \rho))) \leq \pi_i(\mathcal{V}(\varphi(t, \omega, \rho)))$ for all $t \geq 0$, $\omega \in \Omega$ and $\rho \in X$. But since the system contains discontinuities it is not even clear if the population itself will always converge to a fixed point. We also know, from Theorem 2.2 that even when improvements happen they could be arbitrarily bad.

On the other hand, our experimental results show that exponential convergence (at least for functions like the sphere function) seems to be the rule. The most studied Random Dynamical System are the linear cocycles, the definition follows.

Definition 4.3 (Linear cocycle). *Let $\Theta = (\Omega, \mathcal{F}, \mathbb{P}, (\theta_t)_{t \in \mathbb{T}})$ be a metric dynamical system with time $\mathbb{T} = \mathbb{Z}$, $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ be the usual $\mathbb{R}^d$ with the Borel σ-algebra and $A : \Omega \to GL(d)$ be a measurable function with values on the linear group $GL(d)$ of $d \times d$ invertible matrices*

with real coefficients. A linear cocycle is a random dynamical system over Θ *given by*

$$\varphi : \mathbb{Z} \times \Omega \times \mathbb{R}^d \to \mathbb{R}^d \tag{4.12}$$

$$(t, \omega, x) \mapsto A^t(\omega)x \tag{4.13}$$

where $A^t(\omega) = A(\theta_{t-1}(\omega)) \circ A(\theta_{t-2}(\omega)) \circ \cdots \circ A(\omega)$.

There are many theorems concerning about the quantities

$$\gamma_+(\omega) = \lim_t \frac{1}{t} \log ||A^t(\omega)|| \text{ and } \gamma_-(\omega) = \lim_t \frac{1}{t} \log ||A^t(\omega)^{-1}||^{-1}, \tag{4.14}$$

called Lyapunov exponents (see [17] for some examples). The function $\gamma_+(\omega)$ counts how much the cocycle can stretch a vector from $\mathbb{R}^d$ (since that is the definition of the induced norm) while the function $\gamma_-(\omega)$ counts how much the cocycle can shrink a vector from $\mathbb{R}^d$. Some results, like the Oseledets theorem (also in [17]), in fact, calculate those quantities.

The Differential Evolution is not a linear cocycle, but it is piecewise linear. A similar result would give information about the convergence rate of the algorithm, at least for the sphere function.

Another property, that is commonly employed in the study of non-linear Random Dynamical Systems, is Lipschitz continuity (see [5] for an example). When $\varphi(t, \omega, \cdot) = \phi_{\theta_t\omega} \circ \cdots \circ \phi_{\theta_0\omega}$ we say that the Random Dynamical System is an Iterated Random Function and if all ϕ are Lipschitz we say that the System is Lipschitz. Taking $\phi_{\theta_t\omega} = \mathcal{D}(\lambda_t, \cdot)$ where $\omega = (\lambda_k)_{k \in \mathbb{N}}$ we can see the Differential Evolution as an Iterated Random Function, but not a Lipschitz one.

Although it is not Lipschitz, one may assume, for instance, that $||\mathcal{D}(\lambda, \rho) - \rho^*|| \leq ||\rho - \rho^*||$ where $\rho^* = (x^*, \cdots, x^*)$, i.e., that the population isn't always getting further away from the optimum. That is the case for well behaved functions near the optimum, take $f : \mathbb{R}^n \to \mathbb{R}$ with such that f'' exists and is positive definite, for example.

Other feature that needs careful examination is the set of fixed points, Theorem 2.1 shows that high values of F are strongly related to the amount of fixed points. Zaharie ([18]) shows that low values of F are related to premature convergence. Is there a value of F for which the set of fixed points has zero measure at least for radial functions? Our experimental results stand in favor of the existence of such value, but our attempts to prove it have failed.

The main difficulty posed by the Differential Evolution is the elitist selection, that imposes natural discontinuity. Some works ([3] and [4]) remove the elitist selection changing it for a smoother function. They try to approximate the Differential Evolution algorithm with a continuous time model and study the properties of that model. Then, using tools like Lyapunov's second method, they were able to establish analogies with the gradient descending method and to study the behavior of their model near the optimum.

As [10] says, most theoretical approaches to the Differential Evolution only try to study a simpler model of the algorithm or only look at one operation, not capturing the interactions caused by multiple iterations.

As we show, behind its simplicity and elegance, the Differential Evolution has an intrinsically complex behavior, with fixed points, discontinuities and low quality enhancements. An ideal tool to study the algorithm needs to be able to counterbalance those characteristics with the linearity over iterations.

Chapter 5

Conclusion

Our two goals with this work were to better understand the dynamics of the Differential Evolution and to develop a method for helping the choose of algorithm parameters. Theorems 2.1 and 2.2 strike both problems.

By setting the objective function as $f(x) = ||x||^2$ and computing the probability of enhancement after a full iteration we were able to improve Zaharie's work [18] with further understanding of the parameters F and C_r. We show that when the dimension n of the search space increases and the parameters are kept fixed the probability that an enhancement occurs goes exponentially to zero. That computation was also useful for measuring the set of fixed points.

We also developed a heuristics to choose the parameter C_r when F is given. Our experimental results showed the efficacy of our heuristics. The well behaved and calculated probabilities for $f(x) = ||x||^2$ led to investigation of separable and non-separable functions, producing a new Differential Evolution variant based on coordinate transformations to deal with functions with elongated geometry. Together with Zaharie's results, we can propose the following method to choose parameters: (a) If the function one is trying to optimize has an elongated geometry choose the variation with change of coordinates; (b) Take N according with the available computational power; (c) Then, let $F > F_{crit}$ to avoid premature convergence, where

$$F_{crit} = \sqrt{\frac{1}{N} - \frac{C_r}{2N}} < \sqrt{\frac{1}{N}}, \tag{5.1}$$

but take care because high values of F are related with fixed points. Experimental results from the literature [15] recommend $F \in \left[\frac{1}{2}, 1\right]$; (d) Finally, let $C_r = C_r^*$.

At the end, without the assumption that $f(x) = ||x||^2$, we also presented results about the dynamical structure of the algorithm, viewing it as a Random Dynamical System. This may represent a meaningful contribution to the comprehension of the problem itself.

References

[1] Robert G. Bartle. *The Elements of Integration and Lebesgue Measure.* John Wiley & Sons, New York, 1995.

[2] Stephen Chen, James Montgomery, and Antonio Bolufé-Röhler. Measuring the curse of dimensionality and its effects on particle swarm optimization and differential evolution. *Applied Intelligence*, 42(3):514–526, 2015.

[3] Sambarta Dasgupta, Arijit Biswas, Swagatam Das, and Ajith Abraham. The population dynamics of differential evolution: A mathematical model. In *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pages 1439–1446. IEEE, 2008.

[4] Sambarta Dasgupta, Swagatam Das, Arijit Biswas, and Ajith Abraham. On stability and convergence of the population-dynamics in differential evolution. *AI Communications*, 22(1):1–20, 2009.

[5] Persi Diaconis and David Freedman. Iterated random functions. *SIAM Review*, 41(1):45–76, 1999.

[6] Sever Silvestru Dragomir, Ravi P Agarwal, and Neil S Barnett. Inequalities for beta and gamma functions via some classical and new integral inequalities. *Journal of Inequalities and Applications*, 2000(2):103–165, 2000.

[7] Zhongbo Hu, Qinghua Su, Xianshan Yang, and Zenggang Xiong. Not guaranteeing convergence of differential evolution on a class of multimodal functions. *Applied Soft Computing*, 41:479–487, 2016.

[8] David A Levin and Yuval Peres. *Markov Chains and Mixing Times*, volume 107. American Mathematical Society, 2017.

[9] Kim-Hung Li. Reservoir-sampling algorithms of time complexity o (n (1+ log (n/n))). *ACM Transactions on Mathematical Software (TOMS)*, 20(4):481–493, 1994.

[10] Karol R Opara and Jarosław Arabas. Differential evolution: A survey of theoretical analyses. *Swarm and Evolutionary Computation*, 44:546–558, 2019.

[11] Kenneth V Price. Differential evolution: a fast and simple numerical optimizer. In *Proceedings of North American Fuzzy Information Processing*, pages 524–527. IEEE, 1996.

[12] Jani Ronkkonen, Saku Kukkonen, and Kenneth V Price. Real-parameter optimization with differential evolution. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 506–513. IEEE, 2005.

[13] Günter Rudolph. Convergence of evolutionary algorithms in general search spaces. In *Proceedings of IEEE International Conference on Evolutionary Computation*, pages 50–54. IEEE, 1996.

[14] Albert N Shiryaev. *Probability*. Springer, New York, 1996.

[15] Rainer Storn. On the usage of differential evolution for function optimization. In *Proceedings of North American Fuzzy Information Processing*, pages 519–523. IEEE, 1996.

[16] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, 1997.

[17] Marcelo Viana. *Lectures on Lyapunov Exponents*, volume 145. Cambridge University Press, 2014.

[18] Daniela Zaharie. Critical values for the control parameters of differential evolution algorithms. In *Proc. of MENDEL 2002, 8th International Conference on Soft Computing*, pages 62–67, 2002.

Appendix A

This appendix presents some definitions and results about Markov chains used in the previous chapters. For further reference see [8] and [14].

Definition A.1. *Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space, $(E, \mathcal{F})$ be a measurable space and T a set. If $X_t : \Omega \to E$ is a random variable for every $t \in T$, then the family $(X^t)_{t \in T}$ is a stochastic process with time T. If $T = \mathbb{N}$ the process has discrete time. The set E is the state space and if $T = \mathbb{N}$ the sequence $X^0(\omega), X^1(\omega), \cdots$ is a sample sequence for every fixed $\omega \in \Omega$.*

Definition A.2. *A discrete time Markov chain is a stochastic process $(X^t)_{t \in T}$ such that*

$$\mathbb{P}(X^t \in A | X^{t_1}, \cdots, X^{t_k}) = \mathbb{P}(X^t \in A | X^{t_k}) \ a.e. \tag{A.1}$$

for every $A \in \mathcal{F}$, $k \in \mathbb{N}$ and $0 < t_1 < \cdots < t_k < t$. If

$$\mathbb{P}(X^{t+s} \in A | X^t) = \mathbb{P}(X^s \in A | X^0) \ a.e. \tag{A.2}$$

for every $t, s \in T$, then the Markov chain is also homogeneous.

If $(X^t)_{t \in \mathbb{N}}$ is a discrete homogeneous Markov chain then we can understand it by looking at its kernel.

Definition A.3. *The kernel of a discrete homogeneous Markov chain $(X^t)_{t \in \mathbb{N}}$ is a function $K : E \times \mathcal{F} \to [0, 1]$ such that*

1. *$K(\rho, A) = \mathbb{P}(X^1 \in A | X^0 = \rho)$ for every $\rho \in E, A \in \mathcal{F}$;*

2. *$K(\rho, \cdot) : \mathcal{F} \to [0, 1]$ is a measure on $(E, \mathcal{F})$ for every fixed $\rho \in E$;*

3. *and $K(\cdot, A) : E \to [0, 1]$ is measurable for every fixed $A \in \mathcal{F}$.*

We also define the t-th iteration of K by

$$K^{(t)}(\rho, A) = \begin{cases} K(\rho, A), \ if \ t = 1 \\ \int_E K^{(t-1)}(y, A) K(\rho, dy), \ if \ t > 1 \end{cases} \tag{A.3}$$

for every $t \geq 1$.

Then, if X^0 has a law $p(\cdot)$ we have

$$\mathbb{P}(X^t \in A) = \begin{cases} p(A), \text{ if } t = 0 \\ \int_E K^{(t)}(y, A)p(dy), \text{ if } t \geq 1 \end{cases} \tag{A.4}$$

for every $t \in \mathbb{N}$.

Appendix B

Here we state Zaharie's theorem again and transcript her demonstration.

Theorem ([18]). *Let $\rho = (x_1, \cdots, x_N)$ be a scalar population, $U = (u_1, \cdots, u_N)$ the result of the mutation step and $O = (o_1, \cdots, o_N)$ the result of the crossover step between ρ and U, then*

$$\mathbb{E}(\mathit{Var}(U)) = \left(2F^2 + \frac{N-1}{N}\right)\mathit{Var}(\rho) \tag{B.1}$$

and

$$\mathbb{E}(\mathit{Var}(O)) = \left(1 + 2C_r F^2 - \frac{2C_r}{N} + \frac{C_r^2}{N}\right)\mathit{Var}(\rho). \tag{B.2}$$

Proof. Since

$$\mathbb{E}(\mathrm{Var}(U)) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}(x_{i_1} + F(x_{i_2} - x_{i_3}))^2\right) - \mathbb{E}\left(\left(\frac{1}{N}\sum_{i=1}^{N}(x_{i_1} + F(x_{i_2} - x_{i_3}))\right)^2\right) \tag{B.3}$$

we can calculate $\mathrm{Var}(U)$ by calculating the two terms above as following:

$$\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}(x_{i_1} + F(x_{i_2} - x_{i_3}))^2\right) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}(x_{i_1}^2 + 2Fx_{i_1}(x_{i_2} - x_{i_3}) + F^2(x_{i_2} - x_{i_3})^2)\right) \tag{B.4}$$

$$= \mathbb{E}\left(x_{i_1}^2\right) + 2F\mathbb{E}(x_{i_1}(x_{i_2} - x_{i_3})) + F^2\mathbb{E}\left((x_{i_2} - x_{i_3})^2\right) \tag{B.5}$$

Using that $\mathbb{E}(x_{i_j}^2) = \mathbb{E}(\rho^2)$ for all $j = 1, 2, 3$ and that $\mathbb{E}(x_{i_1}x_{i_2}) = \mathbb{E}(x_{i_2}x_{i_3}) = \mathbb{E}(x_{i_1}x_{i_3})$ we write:

$$\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}(x_{i_1} + F(x_{i_2} - x_{i_3}))^2\right) = (1 + 2F^2)\mathbb{E}(\rho^2) - 2F^2\mathbb{E}(x_{i_2}x_{i_3}) \tag{B.6}$$

Now, for the second term we compute:

$$\mathbb{E}\left(\left(\frac{1}{N}\sum_{i=1}^{N}(x_{i_1} + F(x_{i_2} - x_{i_3}))\right)^2\right) = \mathbb{E}\left(\frac{1}{N^2}\sum_{i=1}^{N}(x_{i_1} + F(x_{i_2} - x_{i_3}))^2\right. \tag{B.7}$$

$$\left. + \frac{1}{N^2}\sum_{i\neq l}(x_{i_1} + F(x_{i_2} - x_{i_3}))(x_{l_1} + F(x_{l_2} - x_{l_3}))\right) \tag{B.8}$$

Since $\mathbb{E}(x_{i_1} + F(x_{i_2} - x_{i_3}))$ is independent of i and when $i \neq l$ it holds that x_{i_j} and x_{l_k} are independent for every $1 \leq j, k \leq 3$:

$$\frac{1}{N^2}\mathbb{E}\left(\sum_{i \neq l}(x_{i_1} + F(x_{i_2} - x_{i_3}))(x_{l_1} + F(x_{l_2} - x_{l_3}))\right) = \frac{N-1}{N}\left(\mathbb{E}(x_{i_1} + F(x_{i_2} - x_{i_3}))\right)^2 \tag{B.9}$$

$$= \frac{N-1}{N}\left(\mathbb{E}(\rho)\right)^2 \tag{B.10}$$

Hence:

$$\mathbb{E}\left(\text{Var}(U)\right) = \left(1 - \frac{1}{N}\right)\left((1 + 2F^2)\mathbb{E}(\rho^2) - 2F^2\mathbb{E}(x_{i_2}x_{i_3}) - (\mathbb{E}(\rho))^2\right) \tag{B.11}$$

Now we need to evaluate $\mathbb{E}(x_{i_1}x_{i_2})$, by definition:

$$\mathbb{E}(x_{i_1}x_{i_2}) = \sum_{j,k} x_j x_k \mathbb{P}(i_1 = j, i_2 = k) \tag{B.12}$$

$$= \frac{1}{N(N-1)}\sum_{j \neq k} x_j x_k \tag{B.13}$$

$$= \frac{1}{N-1}\left(\sum_j x_j\left(\frac{1}{N}\sum_{k \neq j} x_k\right)\right) \tag{B.14}$$

$$= \frac{1}{N-1}\left(\sum_j x_j\left(\mathbb{E}(\rho) - \frac{x_j}{N}\right)\right) \tag{B.15}$$

$$= \frac{N(\mathbb{E}(\rho))^2 - \mathbb{E}(\rho^2)}{N-1} \tag{B.16}$$

Therefore:

$$\mathbb{E}\left(\text{Var}(U)\right) = \left(1 - \frac{1}{N}\right)\left(1 + 2F^2\frac{N}{N-1}\right)\left(\mathbb{E}(\rho^2) - (\mathbb{E}(\rho))^2\right) \tag{B.17}$$

$$= \left(2F^2 + \frac{N-1}{N}\right)\text{Var}(\rho) \tag{B.18}$$

To compute $\mathbb{E}(\text{Var}(O))$ we start with

$$\mathbb{E}\left(\text{Var}(O)\right) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} o_i^2\right) - \mathbb{E}\left(\left(\frac{1}{N}\sum_{i=1}^{N} o_i\right)^2\right) \tag{B.19}$$

and then evaluate each term. The first goes by the linearity of the expectation:

$$\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} o_i^2\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(o_i^2) \tag{B.20}$$

$$= \frac{1}{N}\sum_{i=1}^{N} C_r\mathbb{E}(u_i^2) + (1 - C_r)\mathbb{E}(x_i^2) \tag{B.21}$$

$$= C_r\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} u_i^2\right) + (1 - C_r)\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) \tag{B.22}$$

And on the second term we use the independence of o_i and o_j when $i \neq j$:

$$\mathbb{E}\left(\left(\frac{1}{N}\sum_{i=1}^{N}o_i\right)^2\right) = \frac{1}{N^2}\mathbb{E}\left(\left(\sum_{i=1}^{N}o_i^2\right) + \left(\sum_{i\neq l}o_i o_l\right)\right) \tag{B.23}$$

$$= \frac{1}{N}\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}o_i^2\right) + \frac{1}{N^2}\sum_{i\neq l}\mathbb{E}(o_i)\mathbb{E}(o_l) \tag{B.24}$$

Now we need to calculate $\sum_{i\neq l}\mathbb{E}(o_i)\mathbb{E}(o_l)$, since $\mathbb{E}(u_i) = \mathbb{E}(\rho)$ we write:

$$\sum_{i\neq l}\mathbb{E}(o_i)\mathbb{E}(o_l) = \sum_{i\neq l}(C_r\mathbb{E}(u_i) + (1-C_r)x_i)(C_r\mathbb{E}(u_l) + (1-C_r)x_l) \tag{B.25}$$

$$= \sum_{i\neq l}(C_r\mathbb{E}(\rho) + (1-C_r)x_i)(C_r\mathbb{E}(\rho) + (1-C_r)x_l) \tag{B.26}$$

$$= \left(\sum_{i=1}^{N}(C_r\mathbb{E}(\rho) + (1-C_r)x_i)\right)^2 - \sum_{i=1}^{N}(C_r\mathbb{E}(\rho) + (1-C_r)x_i)^2 \tag{B.27}$$

$$= (N^2 - 2NC_r + NC_r^2)(\mathbb{E}(\rho))^2 - N(1-C_r)^2\mathbb{E}(\rho^2) \tag{B.28}$$

Replacing B.28 on B.24 and then B.24 and B.22 on B.19 one obtains B.2, completing the proof. $\qquad\square$